Bolesław NIEMIERKO
Uniwersytet Gdański

KU CZEMU ZMIERZAJĄ EGZAMINY SZKOLNE?

W swym referacie spróbuję wykorzystać teoretyczny model idealnego testu sprawdzającego do dyskusji właściwości egzaminów zewnętrznych, obecnych i przyszłych. Uczynię to w przeświadczeniu, że teoria pomiaru dydaktycznego jest właściwą podstawą analizy procesów edukacyjnych składających się na sprawdzanie i ocenianie osiągnięć uczniów w każdej formie.

Jestem przekonany, że system egzaminacyjny w Polsce rozwija się – aczkolwiek niekonsekwentnie i z wieloma wahaniami – w odpowiednim kierunku. Przyśpieszanie tego rozwoju byłoby ryzykowne, ale krytyczna refleksja nad jego strategiami wydaje się celowa i pożyteczna. Zwłaszcza teraz – w dobie korekt reformy edukacji.

Nie ma etapu rozwoju systemu edukacji, w którym nie warto byłoby zająć się wynikami kształcenia. Są jednak etapy szczególnie sprzyjające takim dyskusjom. Piąta (a wliczając dwa pionierskie spotkanie w Gdańsku i Elblągu – siódma) konferencja z cyklu Diagnostyka Edukacyjna przypada w okresie prób oczyszczenia systemu egzaminacyjnego ze zbyt pośpiesznie przyjmowanych założeń i decyzji organizacyjnych. W takim okresie odwołanie się do teorii pomiaru dydaktycznego może być szczególnie płodne.

Model teoretyczny, jakim się tu posłużę, został nakreślony w książce „Pomiar sprawdzający" w dydaktyce, a więc nie spodziewam się, by był nowy dla uczestników konferencji, w większości doskonale wyedukowanych na kursach i studiach podyplomowych pomiaru dydaktycznego. Mimo to przedstawię go dość dokładnie, bo jego właściwości wydają się zyskiwać zupełnie nowy walor w niezbyt odległej, a możliwej do przewidzenia przyszłości.

IDEALNY TEST SPRAWDZAJĄCY

Idealny test sprawdzający może być nakreślony czterema następującymi postulatami (Niemierko, 1990, s. 113):[1]

A. Test jest krótki, najlepiej złożony z jednego zadania.

B. Rzetelność testu jest zupełna.

C. Test dokładnie reprezentuje dany zakres programu kształcenia.

D. Skala wyników testowania odzwierciedla wymagania, jakie są faktycznie stawiane uczniom przez nauczycieli.

Te cztery postulaty zostały uporządkowane (inaczej niż w cytowanym źródle) od czysto konstrukcyjnego i najłatwiejszego do najgłębszego i najtrudniejszego, a właściwie – niemożliwego do spełnienia. Stanowią niezrównoważony model idealny, gdyż żadnego z nich nie da się spełnić do końca bez znacznego upośledzenia wszystkich pozostałych. Przedyskutujmy je po kolei, by zarysować możliwość optymalizacji układu.

A. Pojedyncze zadanie proste w roli testu spotyka się w szkole tylko w karnym i złośliwym odpytywaniu bieżącym: „Nawet tego nie umiesz? Siadaj! Niedostateczny". Test bardziej przyjazny uczniowi może być monotematyczny, w formie jednego (teoretycznego) zadania rozszerzonej odpowiedzi lub (praktycznej) próby pracy, co także oszczędza uczniom i nauczycielowi wędrówki po programie kształcenia, a dokładniej – po materiale kształcenia.

Aczkolwiek polonistyczny esej maturalny wydaje się być doskonałym przykładem testu monotematycznego, gdyż bywa zaawansowany pod względem teoretycznoliterackim, a może być traktowany jako próba warsztatu krytycznego, publicystycznego lub pisarskiego, to specjaliści pomiaru dydaktycznego dalecy są od uznania go za idealne narzędzie pomiaru. Jego słabości psychometryczne są trojakie:

1. Zależność od materiału kształcenia. Projektodawcy egzaminów dojrzałości pragną uwolnić ich wyniki od przewagi znajomości szczegółowego materiału nad ogólną kompetencją zdających, proponując im zwykle kilka tematów do wyboru. To jednak wcale nie cieszy teoretyków pomiaru. Dowodzą oni, że taki test traci reprezentatywność w wymiarze materiału kształcenia, gdyż dobór – z pozycji ucznia – losowy materiału egzaminacyjnego zostaje zamieniony na dobór celowy: uczeń wybiera najlepiej opanowany, wcześniej przewidywany temat.[2]

2. Zależność od punktującego wynik. Zmienność wyników punktowania między „sędziami" zadań rozszerzonej odpowiedzi jest bez porównania większa niż między punktującymi zadania jednoczynnościowe.[3] Tę zmienność można ograniczyć ćwicząc sędziów w sztuce punktowania zadań i doprowadzając ich do consensusu, ale takie działania budzą uzasadnione wątpliwości, bo do prawdy (merytorycznej, metodologicznej, artystycznej) nie dochodzi się przez głosowanie, a decydowanie, co uczeń

ma zawrzeć w autorskim, twórczym opracowaniu (np. jakie pojęcia, nazwiska, tytuły, opisy, uzasadnienia, oceny) – zbyt pochopnie podejmowane przez nasze komisje egzaminacyjne – prowadzi na manowce.[4]

3. Czasochłonność. Tylko w baśniach dzielny rycerz rozwiązujący „w mig" zagadkę dostaje rękę królewny i władzę nad krajem, a więc jednozadaniowy test jest krótki. Typowe prace maturalne rozciągają się na kilka godzin, a realne próby pracy na wiele dni. Co więcej, czasochłonność dotyczy także czynności obserwatorów, egzaminatorów, jurorów. Muszą oni poświęcić znaczną część swojej kariery zawodowej na tego rodzaju szlachetną działalność.

