

Bolesław NIEMIERKO
Uniwersytet Gdański

CHŁODNE OBLICZE EGZAMINU ZEWNĘTRZNEGO

Rozpocznę nie od chłodu, lecz od pożaru uczuć. Ze względu na swe ogromne znaczenie społeczne — egzaminy szkolne rozpalają wyobraźnię uczestników i obserwatorów. Przeciw egzaminatorom często padają zarzuty niekompetencji i złej woli, zaś wszelkie pomysły ulepszeń są zrazu przeceniane, a potem — równie hałaśliwie — odrzucane. Wtedy szuka się nowych koncepcji i nowych ludzi, zdolnych wzbudzić zaufanie polityków odpowiedniego szczebla.

Od kilku lat z zapalem przygotowujemy egzaminy dla zreformowanego szkolnictwa i wiadomo już, że nie unikniemy rozczarowań. Ten referat jest napisany właśnie w obawie przed zmarnowaniem dorobku nauczycieli, szkół i komisji egzaminacyjnych. Jego tezą jest konieczność ściślejszego współdziałania wewnętrznych i zewnętrznych form egzaminu. By tę konieczność uzasadnić, spróbuję skojarzyć wybrane elementy teorii pomiaru dydaktycznego ze spontanicznym, potocznym doświadczeniem w ocenianiu osiągnięć uczniów.

Twierdzenie, że dwie różne formy egzaminu są w pełni użyteczne we wzajemnym powiązaniu, brzmi mało rewelacyjnie. Bardziej odkrywcze będzie oparcie dowodu na rejestrze słabości tych egzaminów, nie zaś ich mocnych punktów. Inaczej bowiem niż np. w krytyce literatury i sztuki, która w sposób widoczny tym dziedzinom służy — teoretyka pomiaru dydaktycznego uważa się często za bezdusznego apologetę testów, a przynajmniej za entuzjastę surowego egzaminowania, niewrażliwego na losy uczniów i nauczycieli. Czas wyjść poza te stereotypy.

POZBĄDŹMY SIĘ ZŁUDZEŃ

Zacznę od próby rozwiania złudzeń, jakie wiążą się z dwoma rodzajami egzaminu, a dokładniej — z dwoma rodzajami oceniania: codziennym, bieżącym ocenianiem osiągnięć uczniów, które możemy nazwać egzaminem powszednim (ang. *low-stakes examination*), oraz z ocenianiem w specjalnie organizowanych akcjach, mających istotny wpływ na karierę szkolną i zawodową adepta, które będziemy nazywać egzaminem doniosłym (*high-stakes examination*).

Egzamin powszedni może przybrać postać wypracowania domowego, odpytywania, klasówki lub kartkówki. Egzamin doniosły miewa postać ustną komisyjną, pisemnego eseju lub testową, a rzadziej — praktyczną. Te złożone czynności bywają traktowane przez klientelę edukacji z mieszaniną fascynacji i podejrzliwości, jaka byłaby właściwa na pokazie sztuczek magicznych, a nie w procesie planowego działania pedagogicznego. Winą za to obarczam autorytarną przeszłość szkoły i wciąż niewysoki poziom świadomości społecznej dotyczącej zasad pomiaru osiągnięć uczniów.

EGZAMIN POWSZEDNI

Złudzenie 1. Egzamin powszedni przebiega w warunkach dogodnych dla ucznia i nauczyciela. Te warunki to spokój, swojskość, życzliwość i wzajemne zaufanie, jako przeciwstawienie surowej obcości egzaminu doniosłego. Jednak liczne pokolenia uczniów doświadczyły w swych szkołach czegoś zupełnie innego: stresu „wyrwania” do tablicy lub wyczytania nazwiska, „nerwówki” przed klasówką, niecierpliwości i gniewu nauczyciela, podpowiadania i ściągania, przeszkadzania. O tych wszystkich przykrościach zapominamy zbyt łatwo, gdy tylko na horyzoncie pojawi się wizja egzaminu bardziej zobiektywizowanego, zewnętrznego.

Złudzenie 2. Dzięki dobrej znajomości uczniów egzamin powszedni jest zindywidualizowany. Nauczyciel miałby zatem dobrać treść i formę egzaminu do osobowości ucznia, czego na ogół nie przewiduje procedura egzaminu doniosłego. Badania naukowe pokazują jednak, że orientacja nauczyciela w warunkach pracy ucznia jest daleka od doskonałości, a jego diagnoza uzdolnień uczniów jest silnie zabarwiona oceną innych pożądaných cech osobowości¹. Jeszcze trudniej jest znaleźć metodę kształcenia i egzaminowania, która wykazywałaby interakcję z uzdolnieniami uczniów, a mianowicie sprzyjała osiągnięciom mniej uzdolnionych uczniów (Cronbach, Snow 1977). „Indywidualizacja” egzaminu powszedniego okazuje się więc najczęściej tylko obniżeniem wymagań, czym zajmę się w dalszej części tego referatu.

