Anna Dubiecka, Henryk Szaleniec, Dorota Węziak

1 i 2 – Okręgowa Komisja Egzaminacyjna Kraków
3 – Instytut Statystyki i Demografii SGH

Efekt egzaminatora

Wstęp

W egzaminach zewnętrznych w Polsce prace uczniowskie oceniane są przez niezależnych zewnętrznych egzaminatorów przy zastosowaniu procedury oceniania kryterialnego. W tym przypadku najistotniejsze znaczenie dla rzetelności egzaminu (w zakresie zależnym od komisji egzaminacyjnych) ma stabilność posługiwania się kryteriami przez egzaminatorów oceniających prace uczniowskie oraz zapobieganie ewentualnym błędom systematycznym.

Schemat oceniania, procedura oceniania i egzaminator, jeżeli nie są poddawane ciągłemu monitorowaniu, mogą stanowić istotne źródła potencjalnych błędów obciążających wyniki egzaminu. W latach 2005-2006 autorzy przeprowadzili badania mające na celu zarejestrowanie zróżnicowania surowości egzaminatorów sprawdzianu punktujących zadania otwarte w hierarchicznej strukturze koordynacji procesu oceniania. Wyniki tych badań dostarczają informacji, które mogą mieć praktyczne znaczenie dla planowania monitorowania jakości oceniania zadań otwartych. Badania przeprowadzono bezpośrednio przed przystąpieniem egzaminatorów do właściwego oceniania, dzięki czemu można było wykorzystać wstępne wyniki do doskonalenia pracy egzaminatorów. W kolejnych rozdziałach artykułu autorzy przybliżają terminologię związaną z efektem egzaminatora, przedstawiają model zastosowany w badaniach i prezentują wyniki.

Przeprowadzone analizy wykazały istnienie zróżnicowania poziomów surowości egzaminatorów zarówno na poziomie indywidualnym, jak i na poziomie zespołu, czy też ośrodka koordynacji oceniania.

Potencjalne źródła błędu w ocenianiu prac egzaminacyjnych

Akt oceniania oznacza proces, w który zaangażowany jest egzaminator podejmujący decyzję o poziomie lub stopniu, w jakim oceniany demonstruje lub posiada cechę, która jest przedmiotem oceniania. Egzaminator posługuje się schematem oceniania, którego nieodłączną częścią jest skala (określona w modelu oceniania), jako narzędziem do podejmowania decyzji. Egzaminator, posługując się skalą oceniania, przypisuje egzaminowanym pozycje na skali w zależności od poziomu cechy (umiejętności), która jest przedmiotem pomiaru.

Popham (1990) podkreśla, że są trzy potencjalne źródła błędów, które powinny być monitorowane podczas oceniania: schemat oceniania, procedura oceniania i oceniający (egzaminator).

W jaki sposób schemat oceniania może wprowadzać błąd pomiaru w procesie oceniania? Na przykład pomimo starannego procesu moderowania oceniania nie wszyscy egzaminatorzy w ten sam sposób rozumieją to, co jest przedmiotem oceniania. Taka sytuacja ma miejsce, gdy czynność będąca przedmiotem oceny nie jest jasno określona. Inna sytuacja może dotyczyć niezbyt jasnego zdefiniowania skali dla danego kryterium lub sytuacji, gdy poszczególne stopnie skali zachodzą na siebie i są trudne dla egzaminatorów do rozdzielenia (Szaleniec, 2001). Również wtedy, kiedy granice pomiędzy poszczególnymi kategoriami są nieostre, egzaminatorzy często mają dylemat z zakwalifikowaniem odpowiedzi do poszczególnych kategorii wzdłuż skali określonej w schemacie oceniania.

W jaki sposób procedury oceniania mogą wpływać na wprowadzanie błędu? Jest szereg aspektów związanych z procedurami oceniania, które mogą przysparzać problemów oceniającym. Procedura oceniania może wymagać od egzaminatora oceny jednocześnie zbyt wielu czynności (ocenianie eseju, rozprawki, złożonego zadania matematycznego), co powoduje trudność w klarownym pamiętaniu różnic pomiędzy tymi czynnościami i stopniami skali.

Jeżeli ocenianie trwa wiele godzin każdego dnia (jak to ma miejsce w koordynowanym ocenianiu w naszym kraju) i przez wiele dni (soboty i niedziele), niektórzy egzaminatorzy mogą czuć się bardziej zmęczeni niż inni. Niektórzy są bardziej sprawni rano, inni – po południu, jeszcze inni potrafią być tak samo sprawni przez wiele godzin i wiele dni.

W jaki sposób egzaminator może być źródłem błędów w ocenianiu? Egzaminator podczas oceniana zaangażowany jest w bardzo złożony i podatny na błędy proces. Thorndike i Hagen (1977) zwracają uwagę, że proces oceniania przypomina model „czarnej skrzynki". Znane jest to, co na wejściu i rezultat. Sam proces zachodzący pomiędzy danymi na wejściu (obserwacja zapisanego efektu pracy ucznia na egzaminie) i rezultatem oceniania (przyznana liczba punktów) jest mało poznawalny. Procesy poznawcze towarzyszące ocenianiu w złożony sposób zależne są zarówno od przeszłych doświadczeń nauczycielskich, od re-

zultatów szkolenia poprzedzającego ocenianie, jak i od cech osobowościowych danego egzaminatora.

Zanim przejdziemy do opisu badań prowadzących do określenia wpływu egzaminatora na wynik Sprawdzianu w szóstej klasie szkoły podstawowej, spróbujemy najpierw zdefiniować pojęcie efektu egzaminatora.

Scullen, Mount i Goff (2000) definiują efekt egzaminatora jako szeroką kategorię efektów generujących wariancję wyników oceniania, która nie wynika ze zmienności poziomu ocenianej cechy, lecz jej źródłem są egzaminatorzy.

