

STANISŁAW UBERMANOWICZ  
Uniwersytet im. Adama Mickiewicza  
w Poznaniu

Bolesław Niemierko, *Próbna matura 1997 – ewaluacja metodologiczna*

„Znany jest typ *eksperta-tyrana*, który po łaskawym udzieleniu konsultacji stanowczo domaga się realizacji swoich wniosków, obraża się na opornych, donosi do ministerstwa i do prasy, wydziera władzę kierownictwu, tłumi inicjatywę wykonawców, skłóca wewnętrznie i niszczy opiniowany program. Jeżeli – z powodu ogólnego zapóźnienia – nie mamy jeszcze w kraju ekspertów-tyranów (własnych lub zagranicznych), wkrótce się niechybnie pojawią. Bądźmy czujni!”

*Bolesław Niemierko*

Ta sugestywna izofora pochodzi z wykonanego na zlecenie MEN opracowania, w którym Profesor Niemierko występuje w roli metaewaluatora, czyli eksperta oceniającego wyniki prac Zespołów Regionalnych, w ramach Programu *Nowa Matura*. Powstanie ekspertyzy datowane jest na lipiec–wrzesień 1996 r., toteż chyba jedynie jasnowidztwu zawdzięczać można opisanie (jak sądzi ekspert) próbnej matury z marca 1997. W rzeczywistości materiały dotyczą próbnej matury z roku 1996, a ekspertyzę pisano o rok później. Lapsus podwójnego antydatowania jest jednak błahostką wobec nietrafnych ocen i błędnych wniosków sformułowanych w dokumencie. Oto najbardziej znamienne rysy i skazy owej ekspertyzy:

- Pomieszanie ewaluacji matury z badaniem naukowym, nastawionym na parametryczne sprawdzanie realizacji celów odartych z wartości i treści.
- Rugowanie nurtu naturalistycznego, mimo iż teoria ewaluacji traktuje ten typ wartościowania równoprawnie z podejściem naukowym.
- Promowanie ewaluacji technokratycznej, z pozbawieniem nauczycieli i uczniów prawa do współdecydowania o kształcie Nowej Matury.

- Bezpodstawne zarzuty nierzetelności wobec odmiennych modeli pomiaru i nowszych parametrów niż zalecane przez szkołę gdańską.
- Propagowanie jako wzorcowej i godnej powielania techniki badawczej z kardynalnymi błędami metodologiczno-interpretacyjnymi.

Wbrew własnej przestrodze ekspert wpadł w sidła autokracji, wyrządzając szkodę demokratycznemu modelowi kreowania Nowej Matury. Skutki skażenia całego Programu tak jednowymiarowym sterowaniem eksperckim będą fatalne, gdyż proces optymalizacji matury kierowany jest przez metaewaluatora wciąż w tę samą wąską uliczkę, z szyldem „Pomiar sprawdzający”. A gdzie wartości?

#### BADANIE NAUKOWE CZY EWALUACJA MATURY?

Główną słabością ekspertyzy jest to, że uznaje ona ewaluację matury tylko jako badanie naukowe. Tymczasem słowo *ewaluacja* nie jest jakimś nowomodnym synonimem badania naukowego, lecz pojęciem oznaczającym nową strategię analizy i oceny rzeczywistości. W ujęciu demokratycznym oznacza wartościowanie. Wykazać można, iż oba pomieszczone przez metaewaluatora procesy generalnie różnią się w założeniach. Z badań naukowych formułowane są raczej ponadczasowe twierdzenia o stanie rzeczy, natomiast w efekcie ewaluacji stwierdza się, jak oceniany jest stan aktualny. Domeną uczonych jest głównie budowanie teorii poprzez uogólnienia, natomiast rolą ewaluatorów jest wspieranie praktyki poprzez wyszczególnienie i szacowanie cech składających się na daną wartość.