Złudzenie 3. Nauczyciel jasno mówi swoim uczniom, czego będzie od nich wymagał. Według żywiących to złudzenie niewątpliwym obowiązkiem uprzedzenia o treści i formie sprawdzania osiągnięć jest spełniany zawsze przez nauczycieli, a nie jest właściwie spełniany przez zewnętrzne komisje. To zdanie jest prawdziwe tylko przy założeniu, że ograniczymy się do odtwarzania „przerobionego materiału”, a uczniowie zarejestrują nie tylko informację przedmiotową, lecz także preferencje i nawyki egzaminacyjne swoich nauczycieli, czyli „z czego i jak pyta” każdy nauczyciel. Wszakże w zakresie wyższych kategorii celów kształcenia podstawy programowe i standardy wymagań egzaminacyjnych są dokładniejsze niż pytania nauczyciela, a więc na nich właśnie musimy oprzeć *kształcenie według wymagań* (Niemierko 2000a).

Złudzenie 4. Nauczyciel zna prawdę o wynikach kształcenia, choć nie zawsze ją wypowiada. Taki pogląd daje przewagę intuicyjnej ocenie holistycznej, opartej na uogólnionym doświadczeniu w pracy z uczniem, nad oceną analityczną, opartą na budowaniu zadań do rozwiązania przez ucznia. Jakkolwiek intuicja jest w edukacji

¹ W wielkim dziele Artura Jensena *Błąd systematyczny w pomiarze właściwości umysłowych* czytamy: „Najważniejszy zapewne jest fakt, że nauczyciele zwykli osądzać raczej wiek umysłowy ucznia niż jego iloraz inteligencji, tak więc starsi uczniowie w klasie są szacowani zbyt wysoko, a najmłodszy są niedoszacowani. W jednym z naszych studiów na ten temat okazało się, iż wprawdzie opinie nauczycieli korelowały około 0,60 z ilorazami inteligencji według Lorge'a-Thorndike'a, to korelowały około 0,80 z wiekiem umysłowym w wynikach tego testu. Nauczyciele przeceniają także dziewczęta i nie doceniają chłopców, podczas gdy ilorazy inteligencji mierzone testem nie różnicują płci. W ocenach nauczycieli potulne zachowanie i dobre maniery są w pewnym stopniu mylone z oszacowaniem inteligencji. Inne mylące czynniki to ekstrawersja i zapal. Uczniowie, którzy więcej mówią, chętniej podnoszą rękę do odpowiedzi, zgłaszają się na ochotnika itp., są, co zrozumiale, najczęściej przeceniani przez nauczycieli” (Jensen 1980, 173-4).

niezwykle cenna, to często bywa omylna (Bruner 1964). Przeważającym modelem egzaminu powszedniego jest tendencyjne sprawdzanie hipotez, w którym obraz rzeczywistości dostosowuje się do subiektywnych przekonań, a nie odwrotnie, zaś rola emocji w kształtowaniu się przekonań nauczyciela o osiągnięciach uczniów oraz przekonań uczniów o swych własnych osiągnięciach jest bezsporna (Wojciszke 2001). Łatwo sprawdzić, że i uczniowie, i nauczyciele są zwykle nazbyt optymistycznie nastawieni, jeśli chodzi o opanowane wiadomości i umiejętności (Niemierko 1990, 290–291).

EGZAMIN DONIOSŁY

Złudzenie 5. Ujednolicenie narzędzi i procedur usuwa błąd wyniku egzaminu doniosłego. Niespecjaliści pomiaru na ogół zapominają, że egzamin jest próbą określonych zachowań badanego, reprezentującą określoną populację zachowań, nie zaś tylko okazją zademonstrowania wybranych osiągnięć. O precyzji wnioskowania z próby o populacji mówi nam błąd standardowy pomiaru, pochodna współczynnika rzetelności². Obok dokładności punktowania zadań, w tym zwłaszcza zadań rozszerzonej odpowiedzi, jak np. wypracowanie maturalne, treść zadań, różnorodna w granicach określonych standardami wymagań egzaminacyjnych, jest głównym źródłem błędu pomiaru. Ujednolicenie zadań, to jest rezygnacja z wersji równoległych egzaminu, nie podnosi precyzji wnioskowania o populacji zachowań z tej próby, a błąd wyniku egzaminu pozostaje niepokojąco duży (Niemierko 2000b, 12–13).

Złudzenie 6. Specjaliści krajowi formułują zadania egzaminacyjne najlepiej mierzące osiągnięcia uczniów. Taka opinia podnosi autorytet władz centralnych i zatrudnionych przez nie ekspertów, ale nie potwierdza się w praktyce. By zbudować wysoko trafne zadanie egzaminacyjne, trzeba znać nie tylko program, lecz także przebieg kształcenia w określonym mikrosystemie. Analiza zadań dokonywana po egzaminie często wykazuje, że spisują się one dobrze w jednych, a zawodzą w innych szkołach, nie mówiąc o tym, że pojedynczy uczniowie miewają nietypowe układy osiągnięć (Niemierko 1990, 391–396). Zadania wytwarzane przez nauczycieli tych uczniów mogą okazać się dla nich najtrafniejsze.

