Krzysztof Konarzewski
Instytut Psychologii PAN

Egzamin 2002: rzetelność i trafność testu matematyczno-przyrodniczego

Przeprowadzenie po raz pierwszy w historii polskiej edukacji zewnętrznych egzaminów śmiało można uznać za sukces organizacyjny. Trudniej ustalić, kto i czego się z nich dowiedział. Raport CKE (2002) i podobne opracowania (np. Herczyński, Herbst, 2002) na krótko pobudziły chaotyczne spory ideologiczne, ale nie zachęciły do przyjrzenia się samym narzędziom pomiaru, a przecież to one decydują o wartości poznawczej wyników. Ofertę przeprowadzenia stosownych analiz w skali kraju resort odrzucił. Na szczęście dzięki uprzejmości jednej z okręgowych komisji egzaminacyjnej uzyskałem dostęp do wyników 10 tys. anonimowych gimnazjalistów z 411 oddziałów 107 szkół zlokalizowanych w 82 gminach. Niniejszy tekst przedstawia rezultaty badania rzetelności i trafności testu matematyczno-przyrodniczego na tej próbce.

Rzetelność

Rzetelność testu, czyli udział wariancji wyniku prawdziwego w wariancji wyniku pomiaru, jest wysoka (tab. 1). Przed uznaniem tej sprawy za zamkniętą trzeba jednak zauważyć, że cały 50-pozycyjny test składa się z dwóch części. Pierwsza zawiera 25 zadań wyboru (w terminologii komisji – „zamkniętych"), druga – 11 zadań wymagających krótkiej odpowiedzi („otwartych"). 9 z tych 11 zadań rozbito na 2 lub 3 „czynności". Ponieważ każda czynność była reprezentowana przez osobną zmienną zero-jedynkową, z 9 zadań zrobiło się 23.

Tabela 1.
Parametry wyników testowania

Zbiorowość	Średnia	Odch. st.	α Cronbacha	Rozkład
Kraj	28,16	8,89	0,89	normalny
Próbka	28,31	8,67	0,90	normalny

Taką strategię trzeba stosować ostrożnie, by się nie spotkać z zarzutem sztucznego zwiększania długości, a zatem i rzetelności testu. Ma ona pełne uzasadnienie, jeśli czynności składowe zadania są logicznie niezależne od siebie. Można to poznać po tym, że każda daje się rozwinąć w osobne zadanie. Przykładem – zadanie, które wymaga wytyczenia kąpieliska o głębokości do 1,5 m w południowo-zachodniej części jeziora i podania największej głębokości na podstawie mapy poziomicowej. Uczeń dostaje po 1 punkcie za (1) zaznaczenie gdziekolwiek obszaru spełniającego warunek głębokości, (2) zaznaczenie jakiegokolwiek obszaru w południowo-zachodniej części jeziora i (3) odczytanie największej głębokości. Niezależność składowych potwierdza zarówno zróżnicowanie średnich (odpowiednio: 0,4; 0,55; 0,93), jak i niskie współczynniki korelacji (0,08–0,40).

W wielu zadaniach czynności składowe nie są jednak niezależne, lecz tworzą szereg kumulatywny: pierwsza czynność warunkuje drugą, druga – trzecią, tożsamą z odpowiedzią. Weźmy zadanie, które wymaga znalezienia długości drogi na podstawie informacji, że jej część dopełniająca 3/4 do całości jest o 8 km krótsza niż pozostała. Egzaminatorzy uznali, że najpierw trzeba (1) wprowadzić niewiadomą i oznaczyć części (1/4x, 3/4x), potem (2) ułożyć równanie (3/4x – 1/4x = 8), wreszcie (3) rozwiązać je. Analiza wyników wykazała, że poza tym szeregiem znalazło się 2,2% uczniów. Większość z nich (1,6%) dostała punkt za trzecią czynność, nie dostając go za pierwszą lub drugą. Jak to możliwe? Czy rozwiązali zadanie w pamięci? – jeśli tak, to zasłużyli na 3 punkty, a nie na 1. Czy ściągnęli gotową odpowiedź? – jeśli tak, to nie zasłużyli nawet na 1 punkt. A może to pomyłka egzaminatorów?