Czy to wszystko prowadzi do wniosku, że egzamin w postaci swoistego „majstersztyku" powinien być wyeliminowany? Broń Boże, nie. Efektem zwrotnym takiej decyzji byłaby deprecjacja umiejętności składających się na dany wyczyn. Ma on jednak ograniczoną wartość pomiarową i warto go wspomóc bardziej zobiektywizowanymi metodami lub może pozostawić głównie tam, gdzie autorytet sędziego ma większe znaczenie niż obiektywizm procedury – na najwyższym poziomie wymagań egzaminacyjnych, w konkursach i olimpiadach przedmiotowych, na autorskich egzaminach wewnętrznych w określonej szkole.

B. Rzetelność zupełna testu to zero wariancji błędu, a więc brak losowego błędu pomiaru. Taki test zapewniałby identyczny wynik każdej wersji przy każdej poprawnej organizacji pomiaru, bez względu na upływ czasu i dobór punktujących. Ponieważ te właściwości są nierealne, zastanówmy się raczej nad tym, w jakich warunkach stosunek wariancji błędu do wariancji całkowitej, czyli wariancji wyników testowania, może być zminimalizowany.

Za podstawę tego przeglądu przyjmiemy sześciopunktową, podręcznikową listę działań sprzyjających podnoszeniu rzetelności pomiaru (Niemierko 1999, s. 214–215):

1. Wydłużanie testu. Ta droga ograniczania błędu pomiaru jest bardzo skuteczna, bowiem – teoretycznie biorąc – wyniki prawdziwe kolejnych zadań, nieobciążone błędem, sumują się w teście arytmetycznie, a błędy pomiaru, jako losowe i różnoimiennie, algebraicznie, a więc ich względny udział w ogólnym wyniku testowania maleje. Barierą, jaka wkrótce da znać o sobie w tych działaniach, jest „zmęczenie materiału" w postaci wyczerpywania się pomysłowości konstruktora zadań, znużenia uczniów, wzrostu kosztów pomiaru, marnotrawstwa czasu.

2. Zwiększanie spójności testu na podstawie analizy zadań. Ta droga jest urokliwa dla „pomiarowca", bowiem łączy interpretacje jakościowe (stosowność i poprawność

dydaktyczna zadań) z ilościowymi (moc różnicująca, łatwość i frakcja opuszczeń). Wiemy jednak, że oczyszczając test z zadań słabo różnicujących, zwężamy zakres treści, jaki reprezentuje. Najbardziej spójny byłby test zawierający tylko jedno zadanie, powtarzane monotonnie...

3. Zwiększanie liczebności zespołu sędziów. Gdyby dwu sędziów punktowało wypracowanie maturalne, jego wyniki osiągnęłyby rzetelność jednogodzinnego testu polonistycznego, a gdyby zatrudnić dziesięciu sędziów – rzetelność dobrego testu matematycznego (Niemierko 1999, s. 201 i n.). Jest to wszakże najdroższy ze sposobów podnoszenia rzetelności testu, niedostępny – w zwykłych warunkach – nawet dla najzamożniejszych krajów.

4. Podniesienie jakości punktowania zadań przez ulepszanie kluczy i schematów. To pole wysiłków państwowych komisji egzaminacyjnych, będące też – jak już sygnalizowano – polem złudzeń i porażek. Zadania otwarte łatwo na nim tracą swój największy walor: dywergencyjność, czyli możliwość generowania nowych, cennych rozwiązań. Wartość punktową mieć będzie tylko to, co przewidziała dana komisja. Czy do tego dążymy?

5. Zwiększenie niezależności sytuacji dydaktycznej. W praktyce oznacza to wyeliminowanie zakłóceń i odstępstw od regulaminu oraz niedopuszczenie, by uczniowie korzystali z nieprzewidzianej pomocy zewnętrznej. Zaostrzanie rygorów ma, niestety, także skutki emocjonalne. Jeżeli uczniowie przywykli do terytorialnej bliskości kolegów, podpowiadania i ściągania, mogą znaleźć się podczas testowania, w zupełnie nowej, przerażającej sytuacji. Niepokój testowy, jako szczególna postać stresu egzaminacyjnego, może utrudnić skupienie myśli, upośledzić ich przebieg i obniżyć samokontrolę badanego.

6. Podniesienie atrakcyjności zadań i znaczenia wyniku dla ucznia. Czy ktoś z nas jest przeciw ciekawym zadaniom i dużemu zaangażowaniu ucznia? A jednak i tutaj napotkamy opory i bariery. Po pierwsze, uczniowie niełatwo potwierdzą, iż zadania testowe „są związane z ich życiem", mimo urozmaicania, „udziecinniania" i ozdabiania tych zadań przez zespoły autorów.[5] Po drugie, zadania ciekawe, jako nietypowe i zaskakujące, są trudne i słabiej powiązane z innymi zadaniami, a więc ich wkład w rzetelność pomiaru bywa iluzoryczny. Po trzecie wreszcie, egzamin doniosły, a więc taki, którego wynik mocno waży na karierze ucznia i absolwenta, wywołuje stres, który obniża jakość pracy jednych, a podnosi jakość pracy innych uczniów, a więc wynik egzaminu zależy od ich indywidualnej tolerancji stresu.

Wnioskiem, jaki można wysnuć z tej części rozważań, jest utrzymanie dążenia do wysokiej rzetelności w ryzach nie pozwalających na to, by zniszczyło ono wyższe walory pomiaru. Bardzo wysoka rzetelność (wykazywana współczynnikiem o wartości powyżej 0,80–0,90) potrzebna jest tylko wtedy, gdy potrzebna jest nam porównawcza diagnoza bardziej szczegółowych umiejętności (kompetencji), mierzonych równolegle (Niemierko 2000, s. 11–14).