Złudzenie 7. Standardy wymagań egzaminacyjnych dokładnie wyznaczają oczekiwany poziom osiągnięć uczniów. Kto przyjrzał się uważnie

² Błąd standardowy pomiaru może być oszacowany jako pierwiastek wariancji prawdziwej wyników otrzymanych, według wzoru: $s_e = s_x \sqrt{1 - r_{tt}}$, gdzie s_e jest błędem standardowym pomiaru, s_x — odchyleniem standardowym wyników otrzymanych, a r_{tt} jest współczynnikiem rzetelności pomiaru. Podstawiając do tego wzoru typowe wartości współczynnika rzetelności egzaminu zewnętrznego, $0,8 \leq r_{tt} \leq 0,9$, uzyskujemy oszacowanie błędu standardowego: 32–45% odchylenia standardowego wyników otrzymanych (Niemierko 1975, 240–241). To znaczy, że przy zastosowaniu skali 0–30 punktów i odchyleniu standardowym wyników egzaminu $s_x = 6$ punktów przedział ufności pojedynczego wyniku egzaminu na poziomie ufności $p = 0,95$ wyniesie od niespełna 8 punktów do prawie 11 punktów! Z tych oszacowań nie powinniśmy wnioskować, że egzamin jest grą losową, a próby podniesienia jego rzetelności są daremne, uczą nas jednak skromności w ocenianiu wartości pojedynczej próby zachowań ucznia, choćby była ona ujednolicona w całym kraju.

naszym krajowym wydawnictwom, a tym bardziej — spróbował je zastosować do oceniania osiągnięć uczniów, niełatwo zgodzi się z tym twierdzeniem. Najściślejsze nawet opisy czynności uczniów nie są w stanie ustanowić normy pozwalającej na bezbłędne rozstrzygnięcie opanował — nie opanował na ich wyłącznej podstawie. Niezbędne okazuje się albo — na wzór angielski — ustąpienie standardów, albo — jak to się dzieje teraz w kraju — ich kolejna operacjonalizacja, to jest zaopatrzenie w szczegółowe schematy punktowania, aktualne wyłącznie dla pewnego materiału i wybranego zadania. To pierwsze rozwiązanie jest półśrodkiem, gdyż rozrabnia, a nie usuwa wątpliwości. To drugie jest radykalne, ale obciążone arbitralnym materializmem, a więc „sito może gubić diamenty”.

Złudzenie 8. Egzaminowanie doniosłe pozwala na obiektywną ocenę pracy nauczycieli i szkół. To złudzenie jest bodaj najtrudniejsze do wykorzenia ze świadomości administracji szkolnej, rodziców i opinii publicznej. Od lat wiadomo, że rankingi i „tabele ligowe” wyrządzają wiele zła szkołom pracującym w trudniejszych warunkach (Gipps, Stobart 1993, 91–92; Potworowski 2000). Nie skutkują teoretyczne modele ewaluacji z uwypukloną rolą kontekstu, to jest czynników osiągnięć ucznia niezależnych od nauczyciela i szkoły (Stufflebeam 1969), paragrafy kodeksów etycznych (Schmeiser i in. 1995, cz. 6) i standardy pomiarowe uchwalane przez pedagogów i psychologów (*Standards...* 1999, rozdz. 15). W oczach laików każdy egzamin zewnętrzny odkrywa „całą prawdę” o szkole i basta!

W SZKOLNYM MATECZNIKU

Jedno z najciekawszych, a zbyt mało spopularyzowanych badań ostatniego dziesięciolecia nad nauczycielskim ocenianiem osiągnięć uczniów zostało wykonane przez Susan Brookhart z Uniwersytetu Duquesne w Pittsburgu (Brookhart 1993). Odpowiada ono na pytanie: dlaczego nauczyciele niekonsekwentnie stosują się do zaleceń formułowanych w podręcznikach pomiaru dydaktycznego, a w pewnych przypadkach jawnie od tych zaleceń odstępują?

Było to jedno z wielu badań zainspirowanych modelem rozszerzonej trafności teoretycznej, obejmującej nie tylko uzasadnienia pomiaru na podstawie zebranych dowodów, lecz także uzasadnienia na podstawie przewidywanych konsekwencji powziętych decyzji (Messick 1989; Niemierko 1999, 176–178). Brookhart przebadła 84 czynnych nauczycieli szkół podstawowych i średnich, stosując siedem „scenariuszy” na temat wysiłku i zdolności, brakującej pracy domowej oraz przyrostu osiągnięć³. Na tej podstawie sformułowała wnioski, z których dwa następujące są najważniejsze:

³ Oto treść i wyniki jednego ze scenariuszy Brookhart: „Jesteś nauczycielem algebry w szkole średniej o profilu akademickim. Stosujesz testy dwukrotnie w każdym semestrze. Dawid uzyskał stopień 1 (nieodstateczny) w pierwszym teście i dolną granicę stopnia 2 (dopuszczającego) w drugim teście. W tej sytuacji (A) postawisz Dawidowi ocenę semestralną 1 na podstawie jego średnich osiągnięć z dwu egzaminów, (B) postawisz Dawidowi ocenę semestralną 2, ponieważ wykazał przyrost osiągnięć”. 23 respondentów (29%) wybrało decyzję A, uzasadniając to głównie średnim wynikiem egzaminów, ale 56 respondentów (71%) wybrało decyzję B, uzasadniając to najczęściej konsekwencjami i pozytywnym z oceniania.

1. „Sens pojęcia «stopień szkolny» jest dla nauczycieli ściśle związany z zagadnieniem nakładu pracy ucznia. Nauczyciele używają zwrotów w rodzaju «praca, jaką wykonał(a)», by opisać stopnie i ich stawianie. [...] Stopnie są dla nauczycieli czymś, co uczniowie zarabiają — są wynagrodzeniem za wkład pracy na pewnym poziomie” (s. 139).