Proponowane przez metaewaluatora naukowe metody *sprawdzającego pomiaru dydaktycznego* nie są ewaluacją, odcinają się bowiem zupełnie od wartości formy i treści matury. Stawia się cel, a zadanie badacza polega jedynie na sprawdzeniu, czy ów cel został osiągnięty. W takim założeniu w ogóle nie jest ważna ocena, czy sam cel był słuszny. Zalecane w ekspertyzie badania naukowe nad czterema tylko parametrami „matury jako pomiaru” nie są w stanie wnieść nic przydatnego do procesu optymalizacji jej ducha i materii. Rewelacje z takowych badań, w rodzaju stwierdzenia, iż każde wypracowanie z języka polskiego powinno oceniać przynajmniej trzech nauczycieli mogą co najwyżej służyć pozyskiwaniu stopni... naukowych.

Na szczęście profesjonalni ewaluatorzy wyzwolili się już z okowów nadobnej naukowości, pracując przede wszystkim na rzecz społecznej użyteczności oraz realności. W swoich ocenach wartościujących sięgają po opinie wszystkich zainteresowanych sprawą: liderów reformy oświatowej, kreatorów nowych form matury, metodyków przedmiotowych, dyrektorów szkół, a przede wszystkim nauczycieli i uczniów. Włączają owe grupy do współdecydowania w sprawie matury, mimo iż kategorycznie sprzeciwia się temu autorytarny ekspert. A przecież społeczne wartościowanie formatywne jest jedyną drogą uzyskania niezbędnego konsensu wokół tak doniosłego egzaminu, jakim jest matura. Ta droga z całą pewnością nie kłóci się z teorią i praktyką nauk społecznych.

#### PODEJŚCIE SCJENTYSTYCZNE CZY NATURALISTYCZNE?

Ekspert winien traktować równoprawnie wszystkie komplementarne nurty ewaluacji, nawet jeśli sam uznaje wyłącznie jedno podejście. Uzasadnię ową potrzebę wielu wymiarów wartościowania i wyjaśnić różnice między nurtem *scjentystycznym* a odrzucanym *naturali-*

stycznym, posługując się odniesieniem do oceny tak szlachetnego trunku, jakim (poza Polską) jest wino. Otóż scjentyści preferują obiektywne miareczkowanie, stąd oznaczają tylko parametry mierzalne: procent  $C_2H_5OH$ , ilość cukru, kwasowość. Naturaliści natomiast preferują subiektywne doznawanie: oceniają bukiet w całości, a w szczególności smak, aromat i cechy organoleptyczne. Z kolei konsumenci wina doświadczają wszystkiego: mocy, słodkości bądź wytrawności, podziwiają bukiet, smak, aromat, barwę, klarowność, a wreszcie ponoszą koszty. Trzeba więc wyraźnie zaznaczyć, że oba podejścia pomiarowe – obiektywistyczne i subiektywistyczne – są na równi cenne, gdyż wzajemnie uzupełniają się w ogólnej ocenie jakości.

Dotyczy to także Nowej Matury. Nie da się wmówić abiturientom i ich nauczycielom, że statystycznie egzamin był trafny i obiektywny, a oceny rzetelne, jeśli oni w większości mają zupełnie odmienne osobiste odczucia. Co sezon roczniki maturalne wraz z najbliższym otoczeniem przechodzą przez chorobę egzaminacyjną, a tymczasem superekspert zabrania lekarzom pytać pacjentów o to, co im dolega. Oczywiście można wprowadzić sondę, pobrać wycinek, postawić diagnozę i rozpocząć kurację, jednak etyka zobowiązuje do uzyskania uprzedniej zgody na wszystkie te działania. Co więcej, do wyleczenia konieczne jest współdziałanie.

Jedynie ewaluatorzy nie muszą współdziałać, gdyż w ich pracy cenniejsza jest wielość niezależnych metod. Z tego względu niepokoi narzucanie przez eksperta wzorca badań wyłącznie pod kątem poprawności pomiarowej, z pomijaniem całego spektrum immanentnych kontekstów i czynników wpływających na *jakość matury*. Wspomnę choćby: uniwersalizm kulturowy standardów egzaminacyjnych (zakres i słuszność tego, co matura mierzy), wartości treściowo-merytoryczne i etyczne (jakość i prawość substancji matury, obiektywizm sędziów, sprawiedliwość ocen), a wreszcie organizacyjne (dopasowanie czasu pisania do poziomu trudności zadań, trafność i rzetelność egzaminu). Z pewnością nie ujmuje to wszystkich grup cech jakościowych, oczekiwanych od Nowej Matury, lecz jest ilustracją tego, jak wiele niezwykle ważnych komponentów profesor Niemierko w ogóle ignoruje. A co gorsza – z pozycji *guru* niszczy inne modele ewaluacji.

