DOROTA WĘZIAK
SGH w Warszawie

ZASTOSOWANIE WIELOCZYNNIKOWEGO SKALOWANIA RASCHA DO PORÓWNANIA SPOSOBU OCENIANIA EGZAMINATORÓW

Wprowadzenie

Na ocenę uzyskaną przez studenta z egzaminu wpływ mają: poziom jego wiedzy, trudność kolokwium zaliczeniowego, różnice w surowości lub pobłażliwości egzaminatorów sprawdzających prace, sposób stosowania skali punktowej w każdym zadaniu.

Za pomocą wieloczynnikowego skalowania Rascha (ang. *Many Facet Rasch Model*), metody zaproponowanej przez M. Linacre i zaprezentowanej pokrótce w tym opracowaniu, sprawdzone zostanie, czy egzaminatorzy różnią się pod względem srogości oceniania oraz czy ma to wpływ na oceny uzyskiwane przez studentów, a także, czy osoba prowadzącego zajęcia ma wpływ na sposób rozwiązywania poszczególnych zadań.

1. Wieloczynnikowe skalowanie Rascha

Wieloczynnikowe skalowanie Rascha, wykorzystując transformację ogólnej sumy punktów będącej do tej pory podstawą do wystawienia oceny w miarę o charakterze interwałowym, pozwala uzyskać obiektywne oceny zarówno srogości sędziów, trudności zadań, jak poziomu wiedzy studentów.

Model wieloczynnikowego skalowania Rascha zdefiniowany został następująco:

$$\ln\left(\frac{P_{nikj}}{P_{ni(k-1)j}}\right) = B_n - D_i - R_j - F_k$$

gdzie:

P_{nikj} – prawdopodobieństwo przyznania przez *j*-tego sędziego *k*-tej kategorii punktowej na *i*-tą pozycję skali rozwiązywaną przez *n*-tego respondenta,

R_j – oszacowanie srogości oceniania *j*-tego sędziego,

F_k – oszacowanie *k*-tego progu, D_i – oszacowanie trudności *i* – tego zadania,

B_n – oszacowanie poziomu wiedzy *n*-tego respondenta.

Wieloczynnikowe skalowanie Rascha zakłada zatem, że pozycja respondenta na skali jest nie tylko funkcją poziomu wiedzy badanego (*B*) i trudności rozwiązywanego zadania (*D*), ale również sposobu oceniania zadań przez sędziego (*R*). Parametry F_k odpowiadają punktom przejścia między *k* i *k+1* kategoriami punktowymi. Indeksowane są po *k*, ponieważ założono, że struktura skali ocen jest jednakowa dla wszystkich zadań rozwiązywanych przez respondenta.

Efektywne zastosowanie skalowania Rascha uzależnione jest od wystąpienia dwóch podstawowych cech badanego testu lub sprawdzianu:

1. jednowymiarowości, która oznacza, że zależności między pozycjami testu są wyjaśniane przez tylko jedną zmienną latentną (np. poziom wiedzy, kompetencje),
2. lokalnej niezależności (ang. *local independence*), a więc odpowiedź na rozwiązywaną pozycję (zadanie) nie jest zależna od odpowiedzi na inne pozycje (zadania).

Te dwa podstawowe założenia sprawdzane są za pomocą statystyk dopasowania[1], których konstrukcja opiera się na porównaniu obserwowanych w wyniku badania rezultatów z rezultatami oczekiwanymi wynikającymi z założeń metody[2]. Statystyki dopasowania wykorzystywane są również do wskazywania nietypowych wzorów odpowiedzi.

Wszystkie parametry mierzone są logitem[3], a więc są porównywalne względem siebie. Dodatkowo każde oszacowanie otrzymywane jest wraz z odpowiadającym mu standardowym błędem szacunku, co ułatwia pomiar rzetelności narzędzia pomiarowego.