Do najczęściej spotykanych efektów należą:

1. łagodność lub surowość oceniania,
2. efekt halo,
3. tendencja centralna w ocenianiu,
4. tendencja ocen ekstremalnych.

W tym artykule zajmiemy się tylko pierwszą kategorią efektów, gdyż była ona przedmiotem badań prowadzonych przez autorów na populacji egzaminatorów oceniających prace Sprawdzianu w OKE Kraków.

Guilford definiuje **łagodność** oceniania jako stałą cechę związaną raczej z osobowością egzaminatora występującą niezależnie od sytuacji egzaminacyjnej. Rezultat oceny łagodnego egzaminatora powoduje, że przeciętny wynik oceniania jest powyżej średniej wartości ocenianej cechy zgodnie z ustalonymi zewnętrznymi kryteriami.

O **surowości** oceniania będziemy natomiast mówić w przypadku, gdy przeciętny wynik oceniania jest poniżej średniej wartości ocenianej zgodnie z ustalonymi zewnętrznymi kryteriami cechy.

Można wyróżnić trzy podejścia do badania efektu łagodności i surowości oceniania.

1. Porównanie wyniku danego egzaminatora ze środkiem skali.
2. Analiza wariancji w poszukiwaniu statystycznej różnicy oceniania poszczególnych egzaminatorów.
3. Analiza skośności rozkładu częstości wyniku oceniania danej cechy.

W przeprowadzonych badaniach do zmierzenia i porównania surowości egzaminatorów wstępnie autorzy wykorzystywali wszystkie trzy wymienione podejścia. Głównie jednak zastosowano analizę wariancji (hierarchiczny model regresyjny) a ponadto wieloczynnikowe skalowanie Rascha (Many Facet Rasch Model). Wieloczynnikowe skalowanie Rascha wykorzystuje logitową transformację ogólnej sumy punktów tradycyjnie stanowiącej podstawę do przekształcenia oceny w miarę o charakterze interwałowym. Pozwala ono uzyskać wyrażone w tej samej skali oceny zarówno stopnia surowości oceniających, trudności zadań, jak i poziomu wiedzy uczniów (Węziak, 2005; Myford, Wolfe, 2006).

Krótkie wprowadzenie do wieloczynnikowego skalowania Rascha

Zastosowany model wieloczynnikowego skalowania Rascha wiąże prawdopodobieństwo – *P* poprawnej odpowiedzi udzielanej przez poszczególnych egzaminowanych z czterema zmiennymi takimi jak: wskaźnik poziomu umiejętności egzaminowanego – *B*, wskaźnik trudności zadania – *D*, wskaźnik trudności na progach na skali punktowania zadania – *F i* wskaźnik surowości oceniania – *R*. Opis matematyczny modelu przedstawia funkcja:

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

gdzie: P_{nikj} – prawdopodobieństwo przyznania przez *j*-tego egzaminatora *k*-tej kategorii punktowej za rozwiązanie zadania – i przez *n*-tego egzaminowanego,

R_j – oszacowanie surowości oceniania *j*-tego egzaminatora, F_k – oszacowanie *k*-tego progu, D_i – oszacowanie trudności *i*-tego zadania[1], B_n – oszacowanie poziomu wiedzy *n*-tego egzaminowanego.

Wieloczynnikowe skalowanie Rascha zakłada, że wynik uzyskiwany przez egzaminowanego jest nie tylko funkcją poziomu jego wiedzy (*B*) i trudności rozwiązywanego zadania (*D*), ale również sposobu oceniania zadań przez sędziego/egzaminatora (*R*). Parametry F_k w modelu odpowiadają punktom przejścia między *k*-tą i *k+1* kategorią punktową. Kategorie punktowania indeksowane są tylko po *k*, ponieważ założono, że struktura skali punktowej jest jednakowa dla wszystkich zadań rozwiązywanych przez ucznia. W przypadku sprawdzianu mamy do czynienia z różną długością skali dla poszczególnych zadań. To ostatnie założenie można jednak uchylić, ponieważ dostępne obecnie oprogramowanie pozwala prowadzić analizy także w przypadku zróżnicowanego punktowania zadań. Jednak w tym artykule, w szczególności w jego części teoretycznej, pozostaniemy przy modelu sformułowanym przy założeniu jednakowej punktacji zadań egzaminacyjnych.

Prawidłowe zastosowanie skalowania Rascha uzależnione jest od występowania dwóch podstawowych cech badanego testu lub sprawdzianu:

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

[1] W skalowaniu Rascha i innych modelach IRT również przyjęła się konwencja charakteryzowania zadania poziomem jego trudności w przeciwieństwie do klasycznej teorii testu, gdzie raczej zadania charakteryzuje się za pomocą stopnia współczynnika łatwości.

Te dwa podstawowe założenia sprawdzane są za pomocą odpowiednich statystyk dopasowania[2], których konstrukcja opiera się na porównaniu obserwowanych w wyniku badania rezultatów z rezultatami oczekiwanymi wynikającymi z założeń metody[3]. Statystyki dopasowania wykorzystywane są również do wykrywania nietypowych układów odpowiedzi. Jak już wcześniej wspomniano, wszystkie mierzone parametry wyrażane są w tej samej skali, a więc są względem siebie porównywalne. Jednostka tej skali nazywana jest logitem[4]. Dodatkowo każde oszacowanie otrzymywane jest wraz z odpowiadającym mu standardowym błędem szacunku, co ułatwia pomiar rzetelności narzędzia pomiarowego. Zwyczajowo „0" dla skali ustalane jest w punkcie odpowiadającym średniej trudności zadań w teście, a pozostałe parametry szacowane są względem niego. Pozwala to na umieszczenie na jednej skali metrycznej wszystkich typów oszacowanych parametrów i bezpośrednie ich porównanie.