C. Trafność programowa testu jest postulatem oczywistym dla pedagoga, ale nie każdy pedagog zdaje sobie sprawę, że demokratyzacja szkolnictwa silnie ją ogranicza. Swobody programowo-podręcznikowo-metodyczne, jakimi nasi nauczyciele obecnie się cieszą, wykluczają dokładne reprezentowanie programu kształcenia przez jakikolwiek test o szerokim zasięgu.[6]

Główne ograniczenia, jakie tu napotykamy, są następujące:

1. Jednostronność poznawcza pomiaru. Żaden test osiągnięć szkolnych nie reprezentuje w pełni emocjonalnych aspektów programu kształcenia, choć zadania otwarte i praktyczne są im niewątpliwie bliższe niż zadania zamknięte, nacechowane chłodną logiką. Owszem, w motywacji do wysiłku i kulturze obycia testowego można znaleźć pewne pierwiastki wychowawcze, ale nie jest ich więcej niż elementów poznawczych w spontanicznych wyrażeniach radości lub smutku. Epizodyczność uczuć w teście pisemnym stawia go pod tym względem niżej od interakcji nauczyciela z uczniem w toku tradycyjnego odpytywania.
2. Arbitralność testowych struktur pojęciowych. Testy mogą nieźle reprezentować kategorie celów poznawczych, materiał kształcenia i poziomy wymagań, ale zbyt słabo rejestrują indywidualne schematy poznawcze, które są głównym i najtrwalszym dorobkiem ucznia w pracy nad przedmiotem (Kozielecki 1979, s. 174 i n.). Można powiedzieć nawet, że testy łamią te struktury, narzucając, zwłaszcza w formie zamkniętej, systemy pojęć, które są albo dorobkiem wybranych szkół naukowych, albo własnym poglądem pedagoga, autora testu. Współcześnie rozumiany program kształcenia, obejmujący swą dziedzinę w sposób jak najmniej arbitralny, mógłby być lepiej reprezentowany przez zadania skonstruowane przez uczniów niż przez jakikolwiek zbiór zadań standaryzowanych.
3. Arbitralność analiz programowych. Autorom narzędzi pomiaru dydaktycznego, którzy zwykle dysponują planami testów oraz empirycznymi wskaźnikami ich jakości uzyskanymi w toku próbnych zastosowań, brak podobnych dokumentów programowych. Niewielu ekspertów zastosowało opublikowany inwentarz programowy (Niemierko

1998) lub podobne narzędzie, a jedną z przyczyn może być pracochłonność (około 10 godzin) tego przedsięwzięcia. Programy kształcenia recenzuje się w Polsce jak monografie naukowe, a czasem – jak powieści obyczajowe, nie interpretując ich systematycznie i nie dając podstawy analizom trafności programowej testów. By zachować sprawiedliwość, trzeba dodać, że autorzy testów też często nie widzą potrzeby wykorzystania specjalistów przedmiotu do analizy stosowności i poprawności dydaktycznej zadań oraz logicznych aspektów trafności i obiektywizmu pomiaru.

To jasne, że trafność programowa testu nie może być zupełna. Gorzej, że na ogół nie jesteśmy w stanie oszacować jej poziomu i musimy wierzyć autoreklamie konstruktorów testu, którzy tworzą szczegółowe kartoteki zadań na wyłącznie intuicyjnych podstawach.

D. Obiektywizm skali wyników testowania, rozumiany jako zgodność tej skali z wymaganiami stosowanymi w szkołach, pozostaje na razie w sferze życzeń i to podwójnie: ze względu na powszechne zaniedbania w tym zakresie oraz ze względu na wybujałą życzeniowość norm. Ponieważ jednak edukacja szkolna dość wyraźnie, choć powoli, zmierza w kierunku wieloaspektowej standaryzacji, warto wymienić powody, dla których skale wyników testowania mają tak niewiele wspólnego z wymaganiami stosowanymi przez nauczycieli w toku pozapomiarowego sprawdzania osiągnięć uczniów. Powody są cztery:

1. Rozmaitość poziomów wymagań. Nie darmo psychologowie ignorowali deklaracje nauczycieli, budując do swych testów osiągnięć skale różnicowe, to jest skale empiryczne oparte na rozkładach wyników prób standaryzacyjnych. Uważali, że nauczyciele za bardzo się różnią w projektowaniu norm i że brak im teoretycznych podstaw tych działań. Pojawienie się teorii pomiaru sprawdzającego (Glaser 1963) wpłynęło na poprawę sytuacji, ale wciąż dalecy jesteśmy od niezawodnych procedur.[7] Z tego powodu, a także z powodu niepopularności pracochłonnych analiz, w naszym kraju przeważają pseudonormy wymagań, oparte na naiwnych i nieuprawnionych założeniach (Niemierko 1999, s. 275–279),

2. Konieczność ustępstw na rzecz słabych uczniów. Według Susan Brookhart (1993, s. 140–141), „Istnieją 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". Od dawna wiadomo, że nauczyciele akceptują tylko takie

wyniki testowania, które nie burzą ich opinii o uczniach i nie udaremniają strategii kształcenia (Niemierko 1990, s. 387–388). Chronienie najsłabszych uczniów przed załamaniem i rezygnacją z wysiłku należy do tych strategii. Zaznacza się tu wyraźna różnica między wynikami testowania a ocenianiem szkolnym. Badania „minimum kompetencji" w USA dowiodły, że dolny próg wymagań w szkole nie istnieje.[8]

3. Niejawność procedur oceniania. W dawnej tradycji szkolnej mieliśmy tajemnice notesów nauczycieli, dowcipnie opisane przez Kornela Makuszyńskiego. W nowszej – mamy przypadki oceniania pozornie jawnego, w którym stopień jest zakomunikowany, ale jego pełne uzasadnienie pozostaje w świadomości, a czasem tylko podświadomości nauczyciela. Wewnątrzszkolne systemy oceniania rzadko wyjaśniają, czy dwaj uczniowie mogą otrzymać różne stopnie za dokładnie taki sam wynik pomiaru, ale uzyskany przy różnym wysiłku i w różnych okolicznościach. Nauczyciele mogą zatem nie być zainteresowani takimi procedurami zobiektywizowanego oceniania, które poważnie ograniczą ich pole manewru wychowawczego.