2. Opis badania

Dążąc do ujednolicenia zasad zaliczenia przedmiotów Statystyka I i Statystyka II na studiach dziennych i zaocznych w Szkole Głównej Handlowej, zdecydowano się wprowadzić wystandaryzowane kolokwium zaliczeniowe. To rozwią-

---

1 Odpowiednie wzory Czytelnik znajdzie w artykule: Linacre J. M., *What do Infit and Outfit, Mean-square and Standardized mean*?, „Rasch Measurement Transactions" 16:2, Autumn 2002, www.rasch.org/rmt/rmt162f.htm

2 Po pierwsze, jest bardziej prawdopodobne, że osoby, które uzyskały wyższą sumę punktów, rozwiążą poprawniej poszczególne zadania niż osoby z niższą sumą punktów. Po drugie, jest bardziej prawdopodobne, że zadania łatwiejsze zostaną rozwiązane poprawniej niż zadania trudniejsze.

3 Logit – logarytm naturalny szansy.

zanie jednak nie do końca zapewniło porównywalność wyników. Na przeszkodzie temu w dalszym ciągu stoją:

- konieczność przygotowywania dwóch wersji zaliczenia (nie zawsze można ocenić ich porównywalność),
- brak szczegółowych wytycznych na temat oceniania,
- duża liczba egzaminatorów sprawdzających prace studentów,
- ocenianie jednej pracy przez tylko jednego egzaminatora.

W celu sprawdzenia, czy oceny z egzaminu uzyskane u różnych egzaminatorów są porównywalne, wyniki kolokwium zaliczeniowego z czerwca 2005 r. zostały przeanalizowane za pomocą wieloczynnikowego skalowania Rascha.

Do analizy zakwalifikowanych zostało w sposób celowy 107 studentów uczęszczających na wykład ze Statystyki I, prowadzony przez profesora R, i rozwiązujących zestaw A kolokwium zaliczeniowego. 60 ze 107 prac zostało dwukrotnie sprawdzone (przez dwóch niezależnych względem siebie egzaminatorów – R i W), zaś wybranych 28 prac zostało jeszcze sprawdzone po raz trzeci przez egzaminatora P. Jednym z egzaminatorów był profesor R, zaś dwoje pozostałych to jego asystenci P i W. Należy nadmienić, że w przypadku egzaminatorów R i P nie były to pierwsze doświadczenia z oceną tego typu pracy. Danych osobowych studentów nie zatajano w procesie oceny.

Dobór prac do dwu- i trzykrotnego sprawdzania był celowy: wybrano po 12 studentów od wszystkich 5 asystentów w przypadku drugiego sprawdzającego W i po 5 lub 6 prac od wszystkich asystentów dla trzeciego sprawdzającego P. Asystenci prowadzący zajęcia w formie ćwiczeń do wykładu oznaczeni zostali w tekście jako: P, P-C, S, T, W.

Studenci w trakcie kolokwium mieli do rozwiązania 6 niezależnych od siebie i jednakowo punktowanych zadań. Za każde z nich można było uzyskać od 0 do 5 pkt, dopuszczono przyznawanie 0,5 pkt. Przed analizą wszystkie przyznane punkty zostały przemnożone razy dwa w celu uzyskania liczb całkowitych. Tym sposobem zastosowana do oceny skala punktowa miała 11 stopni. Analizę przeprowadzono stosując program Minifac (Facets Student/Evaluation) v. 3.55.1.

3. Analiza wyników kolokwium zaliczeniowego

Wyniki przeprowadzonych analiz pokazały, że chociaż obliczone dla poszczególnych zadań i dla sumy punktów współczynniki korelacji między ocenami przyznanymi przez poszczególnych egzaminatorów okazały się być stosunkowo wysokie (tab. 1.), to ogólne oceny surowości oceniania każdego z egzaminatorów (R_j) różniły się: najsurowszy okazał się egzaminator P, a najłagodniejszy egzaminator R (tab. 2.).