#### EWALUACJA AUTOKRATYCZNA CZY DEMOKRATYCZNA?

Nawet jeśli ekspertyza powstała na zlecenie MEN, to wcale nie oznacza, że miała ona służyć odgórnemu sterowaniu. Słusznie bowiem głównym filarem Programu *Nowa Matura* jest oddolny ruch reformatorski. Jednak sam metaewaluator określił jednoznacznie swą pozycję wyroczni, odbierając prawo głosu podmiotom matury. Przypomina to postawę polityka, Ryszarda Bugaja, który twierdził, że *jeśli nawet istnieją wartości uniwersalne, to i tak nie wiadomo, kto i jak ma je wyznaczać*. Wskutek owej niewiedzy uznawał za słuszne jedynie prawa stanowione... przez siebie. Społeczeństwo udzieliło mu w tej kwestii całkowicie naturalnej wykładni, eliminując w wyborach nie tylko autorytarnego lidera, lecz również jego partię.

W ewaluacji demokratycznej nie można uciekać od tego, co stanowi podstawę wartościowania. Oto w pewnych kwestiach doniosłych i zasadniczych (np. wartość życia) udaje się uzyskać jednoznaczną afirmację społeczną. Wobec innych kwestii, co do których w ogóle nie wolno szukać jedności (np. światopogląd), wartością staje się pluralizm i tolerancja. Zadajmy profesorowi kluczowe pytanie: Czy chce dla Nowej Matury szukać konsensusu społecznego? Jeśli tak, to nie może ograniczać roli nauczycieli wyłącznie do odbioru informacji

o tym, co postanowił Minister za radą ekspertów. Opinie nauczycieli w sprawie jakości matury są bardziej miarodajne niż głos nawet najlepszego w kraju specjalisty od metrologii dydaktycznej.

Matura może służyć Ministerstwu, lecz nade wszystko winna społeczeństwu. Protest profesora przeciwko modelom demokratycznej ewaluacji matury dziwnie nie pasuje w dobie wyzwolenia społeczeństwa. Przecież *podmioty matury* nie mogą być nadal niewolnikami *instytucji matury*. Jeśli ewaluatorzy i decydenci zaczną wreszcie respektować głosy nauczycieli i abiturientów, to w przyszłości zostanie zaaprobowany nawet pomiar częstszy i doskonalszy, polegający na systematycznym przeprowadzaniu diagnostycznych testów ogólnokrajowych, już w klasach III, VI szkół podstawowych oraz na końcu obu etapów szkół średnich. Takie testy byłyby szczepionką ochronną, a nie końcową resekcją, gdy na leczenie za późno.

#### MODEL GDAŃSKI CZY OGÓLNIEJSZY?

Dla metaewaluatora najważniejsze jest to, czy matura *dobrze mierzy*. Uznaje tylko te eksperymenty, które wyznaczają: niezależność, dokładność, rzetelność, trafność oraz obiektywizm matury. Statystyczne wskaźniki *wewnętrznej jakości samego eksperymentu* bezzasadnie utożsamia z jakością pomiarową egzaminu. Nadużywa zwłaszcza współczynnik rzetelności, choć jest to tylko jedna z wielu cech rozkładu empirycznego, a nie rzetelność matury. Do czego prowadzi taki model? Oto w eksperymencie gdańskim najlepszym przypadkiem jest ten, gdzie wszyscy oceniający wystawili identyczną ocenę. W takim pomiarze nikogo nie martwi, że jest to ocena niedostateczna! Wręcz przeciwnie – gdyby jedynek wystawiono więcej, współczynnik rzetelności byłby wyższy, co sugerowałoby, że kryteria są lepsze. Takie oto mogą być skutki sprowadzania jakości matury do powtarzalności ocen, bez uzyskania aprobaty kryteriów nawet wśród sędziów kompetentnych.