Tę i podobne zagadki wyjaśniła rozmowa z personelem komisji egzaminacyjnej. Egzaminatorzy odrzucają założenie szeregów kumulatywnych, ponieważ prowadziłoby do zasady: sprawdzaj wykonanie do pierwszego błędu i ignoruj pozytywne elementy, które wystąpiły w dalszej części pracy. Wszystkie wyodrębnione czynności uznają tedy za niezależne, wskutek czego można przyznać uczniowi punkt, nawet gdy udziela błędnej odpowiedzi. Np. uczeń mylnie przyjmuje, że dopełnieniem 3/4 do całości jest 1/3, układa równanie 3/4x – 1/3x = 8 i znajduje x = 19 km. Egzaminator uzna, że uczeń źle wykonał czynności (1) i (3), ale dobrze – czynność (2), więc przyzna mu odpowiednio 0, 1 i 0, razem 1 punkt. Gdyby uczeń doszedł do poprawnego rozwiązania swojego równania (19,2 km), dostałby odpowiednio 0, 1 i 1, razem 2 punkty.

Taka metoda oceniania zasługuje na uznanie – tym większe, że jest nadzwyczaj pracochłonna. Egzaminator musi zrozumieć bieg myśli ucznia; mechaniczne przyrównanie wykonania do wzoru nie wystarczy. Dobrze to służy rzetelności pomiaru, ponieważ zmniejsza wpływ pomyłek biorących się z przypadkowej nieuwagi, stresu egzaminacyjnego itp. Ma jednak także słabe strony.

Po pierwsze, pozbawia zmienne jednolitego znaczenia. Jest tak po części dlatego, że różni egzaminatorzy mogą z różną wnikliwością analizować uczniowskie wykonania, po części zaś – że schemat oceniania nie uwzględnia wszystkich możliwych

rozwiązań, gdyż pomysłowość ucznia często przekracza wyobraźnię dorosłego. Załóżmy, że uczeń rozwiąże przytoczone wyżej zadanie w jednym zdaniu:

> Różnica między względnymi miarami obu odcinków: 3/4 – 1/4 = 1/2 odpowiada 8 km, zatem cała droga to 16 km.

Przyjęta metoda oceniania wymaga, by uczeń rozwiązujący zadanie inaczej, niż przewiduje schemat, dostał maksymalną liczbę punktów. Egzaminator ustali zatem wartości 3 zmiennych opisujących to wykonanie na 1, 1 i 1, mimo że uczeń nie wprowadził niewiadomej, nie ułożył równania i nie rozwiązał go. Na szczęście takich przypadków jest mało, ponieważ konstruktorzy opracowują schemat punktowania na podstawie próbki wykonania pobieranej w procesie standaryzacji.

Po drugie, metoda zliczania „kwantów wiedzy", czyli niezależnego punktowania elementarnych czynności, prowadzi czasem do wątpliwych równoważności. Weźmy zadanie, w którym trzeba porównać pola powierzchni bocznej czapeczek w kształcie ostrosłupa i stożka[1]. Schemat punktowania rozkłada je na 3 niezależne czynności: (1) obliczenie powierzchni ostrosłupa i (2) obliczenie powierzchni stożka, (3) porównanie obu wielkości. Uczeń może dostać 1 punkt za poprawny sąd a > b, nawet jeśli a i b nie są poprawnymi miarami powierzchni. Tyle samo dostanie uczeń, który poprawnie obliczył jedną z powierzchni. Czy ich wiedza jest taka sama?

Podobnie z dwoma innymi zadaniami, które wymagają zidentyfikowania narządu lub tkanki na podstawie rysunku lub opisu oraz podania ich funkcji. Trudno uznać, że są to 2 niezależne czynności. Funkcja jest jednym z elementów pojęcia czyli treści nazwy. Od biedy można nagrodzić 1 punktem ucznia, który zapomniał nazwy (skrzela), ale wie, do czego służy narysowany narząd (6% przypadków). Z pewnością jednak jego wiedza nie jest równoważna znajomości nazwy bez funkcji (2%).

Doszukiwanie się wartościowych elementów w wykonaniu zadania można doprowadzić do absurdu, np. uznać, że uczeń, który w obliczeniu objętości stożka pomija czynnik 1/3, coś jednak potrafi – mianowicie pomnożyć przez siebie 3,14, r^2 i h – mimo że bezmyślnie utożsamia stożek z walcem. Eric M. Rogers (1986: 98) zwraca uwagę, że niektóre błędy są jak zdechła mysz znaleziona w słoiku marynaty – dyskwalifikują całe rozwiązanie.

Trzeba jednak przyznać, że nierównoważność „kwantów wiedzy" jest znacznie mniej groźna w przetwarzaniu testu gimnazjalnego, który składa się wyłącznie z zadań odtwórczych, niż byłaby na egzaminie maturalnym. Obie wady nie przekreślają więc zalet omawianej metody oceniania, a wysoka rzetelność całego testu potwierdza tę opinię.