Opis przeprowadzonych badań

W każdej sesji egzaminacyjnej przed rozpoczęciem oceniania egzaminatorzy odbywają wstępne szkolenie połączone z oceną przykładowych prac uczniowskich w ramach moderowania oceniania. W celu praktycznego wdrożenia efektów szkolenia oraz na użytek badań przygotowano 15 prac uczniowskich, zróżnicowanych ze względu na poziom rozwiązania zadań. Każdy egzaminator oceniał prace pięciu uczniów. W trakcie badań poszukiwano odpowiedzi na następujące pytania:

1. Jak wielkie jest zróżnicowanie surowości oceniania pomiędzy poszczególnymi egzaminatorami?
2. Czy wynik oceniania zależy od ośrodka koordynacji?
3. Czy wynik oceniania zależy od zespołu, w którym pracuje egzaminator?
4. Czy wynik zależy od treści zadania i jego schematu oceniania?

Badania zostały zaplanowane na okres dwóch lat. Populację stanowili egzaminatorzy zatrudnieni do oceniania sprawdzianu w OKE Kraków. Zestaw 15 prac planowanych do oceny został przygotowany poprzez wybór autentycznych rozwiązań uczniowskich. Do rejestracji wyników punktowania przygotowano

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

[3] Założenia te mówią, że: po pierwsze, jest bardziej prawdopodobne, że osoby, które uzyskały wyższą sumę punktów, rozwiążą poprawnie poszczególne zadania niż osoby z niższą sumą punktów; po drugie, jest bardziej prawdopodobne, że zostaną rozwiązane poprawnie zadania łatwiejsze niż zadania trudniejsze.

[4] Logit – logarytm naturalny szansy.

karty odpowiedzi pozwalające identyfikować trzy zmienne: z1 – egzaminator, z2 – zespół egzaminatorów, z3 – ośrodek koordynacji oceniania.

Ze względu na to, że tylko ci egzaminatorzy, którzy uzyskali wysoką zgodność punktowania z wynikami sędziów kompetentnych, mogli przystąpić do oceniania prac uczniowskich, zapewniono egzaminatorom samodzielność pracy oraz wysoką motywację. Wstępne wyniki analizy dokonanej podczas szkolenia egzaminatorów zostały natychmiast wykorzystane jako informacja zwrotna kierowana do poszczególnych egzaminatorów przed przystąpieniem do oceniania w danej sesji egzaminacyjnej (odpowiednio 2005 r. i 2006 r.). W dalszej części tekstu skupimy się na rezultatach badań z 2006 roku, w których uczestniczyło 1349 egzaminatorów pracujących w 80 zespołach zorganizowanych w 24 ośrodki koordynacji oceniania.

Tabela 1. Opis populacji egzaminatorów uczestniczących w badaniach w 2006 roku w podziale na ośrodki koordynacji

Województwo	Liczba egzaminatorów	Udział (%)	Liczba ośrodków koordynacji	Liczba zespołów	Udział (%)
Lubelskie (LO)	393	29.13	7	23	28.75
Małopolskie (MO)	515	38.18	9	31	38.75
Podkarpackie (PO)	441	32.69	8	26	32.50
SUMA	1349	100	24	80	100

Uczniowie, których prace oceniano w ramach eksperymentu, nazwani zostali symbolami A01 – A05, B01 – B05 oraz C01 – C05. W schemacie oceniania 5 zadań otwartych zostało rozbite na 18 odrębnych czynności. Dla każdej czynności zdefiniowano odrębnie kryterium zaliczenia opanowania jej oraz skalę. Dla 16 czynności została przyjęta skala 0-1 a dla dwóch skala 0, 1, 2. Tabela 2. przedstawia wyniki oszacowania trudności poszczególnych czynności wyrażonej w jednostkach logit.

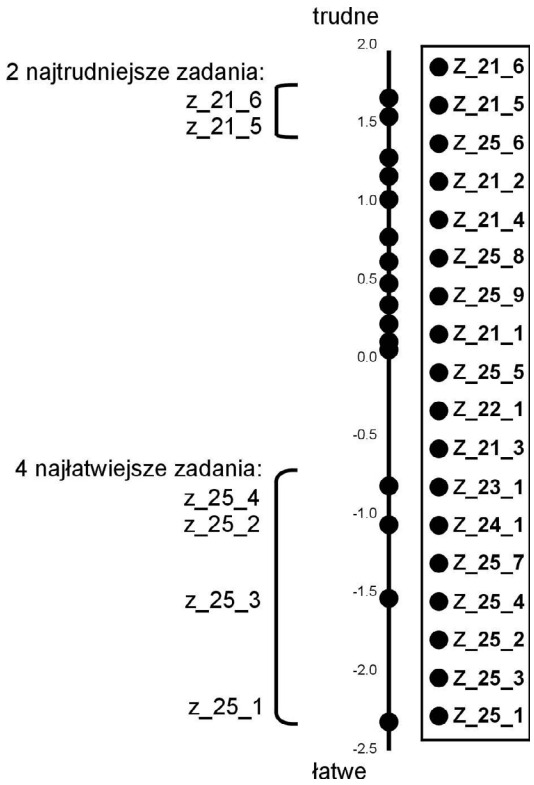
Tabela 2. Zadania otwarte arkusza standardowego sprawdzianu 2006 i ich punktacja

Kod czynności	Liczba punktów	Trudność w jednostkach logit
Z_21_1	0 – 1	0,61
Z_21_2	0 – 1	1,15
Z_21_3	0 – 1	0,33
Z_21_4	0 – 1	1,01
Z_21_5	0 – 1	1,54
Z_21_6	0 – 1	1,66
Z_22_1	0 – 1 – 2	0,34
Z_23_1	0 – 1	0,21
Z_24_1	0 – 1	0,1
Z_25_1	0 – 1	-2,27
Z_25_2	0 – 1	-1,07

Poziomy trudności poszczególnych czynności wchodzących w skład zadań otwartych arkusza standardowego Sprawdzianu 2006 wykorzystanych w badaniach oszacowano na wynikach 10% prostej próby losowej wylosowanej z krajowej populacji wszystkich uczniów piszących arkusz standardowy (A1). Wykorzystano do tego celu wszystkie zadania składające się na Sprawdzian (czyli na podstawie wyników właściwego egzaminu, a nie badań). Dla uproszczenia i większej czytelności, w dalszej części tekstu wyodrębnione w Sprawdzianie części zadań sprawdzające określone czynności będziemy również nazywać zadaniami.