Pomimo ustalonego wcześniej sposobu oceniania poszczególnych zadań, między egzaminatorami R i W wystąpiły różnice. Przeciętna liczba punktów przyznanych za pojedyncze zadanie wyniosła dla egzaminatora R: 3,88, a dla egzaminatora W: 3,69. Te różnice przełożyły się na poziom oszacowania określa-

jącego surowość oceniania (R_j). Jednak najbardziej pod względem oceniania różnił się egzaminator P, z którym nikt sposobu oceniania nie konsultował. Przeciętna liczba punktów przyznanych przez niego kształtowała się na poziomie 3,25. Oszacowania poziomu surowości oceniania wraz z odpowiadającymi im wystandaryzowanymi statystykami dopasowania zawiera tab. 2.

Tabela 1. Współczynniki korelacji między ocenami przyznanymi przez poszczególnych egzaminatorów

Współczynniki korelacji liniowej między punktami przyznanymi przez egzaminatora							
	Zad 1	Zad 2	Zad 3	Zad 4	Zad 5	Zad 6	Suma pkt.
R i P	0,665	0,778	0,807	0,790	0,717	0,871	0,909
W i P	0,572	0,797	0,861	0,883	0,750	0,735	0,923
R i W	0,946	0,749	0,862	0,806	0,766	0,623	0,865

Źródło: Opracowanie własne.

Tabela 2. Porównanie egzaminatorów

Egzaminator	Śr. ocena obserwowana	Średnia ocena skorygowana	Ostrość oceniania (R_j)	Błąd stand.	SIMS	SOMS
R	7,8	8,68	0,11	0,02	3,4	0,9
W	7,4	8,36	0,01	0,02	-1,2	-1,3
P	6,5	7,69	-0,12	0,03	-2,2	-2,1

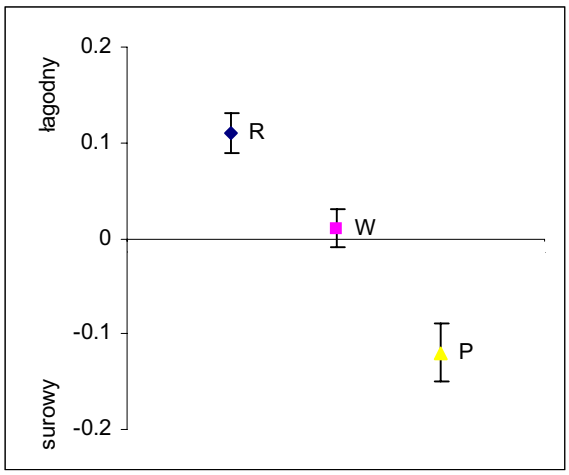
Źródło: Opracowanie własne.

Statystyki dopasowania *SIMS* i *SOMS* w przypadku egzaminatorów R i P wykraczają poza zalecany przedział <-2, 2>[4]. W przypadku egzaminatora R wskazują na zbyt zróżnicowany sposób oceniania w porównaniu do założeń metody, zaś w przypadku egzaminatora P – na zbyt mało zróżnicowany sposób oceniania. Powstaje zatem podejrzenie, że egzaminator R nie stosował w całym procesie oceniania tych samych kryteriów, ale również może to być wynikiem sposobu odpowiadania studentów podlegających ocenie przez tego właśnie egzaminatora.

Wprost przeciwna sytuacja miała miejsce w przypadku egzaminatora P – jego sposób oceniania był zbyt mało zróżnicowany. Mniejsze od -2 wartości statystyk dopasowania mogą również być wynikiem tego, że oceniał akurat studentów o podobnym poziomie wiedzy i sposobie odpowiadania na poszczególne pytania.

Oszacowanie poziomu srogości oceniania *R* każdego z egzaminatorów wykonywane jest w celu późniejszej korekty ostatecznego wyniku studenta, aby jego ocena lub decyzja o zaliczeniu przedmiotu nie zależała od wyboru egzaminatora sprawdzającego prace.