---

[1] W to zadanie wbudowano pułapkę: należało się powstrzymać od obliczenia pól podstawy obu brył. Najwyraźniej zbyt wielu uczniów w nią wpadło, toteż egzaminatorzy po prostu zignorowali ten warunek zadania. Taki rodzaj „twórczego oceniania" zanadto przypomina „twórczą księgowość", by wzbudzić uznanie.

Trafność

Rzetelność można zapewnić w sposób czysto empiryczny, zastępując zadania słabo skorelowane z wynikiem – silnie skorelowanymi. Nie wynika stąd jednak żadne uzasadnienie, dlaczego właśnie ten, a nie inny zbiór zadań ma być podstawą decyzji selekcyjnych. Pytając o takie uzasadnienie, podejmujemy zagadnienie trafności testu.

W podręcznikach metodologii trafność chadza w parze z rzetelnością i stąd zapewne bierze swój początek przekonanie, że oba pojęcia odnoszą się do parametrów narzędzia, które można oszacować metodami statystycznymi. To przekonanie wzmacnia przekład angielskiego *validity* na skromną, techniczną „trafność". Ale angielskiemu słowu znacznie bliżej do naszej „prawomocności". Pytanie *Is it valid?* może się odnosić zarówno do testu inteligencji, jak i czyjegoś małżeństwa: w obu przypadkach idzie o to, czy wynik pewnych czynności spełnia warunki uznania. W odniesieniu do testu inteligencji te warunki są zarysowane mgliście, toteż odpowiedź uciera się w dyskursie polemicznym i nigdy nie jest ostateczna; to odróżnia ją od poprawnie oszacowanej rzetelności. W odniesieniu do testu osiągnięć szkolnych rozziew między uzgadnianiem trafności a ustalaniem rzetelności jest jeszcze większy.

Po czym można poznać, że test osiągnięć jest trafny? Jedyny sposób to wykazać, iż rzetelnie mierzy dyspozycje, które wchodzą w skład pojęcia wykształcenia (Konarzewski, 1999). Samo to pojęcie też wymaga uprawomocnienia, a o to niełatwo, zwłaszcza gdy występuje w postaci zbioru abstrakcyjnych określeń. Na szczęście konstruktorowi testu wystarczy pojęcie uprawomocnione praktycznie, czyli zawarte w realizowanym programie kształcenia. Zadaniem konstruktora jest stwierdzić, czy młodzież nauczyła się tego, co zamierzył twórca programu, a nie dociekać słuszności tych zamierzeń.

Jak jednak badać trafność testu osiągnięć – zawsze zrelatywizowaną do programu kształcenia – gdy szkoły używają różnych programów? Jedyny sposób to rozwinąć wspólne im pojęcie wykształcenia. W zreformowanym szkolnictwie polskim zrobiono to w podstawie programowej, będącej załącznikiem do rozporządzenia ministra oświaty. Ustawa narzuciła ją twórcom programów i autorom podręczników, dlatego uznaje się, iż rzeczywiście jest zawarta w programach i podręcznikach. Do niej musi się zwrócić konstruktor krajowego testu osiągnięć. Nie jest to jednak proste. Nie można układać zadań testowych z podstawą w ręku, ponieważ nie jest ona programem – nie zawiera szczegółowych zamierzeń dydaktycznych. Potrzebne jest ogniwo pośrednie. Nazwano je „standardami wymagań".

Nazwa jest pokrętna (wystarczyłyby same „wymagania"), ale intencja prosta: idzie o „uszczegółowienie wymaganych od ucznia osiągnięć zawartych w podstawie programowej" (CKE, 1999: 28). Podstawa programowa kształcenia ogólnego dla szkół podstawowych i gimnazjów wylicza 8 „umiejętności", których uczeń powinien nabyć – np. „poszukiwania, porządkowania i wykorzystywania informacji z różnych źródeł oraz efektywnego posługiwania się technologią informacyjną"; prócz tego określa „osiągnięcia" w każdym przedmiocie z osobna. W matematyce jest ich 5 (np. „przeprowadzanie nieskomplikowanych rozumowań matematycznych"), w fizyce – 4

(np. „umiejętność obserwowania i opisywania zjawisk fizycznych i astronomicznych"), w chemii – 6 (np. „określanie właściwości różnorodnych substancji oraz ich powiązanie z zastosowaniem i wpływem na środowisko naturalne") itd.