Podczas kalibracji narzucono warunek, aby średnia trudność wszystkich zadań składających się na Sprawdzian, łącznie z zadaniami zamkniętymi, wyniosła 0. W wyniku przyjęcia takiego warunku średnia trudność zadań otwartych, których ocenianie analizowano w badaniach wyniosła 0,255 logita z odchyleniem standardowym 1,036 logita. Wszystkie zadania (z wyjątkiem zadania *z_22_1*) wykorzystane w badaniach wykazywały akceptowalną zgodność z założeniami metody skalowania (wartości *OMS* i *IMS* w granicach <0,7; 1,3>) (T. G. Bond, 2001). Jedynie w przypadku zadania *z_22_1* wychwycono zbytnie zróżnicowanie przyznanych za nie punktów. Przyczyny takiej sytuacji należałoby upatrywać w punktacji przewidzianej za to zadanie. Zakładano punktację 0-1-2 (i taki warunek narzucono przy estymacji modelu), podczas gdy rozkład punktów w rzeczywistości przyznawanych był raczej rozkładem dwupunktowym (0-2).

Skala trudności zadań wykorzystanych w badaniach przedstawiona została na rysunku 1. Zadania uszeregowane zostały według stopnia trudności mierzonego w jednostkach logit. Kolejność zadań w legendzie do rysunku 1. wynika z kolejności wystąpienia zadań na osi trudność – łatwość zadania. Im wyżej na osi znajduje się zadanie, tym było ono trudniejsze. Z zadań wykorzystanych w badaniach najtrudniejsze były zadania z_21_5 oraz z_21_6, zaś zdecydowanie najłatwiejsze było zadanie z_25_1. Pod względem łatwości wyróżniały się również zadania: z_25_3, z_25_2, z_25_4, których położenie na osi trudność – łatwość zadania jest wyraźnie zbliżone do bieguna dolnego (łatwe).



Rysunek 1. Oszacowane trudności zadań otwartych w arkuszu standardowym, Sprawdzianu 2006

Rzetelność wykalibrowanej skali trudność – łatwość zadania była bliska 1, co pozwoliło uznać wyniki uczniów oszacowane za pomocą tejże skali za wiarygodne i wysoce rzetelne. Uzasadnia to również wykorzystanie zadań z tego egzaminu do sprawdzenia jakości oceniania Sprawdzianu w sesji 2006 przez zewnętrznych egzaminatorów zatrudnionych przez OKE Kraków. Do następnych analiz przyjęto założenie, że oszacowanie poziomów surowości oceniania

egzaminatorów mierzone wybranymi zadaniami (otwartymi) w wykalibrowanej skali jest wiarygodne. Dlatego też w procesie kalibracji poziomów surowości egzaminatorów za stałe i znane (takie jak w tabeli 2.) przyjęto poziomy trudności zadań. Przypomnijmy, że szacowany w kolejnym etapie model uwzględniał 3 zmienne:

1 – surowość egzaminatorów,
2 – poziom umiejętności uczniów,
3 – trudność zadań.

Przy czym trudności zadań zostały zakotwiczone na wartościach oszacowanych na wynikach 10% próby losowej wszystkich uczniów piszących Sprawdzian, a więc ich wartości nie były estymowane. Jednocześnie narzucono skalę odpowiedzi osobną dla zadań pytań punktowanych 0-1 (wspólną) oraz osobną dla zadań punktowanych 0-1-2, przy czym w tym drugim przypadku osobną dla każdego z dwóch pytań. Analizę prowadzono, stosując program Facets for Windows Version No. 3.57.0.

Ustalenia metodologiczne – wobec przyjętego modelu

Ponieważ proces oceniania organizowany jest w hierarchicznej strukturze koordynacji obejmującej ośrodki koordynacji, zespoły egzaminatorów i indywidualnych egzaminatorów, zebrane w wyniku badań dane również charakteryzowała hierarchiczna struktura. Egzaminatorzy zgrupowani byli w zespoły, a zespoły podporządkowane były odpowiednim ośrodkom koordynacji. Podczas oceniania przez każdego egzaminatora pięciu tych samych prac uczniowskich egzaminatorzy pracowali samodzielnie bez możliwości konsultowania się. Dlatego też szacując parametry modelu, przyjęto niezależność procesu oceniania egzaminatorów względem siebie w obrębie jednego zespołu. Dopiero po oszacowaniu charakterystyk egzaminatorów (poziomu ich surowości R_j) za pomocą skalowania Rascha porównano wartości wskaźnika poziomu surowości oceniania pomiędzy poszczególnymi zespołami, jak i ośrodkami koordynacji. Dzięki zastosowaniu skalowania Rascha było to możliwe i metodologicznie uprawnione, ponieważ cecha '*surowość oceniania*', choć z natury rzeczy jakościowa i nie bezpośrednio obserwowalna[5], została skwantyfikowana[6].

[5] pomiar cech ukrytych (latentnych) dokonywany jest poprzez mierzenie ich wskaźników; w przypadku oszacowania surowości oceniania egzaminatorów takimi wskaźnikami były liczby punktów przyznane przez poszczególnych egzaminatorów za każde oceniane zadanie egzaminacyjne; przyznane oceny w trakcie procedury obliczeniowej skorygowane zostały o poziom trudności poszczególnych zadań (Di) i poziom wiedzy reprezentowany przez poszczególnych uczniów (Bn).

