mgr **Klemens Stróżyński**
Wydawnictwo Szkolne PWN

STRATEGIA I PROCEDURA BUDOWANIA NORM ILOŚCIOWYCH DLA OCEN SZKOLNYCH

W tytule nie zostało użyte słowo „test". Można jednak przyjąć, że jeżeli przy okazji oceniania osiągnięć uczniów mówi się w ogóle o punktach, to chodzi w istocie o testowe sprawdzanie osiągnięć. Nawet jeżeli sami sprawdzający nie mają tej świadomości.

Wielu nauczycieli dla treściowego zdefiniowania stopni szkolnych w praktyce posługuje się pseudonormą mówiącą, że dla uzyskania oceny *dostateczny* potrzeba zdobycia około połowy punktów - albo wykazania około połowy oczekiwanych wiadomości. Jest to pseudonorma, w żaden sposób nie dająca się uzasadnić teoretycznie, zwłaszcza w odniesieniu do wiadomości i umiejętności, które tworzą skomplikowaną strukturę w umyśle ucznia. W odniesieniu do takiej struktury w ogóle nie wolno posługiwać się pojęciem procentów - może ono dotyczyć tylko zbiorów elementów jednorodnych. W odniesieniu do **punktów** pojęcie połowy z wszystkich może być - mimo nienaukowości - pewnym punktem wyjścia - tylko punktem wyjścia.

Jeśli chodzi o zakres punktów na poszczególne oceny, panuje sporo pomieszania. Częstym zwyczajem jest tworzenie skal (norm punktowych), odwzorowujących w jakimś stopniu rozkład Gaussa (tzw. rozkład normalny). Skale takie przewidują węższy zakres punktowy dla wyższych ocen, najszerszy dla oceny *dostateczny*. Także nie da się w żaden sposób uzasadnić stosowania rozkładu normalnego (opisującego zjawiska przyrodnicze) dla pomiaru umiejętności i wiedzy uczniów, które mają zupełnie inny charakter niż np. rozkład wzrostu czy wagi populacji ludzkiej.

Przykładowa skala, oparta na rozkładzie Gaussa, dla testu 60-punktowego, może wyglądać następująco: dla *bdb* - górne 5 pkt.; dla *db* - następne 10 pkt.; dla *dst* - kolejne 15 pkt.; dla oceny *dopuszczający* (dawn. *mierny*) - następne 10 pkt. Tak więc kto osiągnie więcej niż połowę punktów, otrzymuje ocenę *dostateczny* lub wyższą.

Stosowanie tego rodzaju skal powoduje - z czego mało kto zdaje sobie sprawę - **kreowanie rozkładu wyników**. Wyjaśnię to na przykładzie.

Wyobraźmy sobie grupę 100 uczniów, z których każdy w 100-punktowym teście uzyskał inny wynik (jeden 3 punkty, jeden 25, jeden 99, jeden 100 - itd.). Ustaliłem skalę taką, że *dst* odpowiada normie: 51-80, *db*: 81-95, *bdb*: 96-100. Otrzymam wtedy 30 ocen *dst*, 15 ocen *db*, 5 ocen *bdb*. To rozkład pierwszy, gaussowski. Jeżeli ustalę inną skalę, np. *dst*: 51-60, *db*: 61-80, *bdb*: 81-100, uzyskam następujący rozkład ocen: 10 ocen *dst*, 20 ocen *db* i 20 *bdb*. Zupełnie inny rozkład, zupełnie inne - na pozór - wyniki, jeśli spojrzeć na oceny szkolne. I w jednym, i w drugim przypadku rozkład ocen został wykreowany przez określoną normę punktową.

Chcę tu z całym przekonaniem podkreślić, że **wyniki naszych uczniów, o których tyle dyskutujemy, są w jakimś stopniu wykreowane** przez dość powszechnie przyjmowane podczas oceniania rozkłady gaussowskie - czy to istniejące w postaci norm punktowych, czy stosowane podświadomie, intuicyjnie. W jakim stopniu opisują rzeczywiście osiągnięcia naszych uczniów - tego nie wiem. I bodaj nikt nie wie.

Opieranie zaś norm punktowych na empirii, czyli na wynikach uzyskanych przez - powiedzmy, statystycznie wiarygodną grupę uczniów - też nie zdaje się trafne, ponieważ osiągnięcia uczniów powinny być odniesione do wymagań programowych. Ale wyniki empiryczne mogą być cenną pomocą przy ustalaniu i weryfikacji norm testowych, bo wiążą te normy z praktyką.