Wystawianie ocen za wypracowania naturalne to trudny proces decyzyjny, podejmowany w sytuacji ograniczenia danych, niezbędnych do podjęcia trafnych postanowień. Tymczasem metaewaluator jednoznacznie oskarża nauczycieli o brak obiektywizmu w wystawianiu ocen własnym uczniom. Z góry zakłada, że wykorzystują oni wiedzę o swoich wychowankach w złą wierzę. Nic dziwnego, że przy takim założeniu cały wysiłek reformatorów i ewaluatorów stara się kierować na techniczną stronę samego pomiaru, czemu rzekomo służyć ma kodowanie prac, dublowanie oceniających, wyprowadzenie matury poza szkołę itp.

Ja akcentuję ważkość innych działań. Twierdzę, że oceny za wypracowania wystawiane nawet własnym uczniom mogą być rzetelne i sprawiedliwe, jeśli opierać się będą na autentycznie doskonałych, powszechnie aprobowanych kryteriach. Opowiadam się za pełniejszym schematem modelowania i badania, obejmującym nie tylko ewaluację egzaminu, lecz także społeczną walidację kryteriów. W takim schemacie najpierw konieczne jest opracowanie *wzorca struktury i treści kryteriów*, uwieńczony pomiarem stwierdzającym *jednoznaczny poziom aprobaty* dla każdego z zawartych tam komponentów oceniania. Dopiero po wyłonieniu sędziów kompetentnych, całkowicie zgadzających się z kryteriami, można mierzyć *powtarzalność wystawianych ocen*. Ewaluacja i walidacja w poznańskim modelu ogólniejszym jest wartościowaniem i zatwierdzaniem standardów nie tylko przez powołane do tego instytucje, lecz także przez społeczność związaną z maturą, a zwłaszcza przez nauczycieli oceniających prace egzaminacyjne.

## STATYSTYKI KLASYCZNE CZY NOWATORSKIE?

Profesor Niemierko stosuje tradycyjne miary statystyczne, mylnie zakładając dla skali ocen rozkłady normalne. Bardzo słusznie w ekspertyzie gani za jedynie procentowe relacje z badań, lecz w tym samym dokumencie krytykuje nowsze, pogłębione, stosowane w Poznaniu instrumentarium *Skali ważonych ocen*. To zupełnie tak, jakby uznawał wyłącznie wzór na ruch jednostajny  $S = V/t$ , a odrzucał wersję ogólniejszą  $S = dV/dt$ . Tymczasem w praktyce, podobnie jak ze zmiennością ruchu, najczęściej mamy do czynienia z asymetrią rozkładów ocen. Z tego względu ja stosuję dokładniejsze miary statystyczne, w których celowo wyodrębniam dwa składniki wariancji – dodatni i ujemny. Choć szokuje pojęcie *wariancji ujemnej*, to nie ma błędu, gdyż momenty centralne rzędu drugiego można liczyć jako sumę iloczynu różnic przez moduły z różnic. Służą one do obustronnego wyznaczania odchylenia realnych, które charakteryzują rozkład empiryczny wierniej niż teoretycznie symetryczne odchylenia standardowe.

Ekspertowi obce jest pojęcie *uchybu*, spieszę więc z wyjaśnieniem. Otóż w metodzie ważenia opinii, gdy na jednej „szali” umieścimy propozycje reformatorów matury, a na drugiej opinie podmiotów matury, nie ma bardziej celnego słowa niż „uchyb” na określenie parametru, który oznacza *zakłócenie stanu równowagi wskutek przeniesienia części głosów respondentów na przeciwstawną ogółowi stronę*. Nie jest to ani błąd pomiaru, ani uchybienie proceduralne, jak podejrzewa ekspert. Jest to wskaźnik odpowiadający wprost na pytanie, jak sprawdzają się założenia innowacyjne w rzeczywistości egzaminacyjnej.

Metaewaluator nie zna stosowanego w Poznaniu instrumentarium, lecz mimo to twierdzi, że jest ono nierzetelne. Ja mam klarowniejszą sytuację, znam bowiem klasyczne techniki wykorzystane w Gdańsku i z całą odpowiedzialnością mogę orzec, iż w przykładzie podanym w ekspertyzie jako wzorzec analizy pogłębionej zastosowano je błędnie.