[4] Bond T.G., Fox Ch.M., *Applying The Rasch Model*, „Fundamental Measurement in the Human Science" Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey 2001, s. 209.



Rys. 1. Porównanie surowości oceniania poszczególnych egzaminatorów
Źródło: Opracowanie własne.

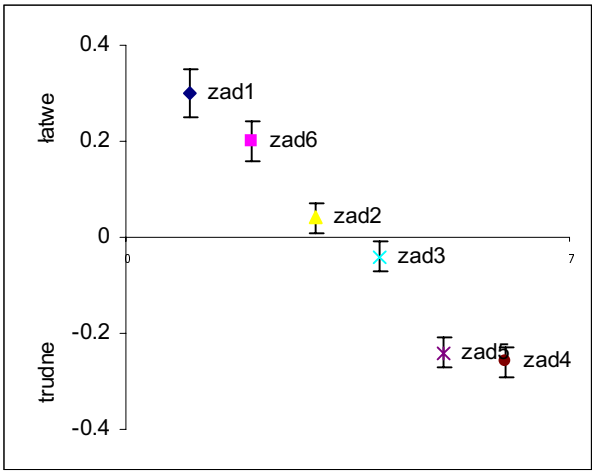
Aby dokładniej zweryfikować kompetencje egzaminatorów, sprawdzono również, czy nie wystąpiły zależności między określonym typem zadań a stosowanym przez egzaminatorów systemem przyznawania ocen. W pierwszym kroku oszacowano poziom trudności poszczególnych zadań z kolokwium egzaminacyjnego D_i. Szczegółowe wyniki prezentuje tab. 3.

Tabela 3. Porównanie trudności zadań egzaminacyjnych

Zad.	Średnia ocena obserwowana	Średnia ocena skorygowana	Trudność zadania (D_i)	Błąd stand.	SIMS	SOMS
zad1	8,8	9,08	0,3	0,05	0,5	-0,2
zad6	8,5	8,91	0,2	0,04	-1,5	-1,3
zad2	7,8	8,45	0,04	0,03	1,8	0,7
zad3	7,4	8,12	-0,04	0,03	-1,2	-0,1
zad5	6,1	6,84	-0,24	0,03	1,8	0,4
zad4	6,0	6,66	-0,26	0,03	0,1	-0,7

Źródło: Opracowanie własne.

Najtrudniejsze okazały się zadania 4. i 5., choć różnica między oszacowaniami ich trudności mieści się w granicach błędu standardowego. Te dwa zadania różnią się natomiast znacząco od pozostałych. W dalszej kolejności ze względu na trudność następowały zadania: 3., 2. i 6., zaś najłatwiejsze dla badanej grupy studentów okazało się zadanie 1. (rys. 2.). Statystyki dopasowania dla wszystkich zadań mieściły się w zalecanych granicach, co oznaczało, że zastosowany zestaw zadań spełniał warunek jednowymiarowości i mógł służyć do pomiaru za pomocą skalowania Rascha.



Rys. 2. Porównanie trudności zadań egzaminacyjnych
Źródło: Opracowanie własne.

Ponieważ założono, że osoba prowadzącego ćwiczenia do wykładu nie ma wpływu na końcową ocenę studenta, jako zmienną pasywną wprowadzono do analizy przynależność studentów do grup ćwiczeniowych. Zbadano, czy ma ona wpływ na sposób rozwiązywania poszczególnych typów zadań, a także, czy sposób oceniania egzaminatorów jest jednakowy wobec wszystkich studentów, bez względu na osobę prowadzącą ćwiczenia. Ten drugi problem zaistniał, ponieważ tylko dwoje z pięciu asystentów prowadzących ćwiczenia zostało egzaminatorami. Powstało zatem podejrzenie o łagodniejsze ocenianie (w sposób zamierzony lub nie) prac „swoich" studentów, wynikające np. z szerszej wiedzy na ich temat, obejmującej m.in. stopień przygotowania do zajęć, aktywność na ćwiczeniach, regularność pracy itp.