4. Brak weryfikacji procedur. Procedury niejawne nie mogą być zweryfikowane, ale i jawnym procedurom brak zwykle tego, co James Popham (1978) nazywa „osądem wspartym informacją". Nawykli do autorytarnych decyzji programowych i metodycznych, zbyt mocno ufamy swoim przewidywaniom i zbyt często musimy na gwałt szukać usprawiedliwień, gdy „rzeczywistość skrzeczy" w postaci dużej liczby negatywnych wyników. O złudzeniach „myślenia życzeniowego" w pomiarze dydaktycznym napisano już wiele (por. Niemierko 1990, s. 289–291). Tu wypada tylko przypomnieć, że w edukacji prawie zawsze konfrontacja z rzeczywistością prowadzi do obniżenia norm wymagań (tamże, s. 344–345).

Jak widać, idealny test sprawdzający jest niemożliwy do skonstruowania nie tylko ze względu na logiczną niespójność zbioru postulatów, lecz także ze względu na niecelowość lub niemożność dokładnego spełnienia żadnego z nich z osobna. I tak:

A. Użyteczny test jednozadaniowy okazuje się bardzo czasochłonny.

B. Test zupełnie rzetelny musiałby być całkowicie jednolity i nieskończenie długi.

C. Dokładna reprezentacja programu wykluczyłaby twórczość dydaktyczną.

D. Wymagania nauczycieli nie tworzą stałych norm osiągnięć uczniów.

Jaki jest więc pożytek z nierealnych postulatów? Po pierwsze, wskazują kierunek dążeń, bo test pozbawiony sprawności organizacyjnej, rzetelności, trafności i obiektywizmu pomiarowego byłby bezwartościowy, a wzrost jego przydatności postępuje w miarę zbliżania się do modelu idealnego. Po drugie, uczą nas wszystkich pokory,

o której tak łatwo zapominamy, przywdziewając togi sędziów osiągnięć uczniów.

Właściwości A (maksymalna zwięzłość) i B (zupełna rzetelność) – gdyby były realne – miałyby znaczenie tylko pomocnicze, właściwości C (trafność programowa) i D (obiektywizm skali) natomiast miałyby znaczenie zasadnicze dla wnioskowania o osiągnięciach uczniów. Dzięki technikom analitycznym, możemy zbliżyć się w testach standaryzowanych do zalet B i C, ale zalety A i D pojawiają się raczej w testach nauczycielskich.

IDEALNY EGZAMIN ZEWNĘTRZNY

Po surowej lekcji, jaką dała nam dyskusja testu sprawdzającego, projektując idealny egzamin zewnętrzny postaramy się powściągnąć cugle teoretyzowania, tak aby zbudować model bliższy rzeczywistości, a przede wszystkim – bliższy uczniowi. Oto jego założenia:

A. Egzamin jest maksymalnie sprawny i ekonomiczny.

B. Zadania egzaminacyjne są naturalnym sprawdzianem umiejętności.

C. Wynik egzaminu obrazuje zarówno odtwórcze, jak i twórcze osiągnięcia uczniów.

D. Egzamin bezwarunkowo wspiera indywidualny rozwój każdego ucznia.

Rozważymy te postulaty po kolei:

A. Walory organizacyjno-ekonomiczne egzaminu. Ich najprostszą miarą jest całkowity koszt egzaminowania jednego ucznia, obejmujący nie tylko budżet akcji pomiarowej, lecz także wykształcenie i utrzymanie personelu oraz wszelkie wydatki administracyjne ośrodka egzaminacyjnego. Cena „4 złote 50 groszy", proponowana szkołom przez Instytut Badań Kompetencji w Wałbrzychu[9], może być uważana za koszt tego rodzaju.

Czy ktoś próbował oszacować koszt egzaminowania jednego absolwenta szkoły podstawowej, absolwenta gimnazjum i maturzysty? Nie wydaje się to nazbyt skomplikowane, ale żadne tego rodzaju dane nie zostały dotychczas opublikowane. Wielka szkoda! Konkurencja między niezależnymi ośrodkami diagnostyki edukacyjnej oraz między takimi ośrodkami a państwowymi komisjami egzaminacyjnymi powinna doprowadzić do obniżenia kosztu egzaminowania przy zachowaniu, a nawet podniesieniu jego jakości.

Nad jakością egzaminów powinna czuwać Państwowa Komisja oraz Rada Naukowa, ale złożona nie z filozofów i rektorów uczelni, lecz z osób, których dziedziną badań jest pomiar dydaktyczny. Chodzi bowiem o ilość wartościowej informacji, jaką uczeń i szkoła mogą uzyskać za daną cenę, nie zaś tylko o obniżenie kosztu egzaminu i jego prestiż.

Dodatkowymi składnikami organizacyjno-ekonomicznej oceny egzaminu powinny być: czas ucznia i nauczyciela niezbędny do jego przeprowadzenia, łatwość komunikacji między zainteresowanymi stronami na temat znaczenia wyników (skale, formularze, komentarze, wzory wnioskowania), lokalne koszty organizacji (np. urządzenia i wyposażenia izb egzaminacyjnych). Im więcej swobody w tej ocenie zostawimy odpowiednim samorządom, tym lepiej.