## POMYŁKA CZY FAŁSZYWE PRZEŚWIADCZENIA?

Wspomniałem już, że współczynnik rzetelności jest jedynie wskaźnikiem jakości wewnętrznej eksperymentu, a nie rzetelności wystawianych stopni. Jeśli gdański eksperyment z jednym sędzią ma niską rzetelność, to na podstawie tak słabego pomiaru nie można niczego orzekać. Nie wolno więc wnioskować, że ocenianie wypracowań przez jedną osobę nie wystarcza. Fakt wzrostu współczynnika rzetelności przy większej liczbie oceniających wynika jedynie z takiej właśnie cechy wewnętrznej parametru, a nie jest efektem potwierdzającym tezę eksperymentu. Jaka zresztą może być przydatność różniących się ocen wielu sędziów, skoro ocenę uczniowi można wystawić tylko jedną?

Mylna jest teza o potrzebie oceniania każdego z wypracowań przez co najmniej trzech nauczycieli. Ów błąd systematyczny stosowanej tam metody bierze się stąd, że oceniający mogą posługiwać się wyłącznie stopniami całkowitymi, podczas gdy badacze stosują technikę porównań odniesioną do zmiennej ciągłej. To zupełnie tak, jakby nakazać mierniczym podawanie wymiarów wyłącznie w centymetrach, po czym „naukowo” wykazać, że wynik uśredniony w milimetrach jest rzetelniejszy. Aby nie być gołosłownym, zacytuję odpowiedni fragment ekspertyzy: „*zastosowanie schematu punktowania zmieniło niektóre oceny o jeden lub nawet dwa stopnie, ale prawie nie wpłynęło na średnie arytmetyczne ocen dziewięciu*

sędziów. (Współczynnik korelacji tych średnich sięga 0,99, ale korelacja tych średnich z ocenami tych prac przez nauczycieli tych uczniów wynosi już tylko 0,61 i 0,60)". Zauważmy, jak beztrzesko i bezpodstawnie miesza się i porównuje korelaty wartości dyskretnych ze średnimi.

Z całą pewnością osoba występująca w roli eksperta bierze na siebie całkowitą odpowiedzialność za propagowanie przykładu badań z tak kardynalnymi błędami metodologiczno-interpretacyjnymi. Pomyłkę można by wybaczyć, lecz niedorzeczności w ekspertyzie tolerować nie sposób. Spróbuję choćby częściowo zniwelować złe skutki nietrafnych opinii i sugestii zawartych w omawianym opracowaniu.

#### PROSTUJMY ŚCIEŻKI

Profesor Niemierko krytykuje narzędzie ewaluacyjne biologów. Oczywiście zgoda, że za mało jest tam stwierdzeń spolaryzowanych ujemnie, którym respondenci winni zaprzeczać. Jednakże co do funkcji superlatywnych sformułowań ankiety, ekspert jest w błędzie. Prawdopodobnie nie zna techniki celowego *justowania estymatorów*, polegającej na dodaniu takich słów, które utrudniają zdecydowaną aprobatę zdań ankiety. W konsekwencji takiego zabiegu wzrasta moc wnioskowania z wypowiedzi. Przecież trudniej respondentom zaakceptować stwierdzenie – *To jest doskonała matura*, niżli – *To jest matura*. Słówka stopniujące jakość (jasno, głęboko, bardziej) pełnią w ankiecie doniosłą rolę, dokładnie odwrotną niż sugeruje to ekspert; wcale nie zachęcają do aprobaty formy testowej. Z danych empirycznych obliczyć można, że *współczynniki konkluzyjności* tych krytykowanych stwierdzeń są bliskie 0,7, co dla ankiety nie będącej skalą standaryzowaną jest dobrym osiągnięciem.