2. „Istnicją dwa standardy sprawiedliwości oceniania: uczeń przeciętny lub lepszy otrzymuje «na co zasłużył», podczas gdy uczeń słabszy od przeciętnego otrzymuje «szansę», jeżeli tylko da się to jakoś uzasadnić. [...] «Nie mogłabym źle ocenić ucznia, który się stara», ponieważ uczeń, który «pilnie pracuje», nie «zasługuje» na złą ocenę. [...] Zalecane procedury oceniania, sugerujące bezkompromisowość, mają ograniczoną wartość dla nauczycieli” (s. 140–141).

Tak więc kryterium motywacyjne odgrywa istotną rolę w amerykańskiej szkole ogólnokształcącej, a nauczyciele pragną zachęcić słabszych uczniów do wysiłku przez podnoszenie ocen. Brytyjczycy Paul Black i David William są także przekonani, że w „czarnej skrzynce” klasy szkolnej taryfa ulgowa dla uczniów napotyających trudności jest konieczna. „Najgorszy scenariusz to taki — piszą Black i William (1998, 12–13) — w którym pewni uczniowie mają niskie stopnie teraz, mieli niskie stopnie przedtem i spodziewają się niskich stopni w przyszłości. [...] Nie można oczekiwać, że uczniowie uwierzą w wartość zmian sposobu uczenia się, jeśli nie doświadczą korzyści ze zmiany”. Również i w tym przypadku trafność oceniania szkolnego jest wzbogacona o kryteria pozaprogramowe i nieprzedmiotowe — wychowawcze i perspektywiczno-rozwojowe.

W badaniach przeprowadzonych w naszym kraju przez Zdzisława Mikołajczaka (1985) nauczyciele fizyki określili udział wymagań programowych na 10 do 90%, ze średnią 61%, wśród ogółu kryteriów bieżącego oceniania osiągnięć uczniów. Inne kryteria to warunki uczenia się w domu i w szkole (13%), zdolności ucznia (12%), motywacja ucznia (7%) oraz polityka edukacyjna regiony i kraju.

Gdy ocenianie szkolne dokonywane wyłącznie według kryterium wymagań programowych nazwiemy ocenianiem dydaktycznym, a ocenianie dokonywane według wielu kryteriów nazwiemy ocenianiem społeczno-wychowawczym (Niemierko 1990, 25–33), to systematyczne porównanie wyników tych dwu procedur doprowadzi nas do spostrzeżenia charakterystycznej prawidłowości statystycznej: „Im większe jest społeczno-wychowawcze znaczenie oceniania [wewnętrznego], tym bardziej uzyskane stopnie szkolne przewyższają wyniki oceniania dydaktycznego” (tamże, 45–47). Ta prawidłowość, określona jako „pierwsza prawidłowość oceniania społeczno-wychowawczego”, powoduje nie tylko to, że duża część ocen stawiana jest w szkole „na wyrost”, w oczekiwaniu, że uczeń z czasem sprosta racjonalnym i oficjalnym wymaganiom, lecz także to, że oceny semestralne i roczne, a tym bardziej oceny na świadectwie ukończenia szkoły — jako bardziej znaczące społecznie i personalnie — są niemal zawsze wyższe niż oceny bieżące⁴.

„Taryfa ulgowa” oceniania może dotyczyć wielu, a nawet większości uczniów. Rozważając zagadnienie programów kompensacyjnych, Allan Ornstein i Francis Hunkins biorą pod uwagę istnienie kilkunastu grup uczniów „zagrożonych” lub „o specjalnych potrzebach

⁴ W eksperymencie nad ocenianiem wypracowań maturalnych z języka polskiego oceny dziesięciu wypracowań u dziesięciu sędziów wzrosły tylko o 0,1 stopnia (z 2,9 do 3,0) między maturą próbną a maturą regulaminową (majową), podczas gdy oceny dokonane przez nauczycieli tych uczniów wzrosły odpowiednio o cały stopień (z 2,7 do 3,7)! (Niemierko, Majkowska 1998, 37–38).

edukacyjnych”, co prowadzi do stwierdzenia, że „ostatecznie może wyjść na to, że około 90% wszystkich uczniów jest w jakiś sposób zagrożonych i wymaga od szkoły dodatkowej pomocy” (Ornstein, Hunkins 1998, 364)⁵. Następuje inflacja stopni szkolnych, gdyż także uczniowie o zwykłych potrzebach edukacyjnych domagają się „sprawiedliwości”, to jest ocen odpowiednio wyższych, a nauczyciel, który nie potrafi uzasadnić „reguł gry”, traci część autorytetu.

Przewaga funkcji motywacyjnej oceny szkolnej nad jej funkcją informacyjną prowadzi do ograniczeń jej znaczenia treściowego, to jest takiego powiązania z treścią kształcenia, by o uczniu uzyskującym tę ocenę można było wnioskować, jaki zakres treści opanował, a jakiego nie opanował (Niemierko 2000a, 85–86). Niemal całkowicie znika możliwość porównywania poziomu pracy szkół na podstawie ocen, jakie otrzymują: w nich uczniowie, co wyraża „druga prawidłowość oceniania społeczno-wychowawczego”: „Stopnie szkolne będące wynikiem oceniania społeczno-wychowawczego są rozłożone w każdym oddziale szkolnym w proporcjach charakterystycznych dla (a) szczebla szkoły i (b) grupy przedmiotów nauczania — bez względu na rzeczywisty poziom osiągnięć uczniów” (Niemierko 1990, 47–51).