B. „Ekologiczna" wartość egzaminu. Egzamin nie powinien zakłócać toku pracy dydaktyczno-wychowawczej szkoły. Chodzi tu nie tylko o jego administrację i większy lub mniejszy stres egzaminacyjny uczniów, lecz o jego wartość znacznie głębszą, wynikającą z czynności wykonywanych przez uczniów dla rozwiązania zadań.

Międzynarodowy światek egzaminacyjno-pomiarowy cierpi na istny zawrót głowy od pomysłów naturalnych, praktycznych, „autentycznych", „ekologicznych" zadań dla uczniów. Uczeń ma znaleźć się w możliwie nisko symulowanej sytuacji, pokonywać naturalne trudności poznawcze i emocjonalne, uzyskiwać wartościowy społecznie wytwór, odbierać egzamin jak przygodę lub ćwiczenie, a nie jako „dzień sądu i gniewu". Sprzyja temu testowanie z wyposażeniem w notatki, podręczniki, kalkulatory, komputer, magnetofon, pracownię przedmiotową oraz wszelkiego rodzaju infrastrukturę działalności. Zadania zamknięte od dawna nie mają dobrej prasy, a utrzymują się na egzaminach głównie z powodu sprawności diagnostycznej.[10]

To wszystko, rzecz jasna, kosztuje. Odpytywanie ustne i testy „papierowe" są dorobkiem ubogiej, a może tylko – oszczędnej szkoły. Całkowicie naturalny egzamin sięga w sferę fantazji: rejestracji mowy ojczystej na wzór „Big Brothera", mowy obcojęzycznej – za granicą, biologii – w ekosystemach, matematyki – w sklepie, w banku, w budownictwie, w pomiarach ziemi i nieboskłonu... Taki egzamin musiałby trwać długo, niemal tak długo, jak zajęcia szkolne. Bo czyż naturalna obserwacja osiągnięć ucznia nie towarzyszy wszelkim ćwiczeniom?

Ani od szkoły, ani – tym bardziej – od zewnętrznego ośrodka, nie można wymagać całkowicie naturalnych egzaminów. Jednak włączenie, choćby pojedynczych, zadań tego typu do testu egzaminacyjnego oraz dysponowanie jego w pełni naturalną wersją, jako kryterium trafności egzaminu „papierowego" lub jego luksusowej odmiany dla zamożnej szkoły, jest postulatem realnym w niedalekiej przyszłości.

C. Zrównoważenie taksonomiczne egzaminu. Chodzi tu o to, by egzamin zawierał – w rozsądnych proporcjach – zadania sprawdzające niższe kategorie celów oraz zadania mierzące twórcze rozwiązywanie problemów.

W dobie powszechnego zainteresowania umiejętnościami („kompetencjami") uczniów pojawiły się już głosy w obronie niższych kategorii celów poznawczych, przywracające konieczną równowagę (Konarzewski 2000a). Z drugiej strony wszakże, egzaminy w formie jaka została zaprojektowana w ramach reformy edukacji, dają uczniom bardzo niewiele okazji do twórczości, o czym świadczy wspomniana już tendencja do zaopatrywania testów w dokładne klucze punktowania zadań rozszerzonej odpowiedzi.

Mieszanie zadań zamkniętych z otwartymi prowadzi do nadmiernego wzrostu znaczenia umiejętności odczytywania instrukcji przez (poddanego stresowi) ucznia, a nie daje korzyści w postaci automatycznego sczytywania odpowiedzi (jak w testach złożonych wyłącznie ż zadań zamkniętych). Dodatkowe zadanie „olimpijskie" umożliwiłoby niektórym uczniom wykazanie się oryginalnym talentem i poszerzyłoby funkcję egzaminu o promowanie twórczości.

Poziom wymagań – lub lepiej: oczekiwań – wykraczających poza program, celujących, może wydawać się marginalnym dodatkiem do sumujących egzaminów końcowych na danym szczeblu szkoły. Ma on jednak niebagatelne znaczenie motywacyjne: zamienia egzamin z wydarzenia emocjonalnie negatywnego (przymus, obawa porażki) na pewien rodzaj „szansy na sukces", przynajmniej dla najlepszych uczniów, którzy zwykle kształtują opinię grupy.

Taki dodatek mocno podrożyłby egzamin, gdyż – jak już sygnalizowano – zadania twórcze powinny być oceniane przez twórczych nauczycieli, na przykład przez wybranych badaczy akademickich lub artystów. Skoro jednak rozważamy idealny egzamin zewnętrzny, to odnotowanie postulatu otwarcia na spontaniczną twórczość ucznia wydaje się konieczne.

Eksperymenty z uzupełnianiem egzaminów przez przeglądy teczek prac uczniów przynoszą mieszane wyniki ze względu na notorycznie znikomą rzetelność pomiaru oraz obciążenie nauczycieli dodatkowymi obowiązkami (Koretz 1998, Stecher 2000). Teczki okazują się natomiast bardzo przydatne do kształcenia nauczycieli w sprawdzaniu i ocenianiu prac uczniów (Klenowski 2000). Czy w kulturze środkowoeuropejskiej – mniej promującej samodzielność ucznia niż anglosaska – projekty wykonane poza szkołą odzwierciedlałyby autentyczną twórczość własną (a nie rodziców) ucznia, to trzeba by potwierdzić eksperymentalnie.

