Sławomir Sapanowski
Okręgowa Komisja Egzaminacyjna w Łodzi

Porównywanie wyników egzaminów – propozycja metody

Egzaminy zewnętrzne funkcjonujące w polskim systemie oświaty, przynoszą rokrocznie tysiące wyników osiągnięć uczniów na różnych poziomach kształcenia. Generuje to nieustającą potrzebę - wśród rzesz rodziców, nauczycieli i uczniów – tworzenia rankingów, zestawień i porównań. Niestety, ze względu na zmieniającą się łatwość arkuszy egzaminacyjnych jest to dla nieprzygotowanych do tego osób zadanie trudne, a nawet w wielu przypadkach prowadzące do błędnych wniosków.

W środowisku nauczycielskim dość dobrze zadomowiła się skala staninowa, która może służyć do porównywania wyników w kolejnych latach, ale obarczona jest jednak wieloma wadami (zbyt krótka, mało precyzyjna). W początkowym okresie funkcjonowania egzaminów dość dobrze spełniła swoje zadanie, uświadamiając wszystkim potrzebę spojrzenia na wynik surowy ucznia z innej (standaryzowanej) strony. Przyszedł jednakże czas na to, aby, posługując się kolokwializmem, powiedzieć: *staniny muszą odejść*.

Automatycznie powstaje pytanie: co zatem w zamian? Może któraś ze skal normalizujących, np. akademicka[1], a może centylowa? Jakie zalety i wady mają wymienione metody?[2]

W statystyce doskonale funkcjonuje sposób polegający na przekształceniu wyników surowych tak, aby średnia wynosiła zero, a odchylenie standardowe było równe jedności (tzw. wynik standardowy). W ten sposób każdy wynik ze skal pierwotnych (w przypadku sprawdzianu od 0 do 40) zostaje wyrażony w postaci wielkości odchylenia standardowego, o jaką jest oddalony od średniej na skali pierwotnej. Dokonuje się tego przekształcania według następującego wzoru:

$$z = \frac{x - \mu}{\sigma}$$

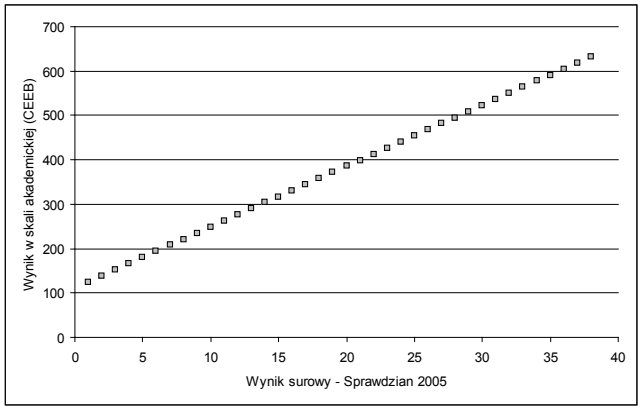
gdzie:
- x oznacza wynik surowy uzyskany na pierwotnej skali pomiarowej,
- μ oznacza wartość średnią wyników surowych w danej grupie,
- σ oznacza wartość odchylenia standardowego wyników surowych w danej grupie.

[1] Skala akademicka stosowana jest w badaniach PISA.

[2] W dalszej części artykułu skupiam się na wynikach sprawdzianu w szóstej klasie szkoły podstawowej, ponieważ różnice w łatwościach na tym egzaminie są największe i pewne własności skal normalizujących są doskonale widoczne. Dotyczy to zwłaszcza roku 2005 (sprawdzian najłatwiejszy) i roku 2009 (najtrudniejszy).

Uzyskane w ten sposób wartości wyników standaryzowanych przyjmują (najczęściej) postać ułamków o wartościach dodatnich lub ujemnych w zależności od tego, czy odchylają się w górę, czy w dół od wartości średniej. Ponieważ posługiwanie się ułamkami i wartościami ujemnymi przy operowaniu wynikami jest często niewygodne, można dokonać prostego liniowego przekształcenia wyników standaryzowanych na skalę o dowolnej wartości średniej i odchylenia standardowego. Dokonuje się tego, mnożąc każdy wynik standaryzowany przez wartość pożądanego odchylenia standardowego i dodając wartość pożądanej średniej. Tym sposobem otrzymujemy różne tzw. skale normalizacyjne:

Skala CEEB (akademicka) = z*100+500
Skala T = z*10+50
Skala IQ = z*15+100

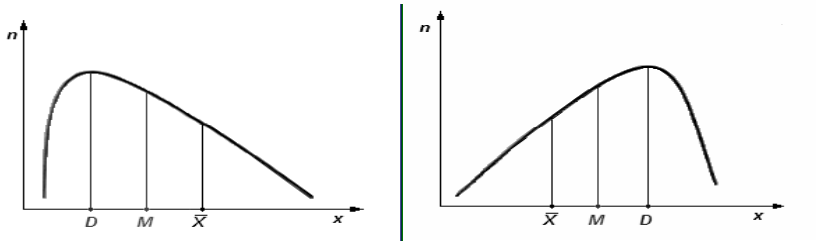


Wykres 1. Wynik sprawdzianu a skala akademicka

Wszystkie tego typu skale oparte są na przekształceniach liniowych, co powoduje, że zmiana wyniku surowego o 1 punkt skutkuje zmianą wyniku na skali znormalizowanej o stałą wartość. Ilustracją tego przypadku jest powyższy wykres (1), na którym przedstawiono zależność między wynikiem ze Sprawdzianu 2005 a znormalizowanym wynikiem w skali „akademickiej".

Powstaje pytanie: czy wiedza i umiejętności uczniów, którzy na sprawdzianie uzyskali wyniki 38 i 39 pkt. różnią się tak samo jak wiedza i umiejętności tych zdających, którzy uzyskali odpowiednio 20 i 21 punktów? Intuicja podpowiada, że raczej nie . Rozważmy sytuację ucznia, który, rozwiązując test, dotarł do momentu, w którym uzyskał 38 pkt. Do rozwiązania pozostały mu jeszcze dwa zadania, są to dwie szanse na uzyskanie dodatkowego punktu. Wykorzysta je lub nie. Inny uczeń w trakcie rozwiązywania arkusza ma już na swoim koncie 20 pkt i ma jeszcze przed sobą 20 zadań (szans) na uzyskanie kolejnego. Jest w „lepszym" położeniu. Rozumowanie to oczywiście jest uproszczone - zakłada, że uczniowie najpierw rozwiązują te problemy, z którymi sobie radzą, a trudne zostawiają na koniec. I często taka strategia postępowania jest stosowana przez uczniów.

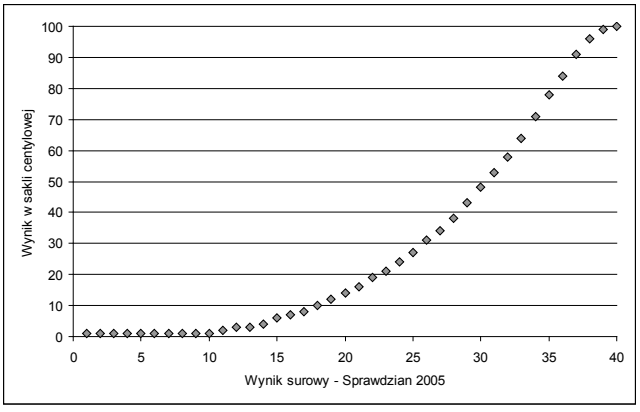
Ponadto skale oparte o wynik standardowy zakładają, że wynik średni z dwóch różnych testów jest równoważny. Przyjrzyjmy się zatem dwu przykładowym rozkładom różniącym się średnią i co ważniejsze skośnością (Czarnotta-Mączyńska i in., 2007).



Wykres 2. Rozkład prawoskośny i lewoskośny (D – dominanta, M – mediana, X – średnia)

Wynik średni w pierwszym rozkładzie (prawoskośnym) jest co najmniej dobry. Zdający uzyskał wynik lepszy od większości rówieśników. W drugim przypadku jest odwrotnie – jest to wynik gorszy. Zrównywanie tych wyników i komunikowanie, że są one równoważne jest chyba pomyłką. Wydaje się, że dużo lepszym parametrem pozwalającym porównywać efekty testowania jest mediana. Zarówno w jednym, jak i drugim rozkładzie uczeń, którego osiągnięcia równają się medianie, pozostawił w polu 50% zdających i dał się „wyprzedzić" przez 50%.

A może lepiej stosować skalę centylową? To lepsze rozwiązanie, ale ze względu na własności skali centylowej przekształcenie wyniku surowego na centyle owocuje zupełnie odwrotnym od oczekiwanego efektem. Uzyskanie dodatkowego punktu w środku skali powoduje dużo większy „skok" niż kolejny punkt na końcu. W świetle wcześniejszych rozważań budzi to poważne wątpliwości.



Wykres 3. Wynik sprawdzianu a skala centylowa

Dylematy dotyczące użycia konkretnej skali lub modelu do porównywania wyników w kolejnych latach może rozwiać skorzystanie z narzędzi, które oferuje probabilistyczna teoria zadania testowego (IRT). W teorii tej uzależnia się prawdopodobieństwo udzielenia poprawnej odpowiedzi na pytanie testowe od pewnej ukrytej cechy (właściwości) bardzo często nazywanej *poziomem umiejętności* ucznia (Verhelst, 2007; Ciżkowicz, 2007), *poziomem mierzonej testem cechy* (Kondratek, 2007), *umiejętności egzaminowanego ucznia* (Szaleniec, 2007) i oznaczanej grecką literą – *Θ (theta)*. Wszystkie te określenia są sobie bliskie i oznaczają mniej więcej to samo. Możemy wykorzystać „poziom *theta*" do porównania wyników uczniów z różnych roczników.