Upraszczając, możemy powiedzieć, że zalety wychowawcze egzaminu powszedniego — aczkolwiek często przeceniane i nadużywane, ale niewątpliwe — okazują się jego wadami ze względu na porównania międzyszkolne. Szkoła skrywa swoje tajemnice przed społeczeństwem, władzami i instytucjami wyższego szczebla, a postęp uzyskiwany przez kodyfikację wewnątrzszkolnych systemów oceniania nie może być ani raptowny, ani równomierny.

KOMU POTRZEBNY JEST ZIMNY PRYSZNIC?

Wiemy już, że egzaminy doniosłe, oparte na ocenianiu dydaktycznym i pozbawione pozaprogramowych kryteriów oceniania, przynoszą gorsze wyniki niż egzaminy powszednie. Wskutek tego większość uczniów może zasadnie „bać się” egzaminów doniosłych, jako tych, które dla większości skończą się rozczarowaniem. Jak dowodzi psycholog Józef Kozielski, człowiek mało dojrzały jest „hubrystyczny”, to jest skłonny do akceptacji wyłącznie dodatnich opinii o sobie i do przeceniania swoich osiągnięć (Kozielski 1986).

Młodzież przeżywa tę zmianę kryteriów bardzo głęboko. Okres dorastania obfituje w próby uzyskania wewnątrzsterowności, to jest samodzielności w dążeniu do własnych celów, stosunki ja — otoczenie wydają się już jako poukładane, a tu nagle dotychczasowe zdobycze psychospołeczne stają się z dnia na dzień nieistotne i trzeba spojrzeć prawdzie w oczy.

⁵ Ornstein i Hunkins wymieniają następujące kategorie uczniów: uczniowie z ubogich rodzin, należący do mniejszości narodowych i rasowych, dzieci emigrantów mówiących innym językiem niż angielski, niepełnosprawni, urodzeni jako wcześniaki, ofiary znęcania się i braku opieki, uzależnieni, ciężarni, członkowie gangów i grup kultowych, dzieci rodziców bez ukończonych szkół lub nastoletnich, uczniowie chorujący na AIDS lub tacy, których rodzice chorują na AIDS, dzieci rodziców rozwiedzionych i uzależnionych od narkotyków lub alkoholu albo uwięzionych.

Wystrój sali, udział osób spoza kręgu dobrze znanych nauczycieli, środki przedsięwzięte dla zapewnienia samodzielności pracy ucznia — wszystko to robi wrażenie zmywy świata dorosłych przeciw dążeniom i uczuciom młodzieży. Co roku szerzą się zatem nadzieje, że matura zostanie złagodzona, a w każdym razie nie będzie w pełni zobiektywizowana.

Sposobem na uniknięcie szoku obiektywnego oceniania osiągnięć, to jest nagłego porzucenia kryteriów pozaprogramowych na rzecz oceniania wyłącznie według wymagań programowych, może być stosowanie skal różnicowych, opartych na porównywaniu wyników między uczniami, nie zaś na porównywaniu wyników z wymaganiami. Wtedy każdy niezerowy wynik surowy pomiaru można uznać za świadectwo osiągnięć, a mniej lub bardziej formalna interpretacja statystyczna tego wyniku — jako „procent opanowanej treści”, centyle lub staniny (Niemierko 1999, rozdz. XIV) — pozwoli zachować częściowo dobre samopoczucie nawet najsłabszym. Jest to wszakże tylko zmiana skali pomiarowej, nie zaś wzrost poziomu osiągnięć.

Niektórzy uczniowie zawsze będą próbowali podnieść swoje samopoczucie oraz wynik egzaminu przez korzystanie z niedozwolonych sposobów uzyskiwania informacji ze specjalnie przygotowanych tekstów oraz od innych egzaminowanych⁶. Egzamin doniosłe nie są dla nich, rzecz jasna, pierwszą okazją do nieetycznych zachowań, są jednak, co bardzo niewychowawcze, kolejną okazją. Z kolei wprowadzanie „policyjnych” form tępienia niedozwolonych form pomocy na egzaminie słabsi psychicznie uczniowie — a więc nie ci, którzy przodują w oszukiwaniu — przyplacają objawami nerwicowymi.

Egzamin zewnętrzny jest negatywnym doświadczeniem dla uczniów słabszych szkół, pracujących w trudniejszych warunkach środowiskowo-rodzinnych, w których przecież — jak stwierdza „druga prawidłowość oceniania społeczno-wychowawczego” — rozkłady ocen szkolnych nie różnią się zasadniczo od rozkładów ocen w szkołach pracujących w bardziej uprzywilejowanych warunkach. Jednakże wielu uczniów „dobrych” i „bardzo dobrych” w słabszych szkołach niższego szczebla potrafi — po słabym występie na egzaminie zewnętrznym — wrócić na pozycje zbliżone do poprzednio zajmowanych, a to zapewne dzięki temu, że kształcenie ma nie tylko aspekt poznawczy, lecz także emocjonalny, i wysokie aspiracje ucznia oraz jego charakter moralny potrafią przełamać trudności wynikające z wcześniejszych braków w wiadomościach i umiejętnościach⁷.