Przeprowadzone analizy wykazały, że nie wystąpiły znaczące różnice w:

1. ocenie studentów należących do grupy prowadzonej przez oceniającego prace egzaminatora, a studentów z innych grup (χ^2 (15) = 2,6, α* = 1,00);
2. systemie oceniania poszczególnych zadań przez poszczególnych egzaminatorów (χ^2 (18) = 11,6, α* = 0,87);

Następnie sprawdzono, czy w poszczególnych grupach ćwiczeniowych wystąpiły predyspozycje do rozwiązywania określonych typów zadań (tab. 4). Analiza pokazała, że takie różnice wystąpiły (χ^2 (30) = 101,6, α* = 0,00).

Po wzięciu pod uwagę osób o tej samej sumie punktów okazało się, że studenci z grup asystenta W znacząco gorzej rozwiązali zadanie 2. oraz zadanie 1., choć w przypadku zadania 1. różnica była najmniejsza spośród wszystkich znaczących. Natomiast ci sami studenci rozwiązali zadanie 4. znacząco lepiej w stosunku do innych grup ćwiczeniowych.

W przypadku asystenta T studenci gorzej niż to przewiduje model poradzili sobie z zadaniami 4. i 6., w porównaniu do osób o tej samej sumie punktów

z innych grup. Podobna sytuacja wystąpiła w przypadku grup asystenta P-C i zadania 5.

Studenci z grup asystentów S i P stosunkowo lepiej radzili sobie z zadaniami 6. i 2. w porównaniu do osób o tej samej sumie punktów z innych grup.

Dysproporcje „na plus" oznaczać mogą, że bez względu na poziom wiedzy studenci stosunkowo lepiej radzili sobie z zadaniem określonego typu. Być może we wskazanych grupach poświęcono tym zadaniom więcej czasu lub dokładniej je przeanalizowano lub wreszcie studenci wiedzieli, że tego typu zadanie może się pojawić na egzaminie i dokładniej się do niego przygotowali. Dysproporcje „na minus" wskazują na istnienie problemów, ponieważ bez względu na poziom wiedzy studenci stosunkowo gorzej sobie radzili z rozwiązaniem danych zadań.

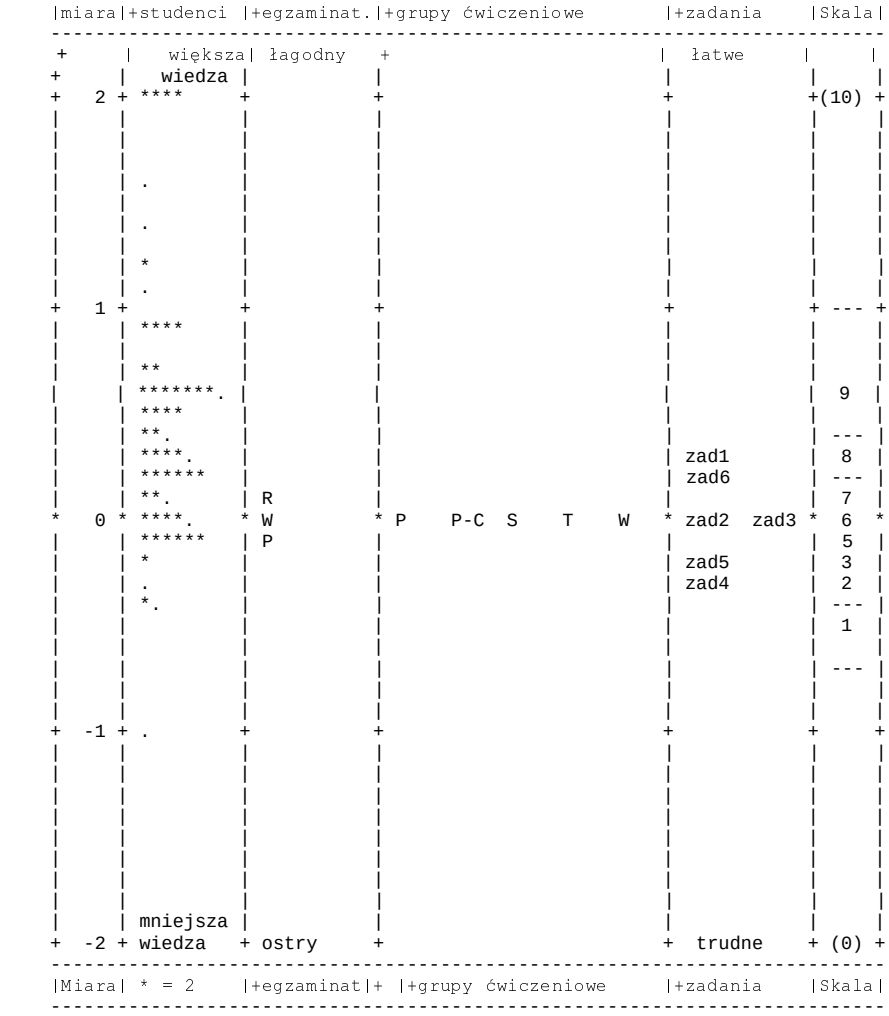
Tabela 4. Różnice w rozwiązaniu poszczególnych zadań a przynależność do grupy ćwiczeniowej