Obliczenie, a właściwie oszacowanie, wartości poziomu umiejętności ucznia na podstawie wyniku testowania jest dla laika dość skomplikowane. Liczbowa wartość theta wyrażona jest poprzez logarytm naturalny ze stosunku prawdopodobieństwa udzielenia przez ucznia prawidłowej odpowiedzi do prawdopodobieństwa udzielenia odpowiedzi błędniej (wynika to z własności funkcji logistycznej stosowanej w teorii IRT) i wyraża się wzorem:

$$\Theta = \ln\left(\frac{P(s)}{1-P(s)}\right)$$

gdzie P(s) to prawdopodobieństwo sukcesu (udzielenia przez ucznia poprawnej odpowiedzi).

O metodzie przeliczenia wyniku punktowego na poziom umiejętności *theta* pisałem przed rokiem w biuletynie XIV Konferencji Diagnostyki Edukacyjnej w Opolu (Sapanowski, 2008). Dzięki niej otrzymujemy przyporządkowanie każdemu wynikowi punktowemu liczby określającej poziom osiągnięć ucznia. Podobnie jak w przypadku wyniku standardowego będą to ułamki o wartościach ujemnych lub dodatnich w zależności od tego, czy wynik ucznia jest niższy czy też wyższy od mediany. Jak już wcześniej wspomniałem używanie tego typu skali jest niewygodne, a czasami wręcz niezręczne[3]. Dlatego zastosujmy przekształcenie liniowe do tej skali, tak aby „upodobnić" ją do najbardziej znanej skali standaryzacyjnej, czyli skali akademickiej. Wartość *theta* przemnożymy[4] przez 25 i dodamy 500, zaokrąglając wynik do całości.

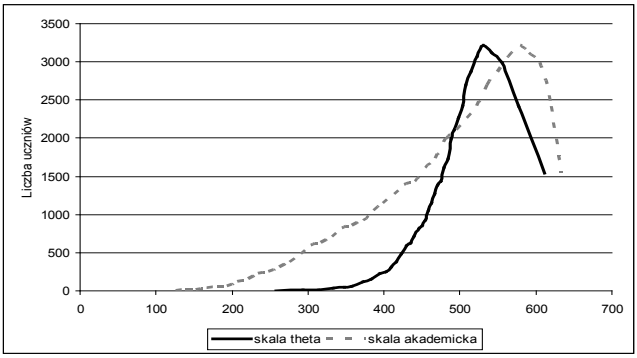
$$Z = \theta * 25 + 500$$

Kiedy popatrzymy na wykres rozkładu wyników w skali akademickiej[5] i w skali *theta* – zauważamy, że rozkład *theta* jest bardzo zbliżony do rozkładu normalnego, co raczej dobrze świadczy o prezentowanej metodzie.

[3] W przypadku ucznia, którego wynik punktowy jest niższy od mediany, komunikujemy zdającemu, że jego poziom umiejętności jest ujemny (mniej niż zero).

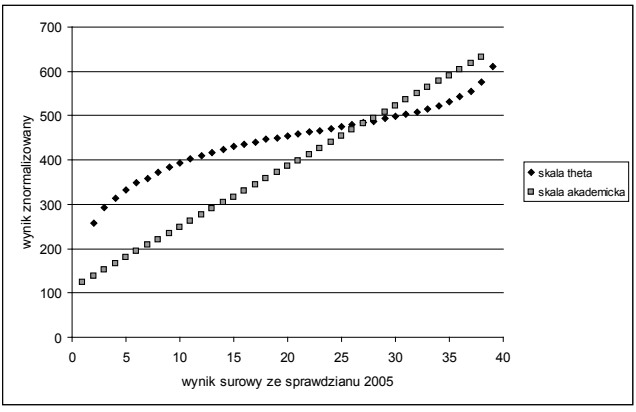
[4] Mnożnika tego nie należy utożsamiać z odchyleniem standardowym w nowym rozkładzie.

[5] Skala akademicka zachowuje parametry rozkładu klasycznego – skośność i kurtyczność.