D. Egzamin jako wsparcie rozwoju. Każdy uczeń wymaga zachęt do wysiłku uczenia się, a słaby uczeń wymaga szczególnie wielu zachęt. Jak napisali Paul Black i Dylan

Wiliam (1998, s. 12–13), „Najgorszy scenariusz to taki, w którym uczniowie mają niskie stopnie, 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". Egzamin może uskrzydlić najlepszych, ale z pewnością nie wolno mu zniechęcić najsłabszych uczniów. Jak to zrobić? W projektach systemu egzaminacyjnego wyrzeczono się jakichkolwiek wymagań wobec absolwentów szkoły podstawowej i gimnazjum, bo nawet 0 punktów potraktowano jako spełnienie obowiązku odbycia egzaminu. Tym samym zamieniono sprawdzanie na różnicowanie osiągnięć uczniów. Czy jednak zajmujący ostatnie miejsca w szeregu otrzymują tak potrzebne im wsparcie?

Dobrym sposobem zachęty do kształcenia się, znacznie lepszym niż rezygnacja z określania wymagań, jest dostosowanie trudności zadań do kilku poziomów osiągnięć uczniów, a więc budowanie testów wielostopniowych. Zaawansowana organizacja wielostopniowego egzaminu to komputerowe testowanie sekwencyjne, w którym kolejne zadania są dobierane na podstawie wcześniejszych rozwiązań.[11] W ten sposób można zapewnić każdemu, w zasadzie, uczniowi przewagę sukcesów nad porażkami w rozwiązywaniu zadań.

Jeszcze większą rolę we współczesnym egzaminie odgrywa komentarz do jego wyniku. Wprawdzie niektóre badania dowodzą, że dla wielu uczniów liczy się tylko wynik egzaminu doniosłego, a nie komentarz,[12] lecz bez objaśnienia „ludzką mową" sensu egzaminu jako prezentacji pozytywnych wyników uczenia się, przebiegającego w określonych warunkach i przy określonych uzdolnieniach i zainteresowaniach danego ucznia, trudno uznać egzamin szkolny za metodę edukacji.

W kilku ośrodkach egzaminacyjnych próbowano zastosować stereotypowy komentarz, zredagowany – nieraz bardzo starannie – dla dużych grup uczniów, którzy znaleźli się w poszczególnych przedziałach skali wyników. Jego oddziaływanie było krótkotrwałe (uczniowie szybko porównali teksty i stwierdzili, że były przygotowane „z góry") i raczej powierzchowne. Tekst adresowany do jednego ucznia, jaki występuje w dojrzałej ocenie opisowej, przemawia silniej, ale musi opierać się na jego dobrej znajomości, nieosiągalnej dla zewnętrznej komisji. Może wprowadzić obyczaj ustnego komentowania wyników egzaminów zewnętrznych – poszczególnym uczniom (a nie klasie) i przez własnych wychowawców (na podstawie dostarczonych danych i własnych spostrzeżeń nauczyciela)?

Cztery postulaty współczesnego egzaminu zewnętrznego – (A) sprawność, (B)

naturalność, (C) twórczość i (D) motywowanie – można podzielić na dwie grupy. Dwa początkowe (A i B) mają charakter technologiczny, a więc właściwe wykorzystanie wyposażenia – w tym głównie urządzeń elektronicznych – może stworzyć odpowiednie warunki do ich spełnienia. W obu przypadkach te urządzenia posłużą zarówno do budowania sytuacji zadaniowej (zawsze w pewnym stopniu symulowanej), jak też do rejestracji i interpretacji czynności ucznia. W tym zakresie postęp będzie wkrótce, jak się przewiduje, lawinowy.

Druga para postulatów (C i D) ma charakter humanistyczny, co sygnalizuje ich doniosłość, lecz zarazem niemożność uzyskania bardzo szybkiego postępu. Dostrzeganie twórczości ucznia nie jest mocną stroną polskiej szkoły, a kwalifikacje nauczycieli i kandydatów na nauczycieli w tym zakresie pozostawiają sporo do życzenia.[13] Także kształcenie egzaminatorów, jakiego dokonano ostatnio w naszym kraju, było zbyt masowe, za mało selektywne i w innym kierunku zorientowane niż interpretacja nietypowych rozwiązań zadań egzaminacyjnych. Jeszcze więcej jest do zdziałania w zakresie takiego komunikowania uczniom wyników egzaminu zewnętrznego, by wszyscy byli pokrzepieni. Niekiedy bowiem odnosi się wrażenie, że pewne instancje nadzoru pedagogicznego i niektóre ośrodki polityczne czerpią siłę i argumentację z porażek, a nie z sukcesów naszych uczniów.

ROBOCZE WNIOSKI

Wnioski też będą cztery, co pozwoli utrzymać klarowność konstrukcji referatu, ale bardziej zwięzłe w uzasadnieniu, bo oparte na przedyskutowanych już postulatach:

1. Jesteśmy uczestnikami szybkiego postępu w zakresie egzaminowania szkolnego, a dalszy taki postęp wydaje się konieczny. Odnotowują go wszyscy fachowi obserwatorzy, w kraju i za granicą.[14] Dotyczy standardów, metod, analiz i sposobu komunikowania wyników oceniania. Nie powinniśmy zatrzymać rozwoju systemu egzaminacyjnego w Polsce, potrzebne są jednak racjonalizacje, usprawnienia, weryfikacje i konsultacje, jakich nie dość mieliśmy w gorączkowych latach startu reformy edukacji.
2. Każda lekcja teorii pomiaru dydaktycznego uczy nas pokory. Nie jesteśmy w stanie zapewnić idealnej bezstronności punktowania, zgodności wewnętrznej, trafności programowej i obiektywizmu wyników egzaminu. Mało tego, w miarę zbliżania się do wzoru, ponosimy rosnące wykładniczo koszty finansowe i społeczne pomiaru. Tylko powierzchownie przyuczeni działacze mogą głosić, że jakiś system egzaminacyjny jest lub będzie niezawodny. Warto jednak systematycznie podnosić jego walory, w czym modele teoretyczne istotnie pomagają.