Na czym może polegać uszczegółowienie tych osiągnięć? Na rozwinięciu ogólnych określeń – wskazaniu rozumowań matematycznych, które uczeń powinien umieć przeprowadzić, wyliczeniu zjawisk fizycznych, które miałby obserwować i opisywać itd. – oraz na określeniu wzoru wykonania wskazanych czynności. Ostatnie z cytowanych osiągnięć mogłoby być uszczegółowione w postaci zbioru standardów typu: „Wie, co to jest wapień (skała węglanowa), jak powstał (z nagromadzenia skorupek mięczaków) i do czego jest stosowany (do wyrobu wapna i nawożenia gleby)". Tak rozumiany standard jest **schematem zadania testowego**, który konstruktor testu może przekształcić w konkretne zadanie.

Z ostatniego zdania płynie ważny wniosek: konstruktor testu nie może być jednocześnie autorem standardów. Po pierwsze, standard jest czymś, co ogranicza konstruktora, instruuje go, jak budować test, żeby był w zgodzie z pojęciem wykształcenia zawartym w podstawie programowej. Konstruktor tworzący standardy ogranicza sam siebie. Po drugie, opracowanie standardów wymaga podejmowania ważnych decyzji o pojęciu kształcenia. W przykładowym standardzie ktoś musi zdecydować, że uczeń powinien coś wiedzieć o wapieniach, a także co uczeń powinien o nich wiedzieć (w przykładzie wymieniono 3 zagadnienia: skład chemiczny, geneza i zastosowanie, pominięto natomiast wiele innych, np. umiejscowienie złóż w Polsce) i ile powinien wiedzieć (w przykładzie wymieniono 2 zastosowania spośród wielu innych). Jest oczywiste, że takie decyzje powinno podejmować to samo ciało, które opracowuje podstawę programową i ponosi za nią odpowiedzialność. Powierzając je konstruktorowi testu, daje się mu pełną kontrolę nad procesem kształcenia w kraju.

Krytyka teoretyczna

Jak sobie poradzili pracownicy Centralnej Komisji Egzaminacyjnej z podwójną rolą autorów standardów i konstruktorów testów? Nie zamierzali wkraczać w kompetencje ministra oświaty, więc zamiast uszczegóławiać osiągnięcia, uogólnili je, czyli sprowadzili do zaledwie 4 standardów. Same zaś standardy zdefiniowali w sposób niewiążący, tak by mogły uprawomocnić każdy test osiągnięć.

Oto lista standardów:

1. Umiejętne stosowanie terminów, pojęć i procedur z zakresu przedmiotów matematyczno-przyrodniczych niezbędnych w praktyce życiowej i dalszym kształceniu.
2. Wyszukiwanie i stosowanie informacji.
3. Wskazywanie i opisywanie faktów, związków i zależności, w szczególności przyczynowo-skutkowych, funkcjonalnych, przestrzennych i czasowych.
4. Stosowanie zintegrowanej wiedzy i umiejętności do rozwiązywania problemów.

Już na pierwszy rzut wydaje się ona wadliwa. Pierwszy grzech to wielosłowie. Przydawka „umiejętne", podobnie jak fraza „niezbędne w praktyce życiowej i dalszym kształceniu" są najzupełniej zbędne, nikomu bowiem nie przyszłoby do głowy, że pożądana jest nieudolność lub stosowanie wiedzy niczemu niesłużącej. Zbędna jest też przydawka „zintegrowana", ponieważ stosowanie zintegrowanej wiedzy niczym się nie różni od stosowania wiedzy.

Drugi grzech to niejasność. „Procedury" to zapewne metody – jeśli tak, należało użyć tego słowa. Trudniej domyślić się, czym są „informacje": synonimem wszelkiej wiedzy, czy tylko jakiegoś jej fragmentu (np. faktów). „Stosowanie" nic nie znaczy, jeśli brakuje mu dopełnienia „do czego". Z komentarza dowiadujemy się, że idzie o stosowanie terminów i pojęć do czytania tekstów i opisywania zjawisk. Pierwsze z tych zastosowań sprowadza się do biernej znajomości terminów i pojęć (znać słowo to przynajmniej tyle co rozumieć komunikat, który je zawiera), ale wprost tego napisać nie można – zabrania ideologia „nowej dydaktyki". W drugim zastosowaniu odrywa się opisywanie od wyjaśniania, które zostaje zesłane do trzeciego standardu. Występuje tam zresztą dwukrotnie – jako „wykorzystywanie zasad i praw do objaśniania zjawisk" i „stosowanie zintegrowanej wiedzy do objaśniania zjawisk" – obok „opisywania przebiegu zjawisk". Nie mogąc nadążyć za epistemologią autorów, zaczynam podejrzewać, że to właśnie było jej głównym celem.