Jakieś strategie przy ustalaniu norm na oceny szkolne trzeba stosować. Jedną z takich strategii, którą polecam, jest zmniejszenie zakresu punktowego dla oceny *dopuszczający* (dawn. *mierny*), ponieważ jest to ocena **pozornie pozytywna**. Ci, co ją otrzymują, nie spełnili wymagań podstawowych, a ocena ta ma tylko dać szanse na uzyskanie w przyszłości stopnia w pełni pozytywnego.

Proponuję, dla świadomego ustalania norm testowych według wybranych strategii, przejść procedurę, którą zastosowałem w przypadku testu „Lekcja inna niż wszystkie". Był to dwudziestozadaniowy, punktowany zero-jedynkowo test rozumienia tekstu dla czternastolatków. Prezentował model warstwowo-liniowy, czyli zadania reprezentowały różne poziomy wymagań, ale podczas przeliczania punktów na ocenę nie było to istotne. Test miał dwie wersje dokładnie równoległe: wersję A z zadaniami czterokrotnego wyboru i wersję B z zadaniami krótkiej odpowiedzi.

Wstępnie ustaliłem „przymiarkowe" normy punktowe na poszczególne oceny, dla obu wersji testu jednakowe. Przyjąłem następującą strategię dla zaproponowania tych norm:

1. Aby nie kreować poprzez same normy gaussowskiego rozkładu wyników, stworzyłem cztery klasy o jednakowej liczbie punktów (3) dla ocen pozytywnych (bez *celujący*).
2. Przyjąłem, że dolna granica oceny *mierny* (od 1 września 1999 r. *dopuszczający*) leży nieco poniżej połowy możliwych do uzyskania punktów.
3. W wyniku tego oceny pozytywne zaczynały się od 9 punktów z 20, a klasy na poszczególne stopnie szkolne liczyły po trzy punkty.
4. Aby zmniejszyć liczbę ocen *mierny* (*dopuszczający*), które są ocenami „przejściowymi", ani w pełni pozytywnymi, ani negatywnymi, przesunąłem jeden punkt z klasy *mierny* do klasy *dostateczny*. W wyniku tego powstała następująca **wstępna tabela norm**:

bdb	18-20	(3 punkty)
db	15-17	(3 punkty)
dst	11-14	(4 punkty)
mrn	9-10	(2 punkty)

Wyniki punktowe obu wersji testu pokazały, iż wersja „A" (z zadaniami wielokrotnego wyboru) okazała się dla uczniów łatwiejsza niż wer-

sja „B" (z zadaniami krótkiej odpowiedzi). W grupie 12 szkół nazwanej ANNA średni wynik wersji „A" to 14,5 punktu, a wersji „B" - 13,2 punktu. W grupie 12 szkół nazwanej ADELA różnica była jeszcze większa: „A" - 14,8 pkt., „B" = 11,9.

Wyniki te potwierdziły założenia teorii testu, która w przypadku porównywania wyników wersji zadań otwartych i zamkniętych przewiduje stosowanie **norm skorygowanych dla zadań zamkniętych** (w związku z tzw. poprawką na zgadywanie). Normę tę wylicza się w sposób następujący:

Norma skor. = norma % KO + (100% - norma % KO) podzielone przez liczbę odp. do wyboru)

Przykładowo, w przypadku zastosowanego testu, wychodząc od wstępnie zbudowanej tabeli norm, na poziomie oceny *dst* (55% możliwych punktów, czyli 11), wyliczenie daje następujący efekt:

Norma skoryg. = 55% + (100% - 55%) : 4 = 55 + (45:4) = 66% [czyli 13 punktów]

Na poziomie oceny *db* (75% możliwych punktów, czyli 15), wyliczenie daje następujący efekt:

Norma skoryg. = 75% + (100% - 75%) : 4 = 75 + (25:4) = 81% [czyli 16 punktów]

Jak widać, norma skorygowana dla 20-punktowego testu zadań zamkniętych, w porównaniu z analogicznym testem zadań otwartych, powinna być podwyższona o jeden punkt na poziomie oceny *db*, a o dwa punkty na poziomie oceny *dst*.