Odwrotnie jest z narzędziem ewaluacyjnym polonistów. Ekspert wyraża aplauz wobec ankiety, podczas gdy zawarte w niej pytania zamknięte w ogóle nie nadają się do podjęcia interpretacji. Czyż nauczyciele biorący udział w eksperymencie mogą nie stosować kryteriów będących *indicatum* (obiektom ewaluacji)? Czyż wolno im zatajać kryteria przed uczniami z tego tylko względu, że sami je nie akceptują? Zdumiewające, jak fachowiec od pomiarów dydaktycznych może chwalić dwoiste stwierdzenie ankiety: *Arkusze oceny i recenzji jest zbędny, niepotrzebnie komplikuje pracę*. Jest to przecież zdanie złożone, dotyczące kwestii wzajemnie sprzecznych. Większość nauczycieli uważa wszak kryteria za konieczne, mimo że są one w pewnym stopniu skomplikowane. Jaką ci respondenci mieli wybrać wypowiedź wobec tak dysonansowego bodźca-stwierdzenia? Oczywiście dostrzegli ową wadę ankiety, co skłoniło ich do obszerniejszych wypowiedzi otwartych. Wyrażone opinie jawnie przeczą temu, co głosi metaewaluator, jakoby *wymagania wielostopniowe do oceny wypracowań budowane były wzorowo*. Okazało się mianowicie, iż nie tylko warstwa merytoryczna kryteriów oceniania wypracowań, lecz nawet ich struktura wymaga jeszcze gruntownej przebudowy. Wzorowy jest jedynie sam zamysł.

Z kolei adresowany do matematyków apel o wprowadzenie „wszelkich form zadań z treścią” wydawał się słuszny, ale tylko do czasu weryfikacji empirycznej. Eksperyment próbnej matury dowiódł, że na końcowym egzaminie nie można sprawdzać kompetencji, której szkoła nie potrafi ukształtować. To właśnie jest przyczyną zarzutu – odpieranego przez eksperta, a stawianego przez innych – budowania „komina bez fundamentów”. Nie wystarczy stwierdzenie fachowca, że komin jest zgodny z normą, jeśli domownicy zatruci są czadem. Warto zbadać, czy sama norma jest słuszna. Przykładem błędu już z założenia jest

namawianie matematyków, ażeby na próbnej maturze porównywali współczynnik rzetelności jednego zadania problemowego z siedmioma zadaniami testowymi. Dla każdego, kto rozumie istotę obliczania i interpretacji współczynnika rzetelności, oczywisty jest bezsens takiej propozycji. Cóż, niektórzy scjentyści nie chcą tego zrozumieć.

#### RÓWNOWAŻMY OPINIE

Ewaluacja matury nie może być jedynie wyrwanym z kontekstu społecznego pomiarem parametrów rzeczywistości, lecz nade wszystko powinna być próbą uzewnętrznienia i analizy tego, jak podmioty matury doświadczają rzeczywistość egzaminacyjną. W szerokim ujęciu określenia „podmioty” mieszczą się nie tylko uczniowie zdający, lecz także nauczyciele poprawiający prace oraz kreatorzy matury: autorzy pytań, tematów, zbiorów zadań i twórcy kryteriów oceniania. Niezbędne jest bilansowanie czynów i skutków, poprzez sprawdzanie tego, na ile propozycje reformatorów matury znajdują powszechne uznanie jako wartości.

Współczesna ewaluacja wykracza poza rutynowe techniki pomiaru i sterowania, stając się integralnym czynnikiem procesu demokratyzacji. Tymczasem aktualny metaewaluator obstaje wyłącznie przy jednej perspektywie widzenia Nowej Matury. Powstaje więc konieczność powołania drugiego eksperta o podejściu komplementarnym, zgodnym z najnowszymi tendencjami wartościowania, promowanymi przez program TERM, a ujętymi w książce *Ewaluacja w edukacji* (red. L. Korporowicz, Oficyna Naukowa, '98). Zresztą prof. Niemierko zastrzega, iż pozostanie tylko przy metodzie badania dokumentów innowacji, a więc przy ekspertyzie, mającej bardziej charakter recenzji wydawniczej niżli ewaluacji. Z takowej perspektywy matura to jedynie pomiar dydaktyczny abstrahujący od treści. Ja oczekuję czegoś więcej. Marzę, ażeby Nowa Matura stała się dobrym sprawdzianem kompetencji będących prawdziwym odzwierciedleniem powszechnie uznawanego uniwersum kulturowego oraz pięknym zwieńczeniem wspólnego wysiłku uczniów i ich mistrzów.