W rozmywaniu standardów ważną rolę grają szeregi budowane z elementów nierównorzędnych. Pisząc „wiedza i umiejętności", autorzy błędnie sugerują, że umiejętność nie jest wiedzą. Podobnie z szeregiem „fakty, związki i zależności". Faktem nazywa się rzeczywisty stosunek zachodzący między jakimiś rzeczami lub zdarzeniami, którego właściwym wyrazem jest zdanie (np. „Napoleon zwyciężył pod Austerlitz", „Atom wodoru ma jeden elektron walencyjny"), nie można go zatem odróżniać od związków i zależności. Czym różnią się od siebie związki i zależności to kolejna zagadka, w której nie pomaga dookreślenie „w szczególności" (można się tylko domyślać, że związki to stosunki przestrzenne i czasowe, a zależności – przyczynowe i funkcjonalne). Co w tym kontekście może znaczyć „wskazywanie" – pokazywanie palcem? Na czym miałoby polegać „opisywanie faktów" – na rejestrowaniu, przytaczaniu, wyjaśnianiu? Dalsze tego przykłady mamy w komentarzu do listy standardów. „Operuje informacją" zostaje rozwinięte w szereg „selekcjonuje, porównuje, analizuje, przetwarza, interpretuje, przedstawia, wykorzystuje", a „opracowuje wyniki" w szereg „ocenia, interpretuje, przedstawia". Czy autorzy wyobrażają sobie niezależne operacjonalizacje tych czynności?

Kuriozalnym przykładem niejasności jest drugi standard. Co znaczy „wyszukiwanie" informacji – poszukiwanie w różnych miejscach czy wydobywanie z danych? Według pierwszej interpretacji idzie o to, by uczeń, który czegoś nie wie, wiedział, gdzie można się o tym dowiedzieć (np. „Wyszukała definicję w *Słowniku psychologii*", „Wyszukał fotografię w Internecie"). Według drugiej – by uczeń potrafił tak przetworzyć dostarczone dane, żeby odpowiedziały na jego pytanie. To znaczenie

lepiej oddaje czasownik „wywnioskować" (np. „Z wykresu wywnioskował, że ankietę wypełniło 350 osób"). Wtedy też staje się jasne, że wyszukiwanie informacji sprawdza większość matematycznych zadań „z treścią", w których trzeba obliczyć czy wywnioskować niewiadomą wielkość z wielkości danych. Która interpretacja jest po myśli autorów? Z pewnością druga, bo w komentarzu mówią wprost o „odczytywaniu informacji" z tekstu, mapy itp. Skąd więc wieloznaczne „wyszukiwanie"? Stąd, że pierwszym znaczeniem nawiązuje do ideologicznej walki ze szkołą pamięci, ale tylko w drugim daje się przełożyć na zadania papierowego testu.

Jest i grzech trzeci – brak jakiejkolwiek myśli taksonomicznej. Wymagane osiągnięcia można dzielić na klasy rozłączne albo skumulowane. Przykładem pierwszej klasyfikacji jest podział na wiadomości i umiejętności traktowane jako podstawowe i równoważne składniki wykształcenia. Wiadomości można dalej dzielić wedle tego, do czego się odnoszą – do nazw (np. uczeń powinien wiedzieć, że ruch o stałej prędkości nazywa się jednostajnym), pojęć (krokodyl to gad zmiennocieplny, jajorodny, drapieżny) lub praw (na ciało pływające w wodzie działa siła wyporu równa sile ciężaru). Podobnie umiejętności: można je dzielić na formalne (uczeń powinien potrafić opisać warunki zadania w postaci układu równań) lub empiryczne (przeprowadzić reakcję zobojętniania). Przykładem klasyfikacji kumulatywnej jest taksonomia ABC Bolesława Niemierki (1990: 215 i nast.): tu druga i dalsze kategorie zawierają w sobie wszystkie poprzednie.

Podział CKE z pewnością nie jest rozłączny. Rozwiązywanie problemów zakłada znajomość pojęć, praw i metod, komunikowanie faktów zakłada znajomość nazw. Podział nie jest też kumulatywny. Czwarty standard mógłby sugerować związek ze „stosowaniem wiadomości", które wieńczy taksonomię ABC, gdyby nie to, że „stosowanie" występuje także w pierwszym i drugim. Czy z tego, że nie ma go w trzecim, należy wnioskować, że „wskazywanie i opisywanie faktów, związków i zależności" odpowiada Niemierkowskim „zapamiętaniu" i „zrozumieniu wiadomości"? Komentarz zdecydowanie temu przeczy. Standard drugi w znaczeniu słownikowym łamie taksonomię ABC, ponieważ mówi o wiadomościach i umiejętnościach (gdzie i jak szukać), które warunkują zastosowanie, ale tylko wtedy, gdy uczniowi brak niezbędnej wiedzy. Ten sam standard w znaczeniu przenośnym jest szczególnym przypadkiem czwartego.