⁶ W badaniach ankietowych po osiemdziesięciu nauczycieli szkół podstawowych i średnich wskaźniki aprobaty twierdzeń „Obyło się bez podpowiadania” i „Obyło się bez ściągania” wyniosły 0,6 i 0,8 dla sprawdzianu kompetencji oraz 1,2 i 1,1 dla matury, w skali +2 („zdecydowanie tak”) — (-2) („zdecydowanie nie”), a w badaniach ankietowych uczniów te wszystkie wartości były wyraźnie niższe i rozkładały się dość symetrycznie między potwierdzenia i zaprzeczenia (Niemierko 1999, 190). Przedmiotem maturalnym wyraźnie górującym w prowokowaniu nieuczciwych form zdawania egzaminów jest matematyka, dla której nauczycielskie wskaźniki aprobaty cytowanych twierdzeń wyniosły tylko 0,8 i 0,9. Pełny optymizm co do samodzielności pracy uczniów na egzaminach zewnętrznych wykazuje tylko znaczna większość dyrektorów szkół (tamże).

⁷ Przed czterdziestu laty, gdy poziom nauczania w szkołach podstawowych wiejskich był wyraźnie niższy od poziomu nauczania w szkołach miejskich, przeprowadziłem analizę danych, która przyniosła następujące wyniki w skali wartości standardowych (\bar{x}): młodzież ze wsi miała wyższe oceny na świadectwach (+0,34), niższe na egzaminie wstępnym (-0,32) i znów wyższe na pierwsze półrocze w I klasie liceum pedagogicznego (+0,10), podczas gdy młodzież z miasta miała odpowiednio -0,44, +0,41 i -0,13, przy czym przewaga osiągnięć młodzieży ze wsi nad młodzieżą z miasta (decydującą się na liceum pedagogiczne) rosła w kolejnych klasach (Niemierko 1969, 87–89).

Jednym z argumentów przemawiających za wprowadzeniem systemu egzaminów zewnętrznych w Polsce było podwójne wyzwolenie szkolnictwa ogólnokształcącego z bezpośredniej zależności od Ministerstwa Edukacji Narodowej: programowe — poprzez zachętę do budowania programów autorskich — i administracyjne — poprzez zezwolenie na zakładanie szkół niepublicznych. Ministerstwo chciało zapewnić sobie — jak to określił Krzysztof Konarzewski — „kontrolę parametryczną” wszystkich szkół i nie dopuścić do nadużyć swobody. Dlatego właśnie sięgnięto po wzory brytyjskie, które wydawały się gwarantować sprawne kierowanie edukacją przez „ujednoczenie kryteriów oceniania” (Marquand 1993, 26). Dziś przeżywamy wątpliwość, czy to wzorowanie się nie poszło za daleko (Potworowski 2000; Konarzewski 2000).

O PEŁNĄ HUMANIZACJĘ EGZAMINÓW

Gdy humanizm określimy jako koncentrację na sprawach ludzkich, w tym zwłaszcza jako poszanowanie swobody, godności i rozwoju człowieka, które miałyby być w szkole uzyskane przez coraz głębsze poznawanie młodzieży, to działalność edukacyjna w zakresie egzaminowania szkolnego mogłaby być poprowadzona w następujących sześciu kierunkach:

1. Należy rozstać się z mitami o bezbłędności egzaminów powszednich i doniosłych. W ustroju demokratycznym te mity nieuchronnie obracają się przeciw edukacji, bo im większą pretensję do nieomyślności mają instytucje oświatowe, tym zacieklej dziennikarze tropią sensację polegającą na rażących odstępstwach od tych wzorów. Podobnie jak Korczak walczył o prawo dziecka do tego, „by było, czym jest”, tak pedagodzy powinni szanować naturę ewaluacji osiągnięć uczniów, a więc realistycznie oceniać egzaminy szkolne i uczciwie informować opinię publiczną o ich wartościach i ograniczeniach⁸. Reklama usług edukacyjnych w tym zakresie musi być wyjątkowo powściągliwa, a otwartość na krytykę — wyjątkowo szczerą.

2. Egzamin powszedni i egzamin doniosły powinny być dokładnie rozróżnione. Żaden z tych egzaminów nie jest „lepszy” lub „gorszy” i żaden nie stanowi kryterium oceny drugiego. Egzamin powszedni jest przesycony emocjonalno-motywacyjnym aspektem kształcenia, co stanowi zarówno o jego sile, jak i słabości. Może wydatnie pomagać nauczycielowi w kierowaniu uczeniem się, ale jego obiektywizm jest ograniczony — najwyraźniej w przypadkach niskich osiągnięć uczniów, które podczas takiego egzaminu są w dobrej wierze akceptowane. Egzamin doniosły jest jednokryterialny, gdyż ujawnia rzeczywisty, nie obciążony względami społeczno-wychowawczymi, poziom opanowania określonych czynności, ale im bardziej jest zobiektywizowany, tym dalej mu

⁸ Wiele postulatów z tego zakresu zawicrają *Standardy pomiaru pedagogicznego i psychologicznego* (Standards... 1999, rozdz. 13). Należą do nich: dokładne wyznaczenie procedury zastosowań pomiaru oraz interpretacji wyników pomiaru dydaktycznego, przestrzeganie przed zbyt daleko idącym wnioskowaniem, staranne dokumentowanie trafności pomiaru, budowanie i aktualizowanie norm krajowych i lokalnych, wystrzeganie się podejmowania decyzji na podstawie wyników jednego egzaminu, zapewnienie uczniom możliwości powtórzenia egzaminu, pełna świadomość zakresu błędów pomiaru, zapewnienie fachowej konsultacji decydemtom itp.

do sytuacji, w której uczeń ma możliwość swobodnie wykazać pełnię swoich osiągnięć w kształceniu.