[6] cecha 'surowość oceniania' powstała w wyniku skalowania Rascha ma charakter cechy ciągłej, interwałowej, o sztucznie ustalonym punkcie zerowym; miara tej cechy wyrażona jest w jednostkach zwanych logitem.

Analizę przeprowadzono, narzucając wartości średnich oszacowań poziomu surowości egzaminatorów równe zero – czyli $\bar{r}_i = 0$. Oszacowania trudności zadań egzaminacyjnych przyjęto za stałe i znane (oszacowane wcześniej). Natomiast oszacowania poziomu umiejętności uczniów pozostawiono bez narzucania żadnych warunków ograniczających.

Przedstawione poniżej wyniki uzyskano, wykorzystując do kalibracji obu skal – '*surowość oceniania*' oraz '*poziom umiejętności*' – wszystkie ich elementy składowe. Oznacza to, że nie usunięto z analizy uczniów, dla których uzyskane oszacowania nie spełniały któregoś z kryteriów dopasowania. Usuwania z analizy egzaminatorów oceniających w sposób bardzo nietypowy nawet nie rozważano, ponieważ to właśnie te informacje leżały w obrębie głównych zainteresowań autorów badań. Pozostawienie elementów znacząco niedopasowanych jest kwestią dyskusyjną. Jednak zdecydowano się na taki krok ze względu na fakt, że uzyskiwane w wyniku skalowania Rascha wyniki porównać można do zespołu naczyń połączonych – usunięcie jednego elementu pociąga za sobą zmianę wartości wszystkich pozostałych i co chyba ważniejsze, również zmianę ich jakości dopasowania, a tym samym dopasowania wszystkich danych do całego szacowanego modelu. Przyjęta ścieżka postępowania nie wyklucza jednak przeprowadzenia analizy w inny sposób.

Analiza wyników badań

Analiza rozkładów trzech zmiennych występujących w modelu

Zestawienie wyników przeprowadzonej analizy w kompleksowy sposób prezentuje rysunek 2. Przedstawiono na nim łącznie wartości wszystkich oszacowanych parametrów modelu. Pierwsza kolumna na rysunku (*skala*) przedstawia wspólną przedziałową skalę dla szacowanych zmiennych (jednostką jest logit), druga kolumna (*uczeń)* pokazuje poziom umiejętności uczniów – im wyżej na osi znalazł się uczeń, tym wyższe jego umiejętności. Trzecia kolumna (*egzaminator*) prezentuje rozkład charakterystyk egzaminatorów. Im wyżej ulokował się egzaminator, tym surowiej na tle grupy oceniał. Czwarta kolumna odpowiada poziomowi trudności zadań egzaminacyjnych – im wyżej na osi znalazło się zadanie, tym było trudniejsze. Dwie ostatnie kolumny prezentują punktację zastosowaną do oceny zadań ocenianych na skali trójstopniowej.

Analizując dane przedstawione na rysunku 2. warto zauważyć stosunkowo niewielkie, w porównaniu do pozostałych rozkładów, zróżnicowanie rozkładu poziomów surowości egzaminatorów. Rozkład ten jest najbardziej zwarty, co jest bardzo pożądanym wynikiem z punktu widzenia oczekiwanych cech egzaminatorów oceniających prace uczniów. Dokładniejsza analiza poziomów surowości przeprowadzona zostanie w rozdziale: Analiza surowości egzaminatorów.

Na podstawie analizy rysunku 2. można zauważyć również, że zróżnicowanie rozkładów poziomu umiejętności uczniów, jak i poziomów trudności zadań

były porównywalne, choć zdecydowanie wśród uczniów wyróżnia się zdający o kodzie A01. Uczeń ten wykonał poprawnie wszystkie zadania z tym, że w zadaniu matematycznym zastosował nietypową metodę obliczania procentów. Możemy przypuszczać, że nauczyciele, którzy nie są matematykami, mogli nie uznawać tej metody za poprawną. Rozkłady trudności zadań i poziomu umiejętności badanych uczniów wyraźnie różnią się od siebie położeniem na skali. Zdecydowanie wyżej na osi skali wykalibrowanej w jednostkach logit znajduje się rozkład poziomów umiejętności uczniów, dla którego wartość średnia wynosi 1,44 logita, podczas gdy średnia trudność zadań otwartych, wykorzystanych w badaniach, wyniosła 1,036 logita. Posługując się wartościami przeciętnymi, możemy uznać, że dla omawianej grupy 15 uczniów zestaw pytań był stosunkowo łatwy, biorąc jako punkt odniesienia ich przeciętny poziom umiejętności.

Rysunek 2. Rezultaty szacowania poziomu umiejętności uczniów, surowości egzaminatorów i trudności zadań

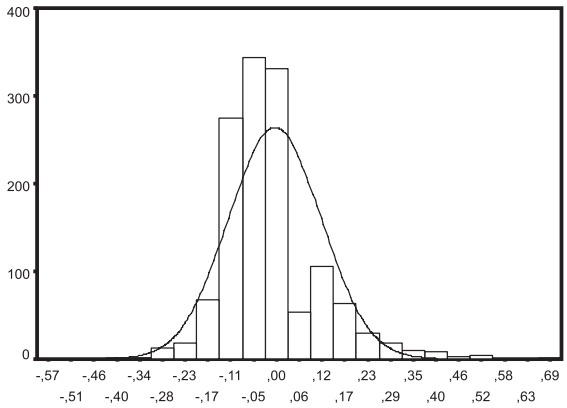
Analiza surowości egzaminatorów

Głównym celem badań było sprawdzenie zróżnicowania surowości punktowania zadań otwartych przez egzaminatorów usytuowanych w hierarchicznej strukturze moderowania procesu oceniania. Ze względu na dużą liczbę egzaminatorów biorących udział w badaniach przedstawione zostaną jedynie zbiorcze wyniki oszacowań ich surowości (tabela 3.).