Nie ulega wątpliwości, że tak chwiejna konstrukcja nie może być oparciem dla konstruktora testu, a to, czym może być, sprowadza się do uprawomocnienia gotowego wytworu *ex post*, przez przyporządkowanie standardów zadaniom, nie zaś zadań standardom. Pożytek z tego żaden, zwłaszcza, że reguły przyporządkowania są wysoce niejasne.

Najłatwiej odkryć regułę przyporządkowania czwartego standardu – zadanie powinno być trudne, mieć złożony tekst, wymagać reprezentacji symbolicznej (równań matematycznych lub chemicznych) i wielu obliczeń. Zapewne dlatego 10-zadaniowy subtest ma wysoką rzetelność (0,84). Reguła dla drugiego standardu też „zaleca się"

prostotą, ale niczym więcej. Przypisano go różnorodnym zadaniom, które łączy tylko to, że odwołują się do rysunku – wykresu, mapy, tabeli, schematu lub obrazka. Jedno zadanie wyłamuje się z tej reguły, ale pozornie – rysunek został w nim bowiem zastąpiony słownym opisem tego, co Robert zobaczył pod mikroskopem. W omawianym zbiorze znalazły się zadania wymagające wywnioskowania informacji z konwencjonalnie zaprezentowanych danych – np. o ilu mniej uczniów interesuje się kolarstwem niż informatyką (ze słupkowego rozkładu liczebności uczniowskich zainteresowań)[2] lub jaka jest największa głębokość jeziora (z mapy poziomicowej). W innych jednak trzeba nazwać to, co pokazuje obrazek, lub wskazać desygnat nazwy; o żadnym wnioskowaniu nie ma tu mowy, ponieważ odpowiedź nie zawiera się w danych. Przy okazji zwróćmy uwagę na dwa zadania o identycznej strukturze: trzeba w nich rozpoznać i nazwać narząd lub tkankę, a następnie podać ich funkcję. Znajomości funkcji tkanki przypisano drugi standard, ale narządu – trzeci. Rzetelność 15-zadaniowego subtestu jest niska (0,61).

Najtrudniej zrozumieć, czym kierowali się autorzy w przypisywaniu pierwszego i trzeciego standardu. W obu grupach znalazły się zadania wymagające układania prostych równań, znajomości nazw, pojęć i praw. Dziwić się temu nie należy, ponieważ definicje obu standardów są skrajnie niejasne i nierozłączne. Pierwszy, 10-zadaniowy, ma umiarkowaną rzetelność (0,78), ale trzeci, też 10-zadaniowy – niedopuszczalnie niską (0,54).

Krytyka empiryczna

Informacja o rzetelności subtestów przenosi krytykę na grunt empiryczny. Gdyby standard odnosił się do dobrze zdefiniowanego aspektu wykształcenia, to reprezentujące go zadania tworzyłyby subtest, którego rzetelność (z uwzględnieniem długości) byłaby nie mniejsza niż rzetelność całego testu. Ten warunek spełniają jedynie standardy pierwszy i czwarty. O dopasowaniu danych do całej klasyfikacji mówi konfirmacyjna analiza głównych składowych. Miara dopasowania okazała się dyskwalifikująca ($p_{\times 2} < 0{,}0001$), a lista sugerowanych zmian – niezwykle długa. Jaką zatem strukturę ma test? Na to pytanie odpowiada eksploracyjna analiza głównych składowych. Zażądałem rozwiązania z czterema składowymi (łącznie wyjaśniły 30% wariancji testu) i rotacją Varimax. Wyniki zawiera tabela 2. By poprawić jej czytelność, opuściłem ładunki bezwzględnie mniejsze od 0,3.

[2] Ciekawe, że pytaniu do tego samego wykresu o procent uczniów interesujących się pływaniem przypisano standard pierwszy. Najwyraźniej autorzy ograniczają „wyszukiwanie informacji" do dodawania i odejmowania.