3. Egzamin powszedni można ulepszyć przez częściowe upodobnienie go do egzaminu doniosłego. Takie zabiegi muszą mieć podwójny charakter: po pierwsze, należy precyzować i popularyzować realistyczne zasady oceniania wewnątrzszkolnego, obejmujące istotną rolę pozaprogramowych kryteriów oceniania, a po drugie, trzeba wprowadzać elementy pomiaru dydaktycznego do sprawdzania osiągnięć w toku lekcji. Obydwa zabiegi są trudne, gdyż w pewien sposób ograniczają nauczyciela. Ujawniając system ulg, nauczyciel traci możliwość stosowania niektórych strategii wychowawczych, może nadto wywołać u uczniów symulację różnorodnych trudności i wzrost rozszczeń, a w konsekwencji inflację ocen szkolnych. Pomiar dydaktyczny nie jest jeszcze przedmiotem kształcenia w uczelniach pedagogicznych, wymaga dodatkowych kursów i nakładu pracy, nadto rządzi się własnymi prawami, które trzeba we współpracy z uczniami szanować. Jednakże postęp w obydwu kierunkach niewątpliwie dokonuje się w Polsce, a niebezpieczeństwo polega raczej na tym, że może dokonywać się zbyt szybko, na zasadzie niszczącego „efektu zwrotnego egzaminu doniosłego” (Nitko 1998).

4. Egzamin doniosły można ulepszyć przez częściowe upodobnienie go do egzaminu powszedniego. To upodobnienie dotyczy zarówno przebiegu egzaminu, który nie powinien być pod żadnym względem zaskoczeniem dla uczniów, jak też — co wydaje się ważniejsze — interpretacji wyników. Rolą diagnostyki edukacyjnej jest wskazanie zmiennych kontekstowych oraz sposobu rejestracji ich wartości, pozwalających na docenienie warunków pracy i wkładu pracy uczniów i nauczycieli (Niemierko 2000b). Pozwoliłoby to na pewnego rodzaju komentarz do osiągnięć uczniów i systematyczne przeciwdziałanie uporczywym tendencjom do „rozliczania” szkół wyłącznie na podstawie wyników egzaminu zewnętrznego. Egzamin doniosły zyskałoby przez to pewne walory ewaluacji kształtującej, jednogłośnie popieranej przez pedagogów.

5. W każdym przypadku należy brać pod uwagę wyniki obydwu rodzajów egzaminu. Dokonano wielu analiz wykazujących, że wysokiej trafności prognostycznej nie gwarantują ani konkurs świadectw, ani egzamin wstępny do szkół wyższego szczebla, traktowane z osobna, natomiast wzrost trafności przynosi prawie zawsze połączenie informacji z tych dwu źródeł⁹. Jest wielkim ryzykiem zachęcanie szkół wyższego szczebla do rezygnacji z innych źródeł informacji o osiągnięciach uczniów niż

⁹ Badania krajowe, których wyniki publikowano głównie w „Życiu Szkoły Wyższej i Dydaktyce Szkoły Wyższej” czekają jeszcze na metaanalizę. Wyniki badań amerykańskich można, na podstawie różnych źródeł, podsumować następująco: wyniki punktowe testów uzdolnień akademickich (SAT) korelują ze średnimi wynikami egzaminów na I roku studiów różnych uczelni w wysokości $r = 0,5$, a stopnie ze świadectwa szkoły średniej korelują ze średnimi wynikami egzaminów na I roku studiów w wysokości $r = 0,6$. Mimo takiej przewagi egzaminów wewnętrznych nikt nie kwestionuje potrzeby stosowania egzaminów zewnętrznych dla kandydatów na prestiżowe uczelnie konkurencyjne (stanowiące zresztą nie więcej niż 15 % ogółu wyższych uczelni amerykańskich; reszta przyjmuje wszystkich chętnych i zdolnych zapłacić za studia), gdyż korelacja ocen na świadectwach i wyników testów uzdolnień wynosi tylko $r = 0,3$, co powoduje, że złożenie wartości dwu zmiennych niezależnych (korelacja wielokrotna z kryterium) sięga $R = 0,7$, czym Amerykanic muszą się już zadowolić. Stosuje się tam także odpowiednik naszych „miejsc rektorskich” w postaci ulg (kwot) dla mniejszości rasowych, wybitnych sportowców, dzieci weteranów wojennych itp.

egzamin doniosły — organizowany centralnie lub przez samą szkołę — bez uzyskania dowodów jego trafności, a dokładniej — bez dokonania rachunku strat, jakie wskutek tej rezygnacji poniesiemy. Stratą dla szkoły jest przyjęcie niewłaściwych kandydatów, a dla kandydata — niedostanie się do szkoły, w której mógłby osiągnąć sukces dzięki cechom emocjonalnym, wykazanym w roli ucznia, a nie objętym jednokryterialnym egzaminem doniosłym.