Tabela 3. Zbiorcze statystyki oszacowań surowości egzaminatorów

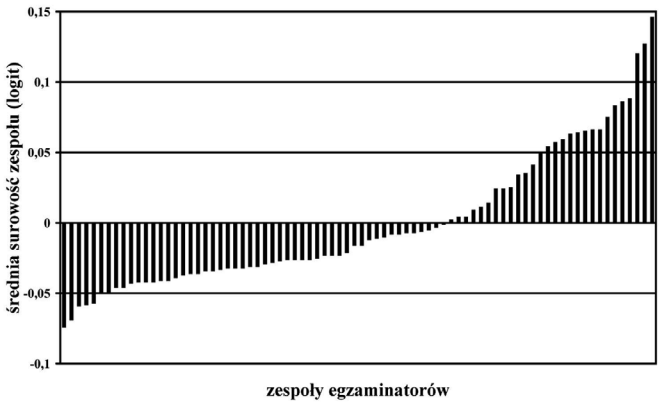
Średnia (w jednostkach logit)	-0,0013
Odchylenie standardowe	0,1168
Współczynnik asymetrii	1,244
Wartość minimalna	-0,4
Wartość maksymalna	0,74
Typowy obszar zmienności[1]	<-0,1181; 0,1155>
Obszar obserwacji nietypowych (1)[2]	<-0,4; -0,2349) oraz (0,2323; 0,74>
Obszar obserwacji nietypowych (2)[3]	<-0,4; -0,35170,3526) oraz (0,3491; 0,74>
Odsetek obserwacji w obszarze nietypowym (1)	1,3% (17 egzaminatorów) oraz 4,3% (58 egzaminatorów)
Odsetek obserwacji w obszarze nietypowym (2)	0,1% (2 egzaminatorów) oraz 1,6% (22 egzaminatorów)

Przeciętny poziom surowości egzaminatorów wyniósł -0,0013, co jest wynikiem narzucenia przy estymacji modelu warunku o średnim poziomie surowości równym 0. Oszacowane poziomy surowości mieszczą się w granicach od -0,4 logita do 0,74 logita, a więc zróżnicowanie poziomów surowości mierzone rozstępem wyniosło 1,14 logita, co jest wynikiem dobrym. Maksymalne odchylenie poziomu surowości od średniego poziomu surowości zaobserwowane zostało tylko u dwóch egzaminatorów (0,71 do 0,74 logita). Ci dwaj egzaminatorzy oceniali wybrane 5 zadań najsurowiej spośród wszystkich badanych. Najłagodniejsi egzaminatorzy zaleźli się w przedziale od -0,40 do – 0,31 logita. Warto zauważyć, że poziom surowości dla 75 (5,6%) egzaminatorów zatrudnionych przez OKE Kraków do oceny sprawdzianu w wiosennej sesji egzaminacyjnej 2006 różnił się więcej niż dwa odchylenia standardowe od średniej, zaś dla 1,7% (24) egzaminatorów o więcej niż trzy odchylenia standardowe. Wśród nietypowych zachowań przeważały odchylenia powyżej średniej, co oznacza, że nietypowe zachowania egzaminatorów występowały raczej w kierunku zbytniej surowości niż łagodności. Natomiast rozkład poziomów surowości wszystkich egzaminatorów charakteryzowała dosyć silna asymetria prawostronna Oznacza to, że wśród egzaminatorów dominowali ci, którzy oceniali łagodniej niż egzaminator o przeciętnej surowości.



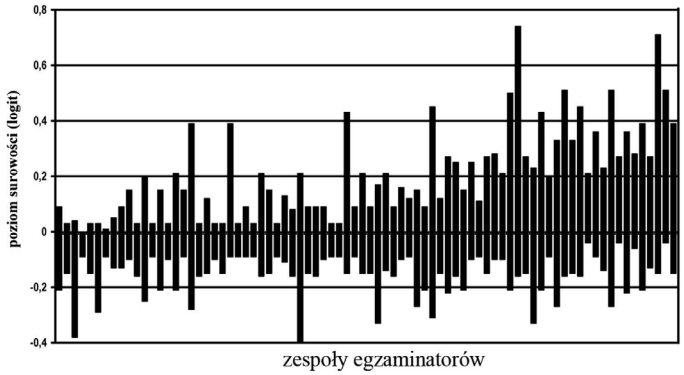
Rysunek 3. Rozkład poziomu surowości egzaminatorów

Aby dokładniej przyjrzeć się poziomom surowości egzaminatorów, dokonano porównań przeciętnych poziomów między poszczególnymi zespołami oraz ośrodkami koordynacji oceniania. Średni poziom surowości oceniania dla całego zespołu prawdopodobnie ma związek z cechami osobowościowymi przewodniczącego i ze sposobem koordynacji oceniania w jego zespole. Taka hipoteza wymaga jednak weryfikacji w dalszych badaniach.



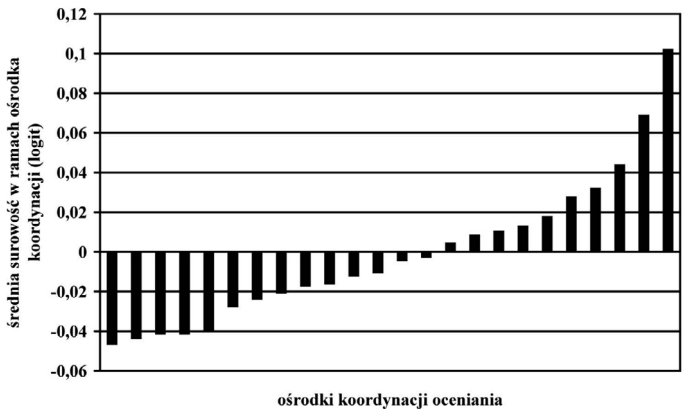
Rysunek 4. Średnie poziomy surowości zespołów egzaminatorów

Pewne światło na to zagadnienie może rzucić analiza rozstępu zmienności surowości w poszczególnych zespołach.