Przyjmując szacunkowo wyliczone wyżej różnice między normą dla testu zadań otwartych i zamkniętych, zastosowałem następującą strategię:

1. Ponieważ wyniki w teście zadań zamkniętych były wysokie, podwyższyłem o jeden punkt dolną granicę oceny *dopuszczający (mierny)* w porównaniu z normą wyjściową, zachowując trzypunktowy zakres klas na pozostałe, wyższe oceny oraz węższy zakres punktowy oceny *mierny*. Tak więc normy dla **wersji wielokrotnego wyboru** są następujące:

bdb	18-20	(3 punkty)
db	15-17	(3 punkty)
dst	12-14	(3 punkty)
dop	10-11	(2 punkty)

2. Dla testu zadań otwartych, w porównaniu z nową normą dla zadań zamkniętych, o jeden punkt przesunąłem w dół dolną granicę ocen *bdb* i *db*, powiększając zakres punktowy oceny *bardzo dobry*, która i tak jest trudna do osiągnięcia. O dwa punkty (czyli **dodatkowo** o jeden) przesunąłem w dół dolną granicę ocen *dst* i *dop*, powiększając zakres punktowy oceny *dostateczny*, aby zachować węższy zakres oceny *dopuszczający*, z wyżej podanych względów. Tak więc normy dla **wersji krótkiej odpowiedzi** są następujące:

bdb	17-20	(4 punkty)

db 14-16 (3 punkty)
dst 10-13 (4 punkty)
dop 8- 9 (2 punkty)

Zastosowanie tak skorygowanych norm na poszczególne oceny dla obu wersji testu **umożliwia porównywanie ocen z tych wersji testu** oraz daje wystarczająco poprawne, czyli z grubsza zgodne z dotychczasowymi osiągnięciami uczniów rozkłady stopni szkolnych.

W powyższe normy wpisane są te strategie, które zostały świadomie przyjęte:

1. Strategia jednakowych klas punktowych dla poszczególnych ocen, aby nie kreować gaussowskiego rozkładu wyników; z zastrzeżeniem:
2. możliwości zwiększenia klasy dla oceny najwyższej, aby skompensować kumulowanie się w tej klasie pomyłek uczniowskich; oraz:
3. ograniczenia klasy dla oceny *dopuszczający*, aby zmniejszyć liczbę ocen **pozornie pozytywnych**.

Powyższe ustalenia wykorzystują zarówno elementy teorii testu, jak wyniki empiryczne uzyskane dla testu „Lekcja inna niż wszystkie". Mogą być wzorem do tworzenia analogicznych norm dla podobnych testów. Zastrzeżenia i obawy dotyczące tak dużego zakresu punktowego dla oceny *bdb* są bezzasadne. Liczba ocen *bardzo dobry* w poszczególnych testowanych klasach nie będzie nadmierna, zostało to setki razy sprawdzone podczas standaryzacji norm punktowych serii „Przyjazne testy" (Wydawnictwo Szkolne PWN), funkcjonującej od roku 1998 w matematyce i biologii w szkołach podstawowych. Natomiast takie unormowanie czyni ocenę najwyższą (w testach według tej koncepcji - *bardzo dobry*) łatwiejszą do osiągnięcia i motywuje uczniów do sięgania po wyższe wyniki.

Zamiast zakończenia

Podczas standaryzacji testu „Lekcja inna niż wszystkie" w dwóch grupach szkół i niejako przy okazji pracy nad budowaniem norm ilościowych testu, udało się sformułować pewną liczbę **wniosków dydaktycznych**, dotyczących treści sprawdzanych za pomocą testów polonistycznych. Prezentacja ich w całości jest niemożliwa ze względu na wymogi objętościowe tego doniesienia, pozwolę sobie na podanie tylko dwóch, bez cytowania zadań testowych, które dały podstawę do sformułowania tych wniosków.

1. W przypadku sprawdzania umiejętności uznanych na podstawie programu za trudne (wnioskowanie, uzasadnianie, odkrywanie, ocenianie) uczniowie osiągają wyniki wyższe niż oczekiwane, jeżeli sprawdzanie dokonuje się na materiale literackim bliskim i zrozumiałym dla uczniów. Może część narzekań na bezmyślność uczniów i brak dociekliwości ma źródło w niewłaściwym doborze materiału do sprawdzania tych umiejętności?

2. Zadania zamknięte bywają trafniejsze dla sprawdzania niektórych umiejętności polonistycznych niż otwarte dzięki temu, że podsuwają możliwości, których uczeń w zadaniu otwartym może po prostu nie wziąć pod uwagę. To, ile możliwości odpowiedzi uczeń bierze pod uwagę, jest mocno uzależnione od systemu dydaktycznego jego nauczyciela. W tym sensie zadania zamknięte mogą dawać równiejsze szanse uczniom w przypadku testów masowych.

Zapis z punktu 2 kieruję do tych, którzy tak pryncypialnie odżegnują się od testów zamkniętych. Zaś zapis z punktu 1 polecam tym, którzy układają testy polonistyczne i dobierają teksty takie, które **im** się podobają i posiadają, ich zdaniem, odpowiedni potencjał pomiarowy. Przypomnę tutaj powiedzenie profesora Bogusława Śliwerskiego, że przynętę powinna lubić ryba, a nie rybak.