6. Zmiany systemowe w zakresie egzaminowania szkolnego wymagają intensywnych badań naukowych. Badania nad funkcjonowaniem systemów egzaminowania nie mogą być zastąpione „konsultacją środowiskową” na temat proponowanych rozwiązań ani nawet ewaluacją ogólnoedukacyjną, ciężącą ku metodom jakościowym, wizytacjom koleżeńskim, układom demokratycznym i przymiarkom-a-niemiarom (Wlazło 1999, 22). Pilnie potrzebne są solidne analizy wielozmiennowe z zastosowaniem metod statystycznych, wytyczające drogi optymalizacji procedur pedagogicznych, które nigdy wszakże nie będą bezbłędne ani samowystarczalne. Obecna konferencja jest, moim zdaniem, pierwszym — i to dość późnym — sygnałem potrzeby takich badań. Dla wielu pracowników komisji egzaminacyjnych byłby to krok następny: już potrafią planować testy i dokonywać szczegółowej analizy zadań, a teraz trzeba będzie nauczyć się głęboko interpretować wyniki egzaminu. Nie obejdzie się bez zacieśnienia współpracy z odpowiednimi placówkami naukowymi. Stanowczo potrzeba teraz naszej edukacji naukowego wsparcia działań w zakresie egzaminu doniosłego!

LITERATURA

- Black P., Wiliam D., 1998, *Inside the black box. Rising standards through classroom assessment*, London.
- Brookhart S.M., 1993, *Teachers grading practices: meaning and values*, „Journal of Educational Measurement” 2.
- Bruner J. S., 1964, *Proces kształcenia*, Warszawa.
- Cronbach L. J., Snow R. E., 1977, *Aptitudes and instructional methods: a handbook for research on interactions*, New York.
- Gipps C., Stobart G., 1993, *Assessment. A teachers guide to the issues*, London.
- Jensen A. J., 1980, *Bias in Mental Testing*, New York.
- Konarzewski K., 2000, *Dylematy oceniania osiągnięć szkolnych*, [w:] B. Niemierko (red.), *Pedagogika, pomiar dydaktyczny i ocenianie szkolne. Skrypt dla uczestników II Podypolomowego Studium Ewaluacji Dydaktycznej na Uniwersytecie Gdańskim*, Gdańsk.
- Kozielecki J., 1986, *Psychologiczna teoria samowiedzy*, Warszawa.
- Marquand J., 1993, *Studium wstępne krajowego systemu oceniania w polskim szkolnictwie ponadpodstawowym*, Warszawa.
- Messick S., 1989, *Validity*, [w:] R. L. Linn (red.), *Educational Measurement*, 3rd edition, Washington.
- Mikołajczak Z., 1985, *Czterostopniowa skala stopni szkolnych w opinii nauczycieli fizyki*, „Edukacja” 1.
- Niemierko B., 1969, *O powodzeniu nauczyciela w pracy dydaktyczno-wychowawczej*, Warszawa.
- Niemierko B., 1975, *Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe*, Warszawa.
- Niemierko B., 1990, *Pomiar sprawdzający w dydaktyce. Teoria i zastosowania*, Warszawa.
- Niemierko B., 1999, *Pomiar wyników kształcenia*, Warszawa.
- Niemierko B., 2000a, *Kształcenie według wymagań*, [w:] K. Kruszewski (red.), *Pedagogika w pokoju nauczycielskim*, Warszawa.

- Niemierko B., 2000b, *Jak utworzyć sieć ośrodków diagnozy osiągnięć uczniów szkół podstawowych i średnich?*, referat wygłoszony na konferencji „Diagnostyka edukacyjna jako dziedzina kompetencji pedagogicznych”, Szczecin.
- Niemierko B., 2000c, *Czy zadanie wyboru wielokrotnego nadaje się do diagnozowania procesów edukacyjnych?*, [w:] B. Niemierko, J. Mulawa (red.), *Diagnoza edukacyjna. Zadania wyboru wielokrotnego. Materiały III Ogólnopolskiej Konferencji Naukowej w Wałbrzychu*, Wałbrzych.
- Niemierko B., Majkowska I., 1998, *Co warta jest wielokryterialna ocena pisemnych prac maturalnych?*, „Edukacja” 3 (63).
- Nitko A. J., 1998, *Model egzaminów państwowych opartych na programie nauczania, sprawdzających i różnicujących, przeznaczonych do dyplomowania i selekcji uczniów*, [w:] B. Niemierko, E. Kowalik (red.), *Perspektywy diagnostyki edukacyjnej*, Gdańsk.
- Ornstein A. C., Hunkins F. P., 1998, *Program szkolny. Założenia, zasady, problematyka*, Warszawa.
- Potworowski J., 2000a, *Pierwiastek angielski w transformacji polskiej oświaty*, [w:] K. Kruszewski (red.), *Pedagogika w pokoju nauczycielskim*, Warszawa.
- Schmeiser C. B., Geisinger K. F., Johnson-Lewis S., Roeber E. D., Schafer W. D., 1995, *Code of professional responsibilities in educational measurement*, Washington.
- Standards... 1999 — *Standards for educational and psychological testing*, Washington.
- Wlazło S., 1999, *Mierzenie jakości pracy szkoły*, cz. 3, Wrocław.
- Wojciszke B., 2001, *Psychologia oceniania: mechanizmy, pułapki, środki zaradcze*, [w:] B. Niemierko, B. Wojciszke (red.), *Psychologia oceniania, standardy edukacyjne, zastosowania statystyki*, skrypt nr 1 dla słuchaczy III Podyplomowego Studium Ewaluacji Dydaktycznej, Gdańsk.