Wykres 4. Rozkład wyników sprawdzianu w skali theta i akademickiej

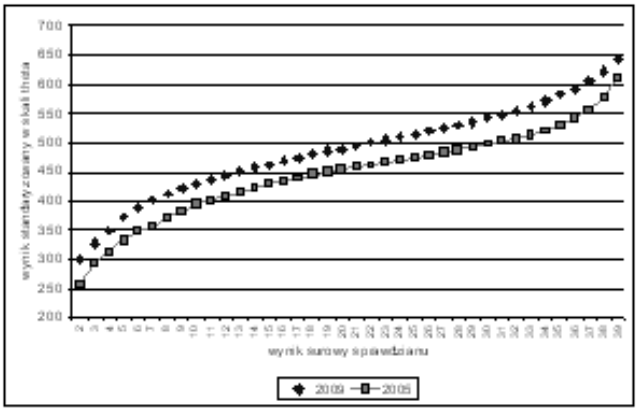
Porównując sposób, w jaki wymienione wyżej skale zamieniają wynik punktowy na wynik znormalizowany, nie sposób zauważyć, że metoda *theta* daje rezultaty zgodne z naszymi wcześniejszymi uwagami. Wyniki uczniów na końcach skali różnią się bardziej niż wynikałoby to z różnicy między surowymi wynikami punktowymi.



Wykres 5. Wynik sprawdzianu, a skala theta

Już na tym etapie wydaje się, że normalizacja (standaryzacja) wyników za pomocą skali *theta* daje lepsze rezultaty niż inne metody. Porównajmy zatem ze sobą wyniki sprawdzianu w szóstej klasie szkoły podstawowej z lat

2005 i 2009. Roczniki te wybrałem nieprzypadkowo, otóż sprawdzian w roku 2005 był najłatwiejszy, a w roku 2009 – najtrudniejszy. Zatem wszelkie różnice będą najlepiej widoczne i najłatwiejsze do przeanalizowania.



Wykres 6. Znormalizowane wyniki sprawdzianu 2005 i 2009

Uczniowie z rocznika 2009 osiągający taki sam wynik punktowy (surowy) jak ich koledzy z rocznika 2005 usytuowani są na skali znormalizowanej zdecydowanie wyżej (przeciętnie o kilkadziesiąt punktów). Jest to oczywiste, biorąc pod uwagę różnicę w trudności obu arkuszy testowych.

Zestawienie fragmentu danych w tabeli (1) pozwala zauważyć, że zdający, którzy mają w skali znormalizowanej wynik 430, uzyskali na sprawdzianach odpowiednio 10 i 15 pkt. Z kolei uczniom z wynikiem 515 odpowiada wynik punktowy 25 (w roku 2009) i 33 (w roku 2005). Są to wyniki porównywalne.

Tabela 1.

Wynik surowy na sprawdzianie		10	15	25	33
Wynik standaryzowany	2009	**430**	463	**515**	563
	2005	394	**430**	475	**515**

Przeliczanie wyników surowych na wyniki znormalizowane jest koniecznością. Nie jest możliwe analizowanie rozwoju grup uczniowskich, szkoły czy też gminy na podstawie tylko średniej i odnoszenie jej do wyników krajowych. Nakłady ponoszone przez organy prowadzące nie zawsze zaowocują świetnymi osiągnięciami uczniów. Wójt czy starosta chciałby mieć pewność, że rezultaty jego działań przekładają się na podniesienie poziomu umiejętności uczniów z terenu gminy lub powiatu. Analiza taka jest możliwa, ale tylko przy wykorzystaniu skal normalizujących, a jeszcze lepiej (taką mam nadzieję) przy użyciu skali *theta*.

Bibliografia:

1. Czarnotta-Mączyńska J., Firsiuk M., Lipska M, Lisiecka Z, *Analiza i interpretacja wyników oceniania i egzaminowania,* [w:] *Teoria i praktyka egzaminowania*, Wydział Badań i Ewaluacji CKE, Warszawa 2007.
2. Verhelst N, *Probabilistyczna teoria wyniku zadania testowego*, [w:] *Biuletyn badawczy CKE* Nr 9/2007, CKE, Warszawa 2007.
3. Ciżkowicz B., *Klasyczna a probabilistyczna teoria testu*, [w:] *Biuletyn badawczy CKE* Nr 9/2007, CKE, Warszawa 2007.
4. Kondratek B., *Teoria odpowiadania na pozycje testowe oraz klasyczna teoria testów*, [w:] *Biuletyn badawczy CKE* Nr 9/2007, CKE, Warszawa 2007.
5. Szaleniec H., *Klasyczna i probabilistyczna teoria analizy zadań egzaminacyjnych*, [w:] *Biuletyn badawczy CKE* Nr 9/2007, CKE, Warszawa 2007.
6. Sapanowski S., *Oszacowanie umiejętności „theta" oraz wyskalowanie osi w metodzie IRT,* [w:] *XIV Konferencja Diagnostyki Edukacyjnej*, Opole 2008.