3. Wolna konkurencja i badania naukowe powinny być motorami dalszego rozwoju systemu egzaminacyjnego w Polsce. Nasz system egzaminacyjny zabrnął w szkodliwą centralizację.[15] Jak wykazuje doświadczenie wielu krajów, konkurencja między ośrodkami niezależnymi finansowo od Ministerstwa Edukacji, miarkowana nadzorem fachowej Komisji Państwowej i Rady Naukowej, sprzyja postępowi organizacyjnemu. Bardzo ważne jest także wspomaganie tych ośrodków przez wyższe uczelnie oraz instytuty badawczo-rozwojowe.

4. Profesjonalizacja dziedziny egzaminowania szkolnego jest nieunikniona. Na początku reformy mógł wystarczyć zapał, wysłuchanie serii wykładów, jaka taka znajomość języka angielskiego i parę wizyt studyjnych na Zachodzie. Obecnie to za mało, by projektować postęp egzaminacyjny. Byłoby dobrze, gdyby niektóre nominacje polityczne z poprzedniego czterolecia zostały zastąpione konkurencją dorobku teoretycznego i organizacyjnego w dziedzinie pomiaru dydaktycznego i diagnostyki edukacyjnej.

Przypisy

1. To zestawienie jest pochodną listy pożądanych cech testu różnicującego, sporządzonej przez Artura Jensena (1980, s. 262 i n.), obejmującej następujące właściwości: 1) test mierzy jakąś ogólną lub znaczącą właściwość, 2) wariancja jego wyników w danej populacji jest maksymalna, 3) różnicowanie badanych osób jest maksymalne, 4) wyniki testowania tworzą skalę przedziałową, 5) rzetelność testu jest zupełna, 6) interkorelacja zadań jest zupełna, 7) test jest krótki, nie wymagający dłuższego czasu testowania, 8) jego wynik są odtwarzalne w sensie skalogramu Guttmana. Łatwo zauważyć, że tylko pierwsza z tych właściwości ma charakter jakościowy, a wszystkie pozostałe mogą być wyrażone ilościowo.

2. Każdy oryginalny temat wykazuje swoistość, przejawiający się niską korelacją między wynikami uczniów uzyskanymi przy zmianie tematu. Z tego powodu miary wewnętrznej zgodności (w tym alfa Cronbacha) wyników cząstkowych zadania monotematycznego nie nadają się na miary rzetelności pomiaru, bo przeceniają korelację między tematami, o jakiej chcielibyśmy wnioskować.

3. W badaniach nad punktowaniem wypracowań maturalnych systematyczne różnice między sędziami oszacowano na 4% wariancji, ale interakcja „uczniowie x sędziowie" sięgnęły aż 25% wariancji wyników. Odpowiednie wartości dla badań amerykańskich, w których występowali niżej wykwalifikowani sędziowie, wyniosły 7% i 18% (Niemierko 1999, s. 208). W innych badaniach wykazano, że w toku kilkudniowej pracy sędziowie istotnie podnoszą wymagania (Congdon i McQueen 2000). „Ideałem byliby sklonowani egzaminatorzy, w jednakowy sposób reagujący na prace maturzystów" – ironizuje Jan Potworowski (2000).

4. Są dowody na to, że twórczość ucznia dostrzega tylko twórczy nauczyciel, a nauczyciel o niskich lub pozaprzedmiotowych kwalifikacjach jest na ogół formalistą w ocenianiu (Jakubowicz 1991, Hamryszak 1994).

5. Takie są doświadczenia Okręgowej Komisji Egzaminacyjnej w Gdańsku.

6. Do takich konkluzji dochodzi wielu autorów obcych (Haertel 1999) i polskich (Poręba-Konopczyńska 2001).

7. Temu zagadnieniu poświęciłem spory fragment książki „Pomiar sprawdzający w dydaktyce" (1990, s. 332–358). Od czasu jej wydania postęp jest tu niewielki, a spory zaostrzyły się (Hambleton i in. 2000).

8. Podsumowanie problematyki pomiaru „minimum kompetencji" w USA przedstawiła Chris Pipho (1978).

9. Oferta Instytutu Badań Kompetencji z dnia 15 października 2001. Obejmuje ona test złożony z 30–48 zadań zamkniętych i 1–20 zadań otwartych oraz ankietę dla ucznia. Instytut podejmuje się przygotować i dostarczyć materiały, sprawdzić wyniki, przeanalizować je i dostarczyć dane z badań oraz indywidualne zaświadczenia.

10. Tym zagadnieniom poświęcona była III (V) Ogólnopolska Konferencja Diagnoza edukacyjna w Wałbrzychu (zob. Niemierko 2000; Konarzewski 2000b).

11. Wprowadzenie do zagadnień testowania sekwencyjnego przedstawiłem w książce „Pomiar sprawdzający w dydaktyce" (1990, s. 109–111 oraz 285–289).

12. W badaniach Ruth Butler, przeprowadzonych w Izraelu w latach 1987–1988, wprowadzenie komentarza do stopnia ani nie podniosło motywacji uczenia się, ani nie zmieniło orientacji uczniów na sukces egzaminacyjny, nie zaś na ulepszenie uczenia się (Knight 2000). Niewątpliwie potrzebne są dalsze, intensywne badania nad typami komentarzy dydaktycznych i ich znaczeniem dla uczniów.

13. Dwa interesujące studia na ten temat przedstawiła Anna Malenda (1994, 1998).

14. „Zadziwiająco wielki postęp został dokonany w zakresie oceniania szkolnego w zaskakująco krótkim czasie. Warto zastanowić się nad tym, jak wiele zmieniło się w tym zakresie" – tak zaczyna swój przeglądowy artykuł „Zakamarki oporu wobec rewolucji w ocenianiu" Gregory Cizek (2000, s. 16).