Tabela 2.
Rotowana macierz czynnikowa zadań testowych

Nr	Treść zadania	Standard	Średnia	I	II	III	IV
1	2	3	4	5	6	7	8
Zadania krótkiej odpowiedzi: matematyka							
29B	Ułożenie równania	IV	0,33	0,78			
29C	Rozwiązanie równania	IV	0,29	0,77			
29A	Oznaczenie warunków zadania	IV	0,40	0,74			
26C	Obliczenie wysokości słupa wody	I	0,33	0,64			
33C	Porównanie powierzchni	IV	0,25	0,64			
33A	Obliczenie powierzchni ostrosłupa	IV	0,31	0,63			
33B	Obliczenie powierzchni stożka	IV	0,33	0,60			
26A	Obliczenie pola podstawy	I	0,54	0,59			0,33
32A	Obliczenie pola deltoidu w skali	I	0,43	0,56			
26B	Obliczenie objętości wypływającej wody	I	0,61	0,54			0,37
32B	Obliczenie pola rzeczywistego deltoidu	I	0,21	0,53			
Zadania wyboru							
20	Obliczenie średniej szybkości	I	0,83		0,55		
5	Obliczenie ceny znaczka	I	0,86		0,52		
19	Obliczenie temperatury	III	0,69		0,50		
3	Obliczenie procentu na podstawie wykresu	I	0,80		0,49		
23	Obliczenie proporcji „co n-ty"	I	0,92		0,44		
15	Obliczenie warunku zrównania kosztów	III	0,89		0,43		
25	Wybór właściwości atomu na podstawie rysunku	II	0,79		0,42		
16	Obliczenie sumy trzech półokręgów	I	0,46		0,41		
4	Obliczenie liczby znaczków	I	0,78		0,40		
10	Wybór substancji – źródła energii życiowej	I	0,71		0,40		
22	Wybór planety spełniającej dwa warunki	II	0,89		0,38		
17	Odczytanie różnicy wysokości z mapy poziomicowej	II	0,77		0,38		
24	Wybór równania	III	0,82		0,37		
18	Wybór kształtu góry na podstawie mapy poziomicowej	II	0,68		0,36		
9	Obliczenie czasu rozpoczęcia ruchu	IV	0,51		0,34		
21	Obliczenie liczby 3-elementowych ciągów binarnych	IV	0,42		0,32		
1	Obliczenie liczby przypadków z wykresu	II	0,96		0,30		
14	Wybór nazwy grupy COOH	I	0,77				
6	Wybór opisu krokodyla	I	0,81				
12	Wybór zginacza na podstawie rysunku	II	0,66				
11	Wybór rodzaju ruchu na podstawie wykresu	III	0,53				
8	Wybór liczby osi i środka symetrii figury	I	0,52				

1	2	3	4	5	6	7	8
Zadania krótkiej odpowiedzi: szkolna wiedza przyrodnicza							
31B	Narysowanie poziomych wektorów sił	III	0,12			0,58	
31A	Narysowanie pionowych wektorów sił	III	0,17			0,56	
35C	Rozwiązanie proporcji	IV	0,21	0,49		0,54	
35B	Ułożenie proporcji	IV	0,24	0,49		0,54	
35A	Zapisanie reakcji chemicznej	III	0,15	0,32		0,40	
36A	Podanie nazwy tkanki na podstawie opisu	II	0,24			0,46	
36B	Podanie funkcji tkanki	II	0,44			0,39	
7	Wybór pory roku na półkuli południowej	III	0,25			0,30	
13	Wybór nazwy drzewa na podstawie rysunku	II	0,21				
Zadania krótkiej odpowiedzi: praktyczne zastosowania wiedzy przyrodniczej							
28A	Podanie nazwy narządu na podstawie rysunku	II	0,92				0,48
28B	Podanie funkcji narządu	III	0,96				0,48
27	Podanie wyjaśnienia zjawiska na podstawie wykresu	II	0,70	0,31			0,44
30C	Podanie największej głębokości	II	0,93				0,43
30B	Wskazanie południowo-zachodniej części jeziora	II	0,55				0,42
34	Obliczenie liczby żarówek w obwodzie	III	0,72	0,33			0,39
30A	Wskazanie obszaru nie głębszego niż 1,5 m	II	0,40				0,33
2	Obliczenie różnicy na podstawie wykresu	II	0,98				

Wyniki analizy pozwalają w następujący sposób przedstawić strukturę testu:

Jedna składowa obejmuje wyłącznie zadania wyboru. Poza nią znalazły się tylko trzy zadania mające tę formę, z tego dwa (2 i 13) słabo związane z całym testem. Istnieje zatem pojedyncza dyspozycja odpowiedzialna za wykonanie zadań wyboru, całkowicie ignorująca podziały na przedmioty oraz standardy. Widzę w tym potwierdzenie hipotezy, że typowe testy wyboru mierzą co innego niż testy twórcze, a nawet testy krótkiej odpowiedzi (Konarzewski, 2000). Co mierzą? W dużej części zapewne **biegłość testową**: uważne czytanie treści zadań w celu znalezienia ukrytych wskazówek, domyślanie się odpowiedzi przez eliminowanie niedorzecznych opcji, zgadywanie