Rysunek 5. Rozstęp poziomu surowości egzaminatorów w poszczególnych zespołach

Jak można zauważyć, analizując rysunek 5., w niektórych zespołach rozstęp surowości egzaminatorów jest stosunkowo mały. Takie zjawisko częściej zostało zaobserwowane w zespołach o niskiej średniej surowości zespołu egzaminatorów. W przypadku jednego zespołu okazało się, że żaden z jego członków nie charakteryzował się surowością o dodatnim oszacowaniu, co więcej, poziom surowości dla każdego z nich był poniżej średniej, co oznacza, że wszyscy egzaminatorzy z tego zespołu charakteryzowali się mniejszą niż przeciętna surowością (większą łagodnością). I jeszcze na koniec rzut oka na zróżnicowanie surowości oceniania pomiędzy ośrodkami koordynacji.



Rysunek 6. Zróżnicowanie średniej surowości egzaminatorów pomiędzy ośrodkami koordynacji oceniania

Wyniki analizy regresji

W celu zbadania, w jakim stopniu poziomy surowości egzaminatorów wynikają z przynależności poszczególnych egzaminatorów do zespołów i ośrodków koordynacji, oszacowany został model regresji hierarchicznej, w którym zmienną objaśnianą była '*surowość egzaminatora*'. Oszacowany model uwzględniał jedynie efekt stały i nie zawierał żadnej zmiennej objaśniającej.

Tabela 4. Podział wariancji wskaźnika surowości oceniania na indywidualną i wynikającą z przynależności do ośrodka koordynacji oceniania i do zespołu egzaminatorów

Efekt stały	Współczynnik	Błąd standardowy	Krytyczny poziom istotności
γ_{00} – wyraz wolny	0.001039	0.007509	0.891
Efekt losowy	Wariancja	Błąd standardowy	Krytyczny poziom istotności
Wariancja na poziomie indywidualnym	0.012091	0.000480	0.000
Wariancja na poziomie grupowym – ośrodek	0.000971	0.000386	0.012
Wariancja na poziomie grupowym – zespół w ramach ośrodka	0.000525	0.000236	0.026

Przeprowadzone analizy wykazały istnienie zróżnicowania poziomów surowości egzaminatorów zarówno na poziomie zespołu, jak i na poziomie ośrodka, co sugeruje istnienie trójstopniowej hierarchii w danych. Wyniki prezentuje tabela 4. Przedstawione w niej wariancje pokazują wielkość zróżnicowania surowości egzaminatorów między sobą (wariancja na poziomie indywidualnym), między zespołami w obrębie ośrodków (wariancja poziomie drugim), między ośrodkami (wariancja na poziomie trzecim). Okazało się, że 88,99% zróżnicowania wynika z różnic indywidualnych między samymi egzaminatorami, 3,68% z przynależności do zespołu podlegającego odpowiedniemu ośrodkowi koordynacji, zaś 7,15% z podlegania ośrodkowi koordynacji.

Wartość wyrazu wolnego w tabeli 4. odpowiada przeciętnej wartości poziomu surowości egzaminatorów. Wartość ta różni się nieznacznie od średniej surowości policzonej bez uwzględniania hierarchicznej struktury danych (-0,0014). Różnica jest mała, bo zróżnicowanie na poziomach zespołów i ośrod-

ków jest niewielkie, nie ma zatem w tym przypadku większego znaczenia, w jaki sposób obliczana jest wartość przeciętna. Podkreślić trzeba jednak, że gdyby wartość współczynnika korelacji międzygrupowej była wyższa, wtedy bardziej wiarygodnym miernikiem tendencji centralnej byłaby średnia liczona z uwzględnieniem struktury hierarchicznej.

Przeprowadzone analizy pokazały, że choć egzaminatorzy nie różnili się mocno między sobą poziomami surowości w wartościach bezwzględnych, to jednak niepokojący pozostaje fakt, wykrycia wpływu przynależności do odpowiedniego zespołu i odpowiedniego ośrodka koordynacji na poziom surowości oceniania. Choć wpływ ten był niewielki, to jednak okazał się statystycznie istotny. Istnieje możliwość wyjaśnienia źródeł zróżnicowania w poziomach surowości egzaminatorów, w tym celu jednak konieczne byłoby uwzględnienie dodatkowych informacji deskryptywnych o egzaminatorach, zespołach egzaminatorów, jak i ośrodkach koordynacji a także wszelkich innych podmiotach, co do których istnieje podejrzenie, że mogły mieć wpływ na zróżnicowanie poziomu surowości oceniania.

Podsumowanie

Przeprowadzona analiza wyników badań pokazała, że choć występują różnice między poziomami surowości egzaminatorów, to całkowita rozpiętość poziomów surowości wyniosła tylko 1,14 logita. Dodać trzeba, że 98,3% badanych egzaminatorów charakteryzowało się poziomem surowości w granicach dwóch odchyleń standardowych od średniej, a więc takich, których nie uznaje się za nietypowe. Chociaż zróżnicowanie poziomów surowości było niezbyt duże, to przeważali w badanej grupie egzaminatorzy oceniający łagodniej niż egzaminator przeciętny. Np. uczeń, którego praca została zakodowana C05 powinien uzyskać wynik 11 punktów. Taki wynik przyznało tej pracy 65 procent z 431 egzaminatorów. 13 procent egzaminatorów przyznało o 1 punkt mniej i 12 procent o jeden punkt więcej. O trzy punkty mniej oceniło tę pracę tylko 2 egzaminatorów, czyli 0,5 procent. W tym zespole zarejestrowano jednocześnie15 najłagodniejszych egzaminatorów (3,5 procent), którzy za tę pracę przyznali co najmniej 3 punkty więcej, niż powinien ten uczeń uzyskać.