15. Jan Potworowski (2000) wyjaśnia to powierzchownym naśladowaniem angielskiego systemu oświatowego. [Zauważa, że] „Uzasadnienie polskich sprawdzianów stanowi dziwaczną mieszaninę dwóch sprzecznych zamierzeń. Jednym jest chęć zaprowadzenia centralnej kontroli na wzór Anglii, (...) wprowadzenia tego samego sprawdzianu dla uczniów w całym kraju. Drugim – ochota uczynienia ze szkół miejsca atrakcyjnego dla dzieci, szczęśliwego i pozbawionego napięć". (s. 56) [Jednakże] „Kopiowanie centralizmu angielskiego, nawet w najlepszych intencjach, to recepta na niekończący się zamęt w kraju, nad którym wciąż unoszą się wspomnienia totalitaryzmu, a demokracja jest tak młoda". (s. 61)

Literatura

1. Black P. i D., 1998, *Wiliam Inside the black box. Rising standards through classroom assessment*, King's College London, London.

2. Brookhart S.M., 1993, *Teachers' grading practices: meaning and values*, „Journal od Educational Measurement", nr 2.

3. Cizek G.J., 2000, *Pockets of resistance in the assessment revolution*. „Educational Measurement: Issues and Practice", nr 2.

4. Congdon P. J., McQueen J., 2000, *The stability of rater severity in largescale assessment programs*, „Journal of Educational Measurement", nr 2.

5. Glaser R., 1963, *Instructional technology and the measurement of learning outcomes*, „Educational Psychologist", s. 519–521.

6. Haertel E. H., 1999, *Validity arguments for highstakes testing: In search of the evidence*, „Educational Measurement: Issues and Practice", nr 4.

7. Hambleton R.K. et al., *A response to „Setting reasonable and useful standards" in the National Academy of Sciences „Grading the Nation's Report Card"*, „Educational Measurement: Issues and Practice" 2000 nr 2.

8. Hamryszak J,. *O sprawdzaniu i ocenianiu osiągnięć uczniów w zakresie plastyk,i* [w:] B. Niemierko (red.) Diagnostyka edukacyjna. Gdańsk 1994, Wyd. UG.
9. Jakubowicz S., 1991, *Poprawianie klasówki – sztuka czy rzemiosło?*, „Kwartalnik Pedagogiczny", nr 1.
10. Jensen A., 1980, *Bias in mental testing*, The Free Press, New York.
11. Klenowski V., 2000, *Portfolios: promoting teaching*, „Assessment in Education: principles, policy and practice" nr 2.
12. Knight P., 2000, *The value of a programmewide approach to assessment*, „Assessment in Higher Education", nr 3.
13. Konarzewski K., 2000 a, *W obronie encyklopedyzmu*, „Klocki Autonomiczne", nr 5.
14. Konarzewski K., *Miejsce testów wyboru w kulturze oświatowej* [w:] B. Niemierko i J. Mulawa (red.), Diagnoza edukacyjna. Zadania wyboru wielokrotnego, Wałbrzych 2000b, Instytut Badan Kompetencji.
15. Koretz D., 1998, *Largescale protfolio assessments in the US: evidence pertaining to the quality of measurement*, „Assessment in Education: principles, policy and practice", nr 3.
16. Kozielecki J., 1979, *Koncepcje poznawcze czlowieka*, Żak, Warszawa.
17. Malenda A., 1994, *Rozwiązywanie zadań matematycznych przez uczniów szkół podstawowych i przez studentów matematyki* [w:] B. Niemierko (red.) Diagnostyka edukacyjna, Wyd. UG, Gdańsk.
18. Malenda A., 1998, *Ocenianie rozwiązań zadań matematycznych z zakresu szkoły średniej przez studentów kierunków nauczycielskich matematyki* [w:] B. Niemierko i E. Kowalik, Perspektywy diagnostyki edukacyjnej, Wyd. UG, Gdańsk.
19. Niemierko B., 1990, *Pomiar sprawdzający w dydaktyce. Teoria i zastosowania*, PWN, Warszawa.
20. Niemierko B., 1998, *Inwentarz programowy* [w:] A. C. Ornstein i F .P. Hunkins, Program szkolny: założenia, zasady, problematyka, WSiP, Warszawa.
21. Niemierko B., 1999, *Pomiar wyników kształcenia*, WSiP, Warszawa.
22. Niemierko B., *Czy zadanie wyboru wielokrotnego nadaje się do diagnozowania procesów edukacyjnych?* [w:] B. Niemierko i J. Mulawa (red.) Diagnoza edukacyjna.
23. Zadania wyboru wielokrotnego, Wałbrzych 2000, Instytut Badan Kompetencji.
24. Pipho C. (red.), 1978, *Minimum competency testing*, Phi Delta Kappan. Special Issue.
25. Popham W. J., 1978, *Setting performance standards*, IOX, Los Angeles.
26. Poręba-Konopczyńska, 2001, *Moje refleksje nad ocenianiem na podstawie układu rzetelności i trafności wyników sprawdzania zaproponowanego przez Petera T. Knighta*, [w:] B. Niemierko i K. Szmigel (red.) Teoria i praktyka oceniania zewnętrznego, Pandit, Kraków.
27. Potworowski J., 2000, Pierwiastek angielski w transformacji polskiej oświaty [w:] K. Kruszewski (red.), Pedagogika w pokoju nauczycielskim. Warszawa, WSiP.
28. Stecher B., 1998, *The local benefits and burdens of largescale protfolio assessment*, „Assessment in Education: principles, policy and practice", nr 3.