Różnica między obserwowaną średnią l. punktów a oczekiwaną średnią l. punktów	Oszacowanie obciążenia	Błąd standardowy	Grupa ćwiczeniowa	Zadanie
-1,67	0,30	0,06	W	zad2
-0,56	0,19	0,08	W	zad1
-1,31	0,18	0,06	T	zad4
-1,13	0,15	0,06	P-C	zad5
-0,91	0,15	0,06	T	zad6
1,63	-0,29	0,08	W	zad4
0,92	-0,34	0,13	S	zad6
0,92	-0,45	0,15	P	zad2

Źródło: Opracowanie własne.

Zbiorcze zestawienie wszystkich oszacowanych parametrów wieloczynnikowego modelu Rascha prezentuje rys. 3. Kolejne kolumny na wykresie odpowiadają szacowanym parametrom, zaś poszczególne elementy usytuowane są na jednej osi, której jednostką jest 1 logit (pierwsza kolumna).

Oszacowania poziomu wiedzy studentów, skorygowane w stosunku do surowej sumy punktów uzyskanych z kolokwium, przedstawione zostały w drugiej kolumnie. Korekta uwzględniała poziom srogości egzaminatorów oceniających pracę danego studenta i została wprowadzona, aby zapobiec sytuacji, w której studenci o zbliżonym poziomie wiedzy uzyskują różny wynik w zależności od wyboru oceniającego pracę egzaminatora.



Rys. 3. Mapa

Źródło: Wydruk programu Minifac.

Na rysunku 3 grupy ćwiczeniowe (P, P-C, S, T, W) usytuowane są na jednakowym poziomie, ponieważ do analizy zostały wprowadzone jako zmienne pasywne. Wielkości wszystkich pozostałych parametrów zostały oszacowane metodą największej wiarygodności.

Podsumowanie

W wyniku przeprowadzonych analiz udało się: (1) oszacować względny poziom srogości egzaminatorów i względny stopień trudności zadań egzamina-

cyjnych, (2) zweryfikować negatywnie pogląd, że egzaminator „lepiej" ocenia „swoich" studentów, a także (3) pogłębić wiedzę o stopniu przyswajania wiedzy przekazywanej studentom przez asystentów w trakcie ćwiczeń.

Kolejny etap prac będzie obejmował:

1. włączenie do analizy wyników studentów rozwiązujących zestaw B,
2. próbę transformacji logitowej skali wyników studentów na powszechnie wykorzystywaną skalę punktową wraz z opracowaniem progów punktowych odpowiadających odpowiednim ocenom,
3. próbę opracowania zasad komunikacji wyników studentom.