i zwykłe ściąganie. Wiadomo, że częste testowanie uczniów prowadzi do podwyższenia wyników testowania, w czym ideolodzy pomiaru widzą poprawę jakości kształcenia. W świetle omawianego wyniku możemy postawić hipotezę, że na tę poprawę składa się głównie wzrost biegłości testowej. Dodajmy, że zadania wyboru są łatwe (średnia 0,73), toteż składają się na 62% średniej krajowej. Gdyby zwyciężyli zwolennicy budowania testów wyłącznie z takich zadań, krzywa rosłaby z roku na rok, podziały społeczne niknęły w oczach i tylko wykształcenie kolejnych roczników pozostawiałoby coraz więcej do życzenia.

Pozostałe składowe obejmują zadania związane z przedmiotami. Jedna wyodrębnia wiedzę matematyczną (w odróżnieniu od biegłości rachunkowej), druga i trzecia – przyrodniczą: typowe zadania szkolne (np. rozkład sił, równanie reakcji chemicznej) oraz zadania zbliżone do problemów praktycznych (np. wyjaśnienie śnięcia ryb w akwarium, znalezienie liczby żarówek w obwodzie szeregowym). Składowe matematyczna i przyrodnicza-szkolna obejmują zadania trudne (odpowiednio – 0,37 i 0,23), które w największym stopniu reprezentują szkolne wykształcenie. Składowa przyrodnicza-praktyczna obejmuje zadania najłatwiejsze (0,77); prawdopodobnie reprezentuje wiedzę wywodzącą się tyleż ze szkoły, ile z życia codziennego.

Tabela 2 nie pozostawia żadnych wątpliwości, że empiryczna struktura testu ignoruje standardy CKE. W składowej matematycznej pierwszy i czwarty występują niemal w równej liczbie, w składowej przyrodniczej-szkolnej – drugi, trzeci i czwarty, a w składowej przyrodniczej-praktycznej – drugi i trzeci. Jeszcze raz się przekonujemy, że standardy nie mają wiele wspólnego ani z teorią dydaktyczną, ani z empirią.

Na koniec istotne zastrzeżenie. Pojedyncza dyspozycja, która okazała się kontrolować zadania wyboru, z pewnością nie jest ogólna: odnosi się do zadań wyboru znajdujących się w analizowanym teście, a nie do wszelkich zadań mających tę formę. Czy jest to osobliwość tego jednego testu i warunków testowania, czy stała właściwość testów osiągnięć konstruowanych i przeprowadzanych przez okręgowe komisje egzaminacyjne – pokaże przyszłość, jeśli tylko pogłębiona analiza wyników testowania stanie się dorocznym zwyczajem egzaminatorów lub ich mocodawców.

Bibliografia

CKE (1999), *Standardy wymagań egzaminacyjnych. Projekt*, Wydawnictwo CKE, Warszawa.

CKE (2002), *Egzamin gimnazjalny 2002. Sprawozdanie*, Wydawnictwo CKE, Warszawa.

Dolata R. (2002), *Procedury rekrutacji i dzielenia uczniów na oddziały w gimnazjach: ocena z perspektywy nierówności społecznych w edukacji*, [w:] E. Wosik (red.), *Zmiany w systemie oświaty. Badania empiryczne*, Wydawnictwo ISP, Warszawa.

Herczyński J., Herbst M. (2002), *Pierwsza odsłona. Społeczne i terytorialne zróżnicowanie wyników sprawdzianu szóstoklasistów i egzaminu gimnazjalnego przeprowadzonych wiosną 2002 roku*, Wydawnictwo Fundacji „Klub Obywatelski", Warszawa.

Konarzewski K. (1999), *Dylematy oceniania osiągnięć szkolnych*, „Kwartalnik Pedagogiczny", 2, s. 29–50.

Konarzewski K. (2000), *Miejsce testów wyboru w kulturze oświatowej*, [w:] B. Niemierko, J. Mulawa (red.), *Diagnoza edukacyjna: Zadania wyboru wielokrotnego*, Wydawnictwo IBK, Wałbrzych.

Konarzewski K. (2001), *Reforma nadzoru pedagogicznego*, [w:] K. Konarzewski (red.), *Szkolnictwo w pierwszym roku reformy systemu oświaty*, Wydawnictwo ISP, Warszawa.

Niemierko B. (1990), *Pomiar sprawdzający w dydaktyce. Teoria i zastosowania*, PWN, Warszawa.

Rogers E. M. (1986), *Doskonalenie nauczania fizyki poprzez konstrukcje i dyskusje różnych typów sprawdzianów*, Wydawnictwo IKN, Wrocław.