W badaniach sprawdzono również, czy egzaminatorzy nie byli stronniczy w procesie oceniania. Rozumowanie przeprowadzono dwukierunkowo, sprawdzając stronniczość w stosunku do uczniów, jak i do zadań. Okazało się, że najczęściej ze stronniczym (obciążonym) sposobem oceny spotkał się uczeń A01, który w zadaniu 21. w nietypowy sposób obliczał procenty. W przypadku zadań były to: z_21_5 wymagające obliczenia, jaki procent powierzchni działki przeznaczono na pasiekę, z_21_6 – obliczenie powierzchni pasieki oraz z_25_2 – polegające na wyjaśnieniu, dlaczego wybrana do opisu przez ucznia osoba zasługuje na szacunek.

Optymistyczny jest jednak fakt, że nie wykryto żadnych regularności w interakcjach egzaminator-zadanie. Oznacza to, że jeśli nawet któryś egzaminator bywał stronniczy, nie dotyczyło to żadnych wybranych zadań ani żadnych wybranych uczniów.

Jakie znaczenie mogą mieć przeprowadzone badania dla doskonalenia egzaminów zewnętrznych w Polsce? W kontekście otrzymanych wyników warto przemyśleć koordynację oceniania w skali kraju. W badaniach uczestniczyli egzaminatorzy Sprawdzianu, którym poświęcono szczególnie wiele troski w przygotowaniu ich do porównywalnego oceniania. Powstaje pytanie, jak ten problem wygląda na poziomie maturalnym i egzaminów potwierdzających kwalifikacje zawodowe? Z obserwacji poczynionych podczas wglądów maturzystów do swoich prac wynika podejrzenie, że na poziomie maturalnym, w niektórych przedmiotach zróżnicowanie surowości oceniania pomiędzy zespołami może być znacznie większe.

Aby zminimalizować efekt surowości/łagodności egzaminatora warto już dziś podjąć kilka działań, które można podzielić na 3 kategorie: szkoleniowe, organizacyjne i statystyczne.

1. Szkolenie egzaminatorów, aby byli świadomi występowania efektu surowości/ łagodności oceniania i umieli przeciwdziałać tej tendencji. Przy tej okazji warto przedstawić, z jakimi konsekwencjami dla egzaminowanych wiąże się brak kontroli tego efektu.
2. Zwrócenie uwagi na klarowne zdefiniowanie poszczególnych umiejętności, których opanowanie badano podczas egzaminu. Jeżeli to jest możliwe, powinniśmy dostarczyć jak najwięcej przykładów prac uczniowskich, które pozwoliłyby na uświadomienie, jakiemu poziomowi umiejętności odpowiadają poszczególne stopnie stosowanej skali.
3. Zgromadzenie do szkoleń przykładów prac egzaminacyjnych, które mogą egzemplifikować poszczególne stopnie skali wykorzystywanej w ocenianiu każdego zadania.
4. Skompletowanie zespołów nauczycieli oceniających w ten sposób, aby surowym egzaminatorom towarzyszyli raczej łagodniejsi weryfikatorzy i na odwrót.
5. Monitorowanie efektu egzaminatora w poszczególnych sesjach egzaminacyjnych.
6. Wykorzystanie metod statystycznych do korekty oceniania skrajnie surowych i skrajnie łagodnych egzaminatorów.

Bibliografia:

1. Guilford J. P., *Psychometric methods*, New York, McGraw, 1954.
2. Linacre J. M., *Many Facet Rasch Measurement*, Mesa Press, Chicago 1994.
3. Myford C. M., Wolfe W. W., Detecting and Measuring Rater Efects Using Many-Facet Rasch Measurement in, *Journal of Applied Measurement*, Constructing Variables Volume 5, nr 2, 2004.
4. Noizet G., Caverni J. P., *Psychologiczne aspekty oceniania osiągnięć szkolnych*., PWN, Warszawa 1988.
5. Popham W. J., Modern Educational Measurement. W, A practitioner's perspective. Englewood Cliffs, New York, Prentice Hall 1990.
6. Scullen S. E., Mount M. K., Goff M., Understanding the latent structure of job performance ratings. Journal of Applied Psychology, 85, 2000.
7. Szaleniec H., *Zastosowanie teorii analizy zadania testowego (IRT) w procesie oceniania zewnętrznego*, [w:] Niemierko B., Szmigel M. K. [red]. Teoria i praktyka oceniania zewnętrznego. IV ogólnopolska konferencja z cyklu „Diagnostyka Edukacyjna". Pandit, Kraków 2001.
8. Trevor G., Bond, Christie M. Fox., Applaying The Rasch Model, Fundamental Measurement in Human Sciences. Lawrence Erlbaum Associates, Publischer, London 2001.
9. Thorndike R. L. Hagen E. P., Measurement and evaluation in psychology and education, John Wiley and Sons, New York 1977.
10. Węziak D., *Zastosowanie wieloczynnikowego skalowania Rascha do porównania sposobu oceniania egzaminatorów,* [w:] Niemierko B., Szyling G. [red]. Holistyczne i analityczne metody diagnostyki edukacyjnej. Perspektywy informatyczne egzaminów szkolnych. XI krajowa konferencja z cyklu „Diagnostyka Edukacyjna". Fundacja Rozwoju Uniwersytetu Gdańskiego, Gdańsk 2005.

Objaśnienia:

1. Typowy obszar zmienności zdefiniowano jako obszar zawierający wyniki odchylające się od średniej o nie więcej niż jedno odchylenie standardowe.
2. Obszar obserwacji nietypowych (1) zdefiniowano jako obszar zawierający wyniki odchylające się od średniej o więcej niż dwa odchylenia standardowe.
3. Obszar obserwacji nietypowych (2) zdefiniowano jako obszar zawierający wyniki odchylające się od średniej o więcej niż trzy odchylenia standardowe.