dr Henryk Szaleniec
Okręgowa Komisja Egzaminacyjna
w Krakowie

KRZYWA INFORMACYJNA ZADAŃ JAKO NARZĘDZIE W KONSTRUOWANIU ARKUSZA EGZAMINACYJNEGO

Autor prezentuje teorię wyniku zadania testowego (*Item response theory* – IRT) jako metodę wyboru zadań do arkusza egzaminacyjnego. Atutem teorii IRT jest niezależność parametrów osiągnięć uczniów od parametrów zadań oraz niezależność parametrów zadań od próby egzaminowanych uczniów. Oznacza to, że bazująca na IRT analiza zadań, po ich próbnym zastosowaniu, może pomóc wykryć usterki zadań, które mogłyby ujawnić się dopiero na egzaminach właściwych. Daje to ramę do budowania arkuszy, przy pomocy których precyzja oszacowania osiągnięć byłaby dostosowana do danej populacji uczniów.

Jednym z kluczowych zadań okręgowych komisji egzaminacyjnych jest przygotowanie wysokiej jakości arkuszy egzaminacyjnych. Konstruowanie arkusza egzaminacyjnego zgodnie z klasyczną teorią testu, obejmuje między innymi wybór zadań zgodnie z pożądanymi ich właściwościami pomiarowymi, takimi jak trafność wewnętrzna, trudność i moc różnicująca. Poziom trudności zdeterminowany jest zwykle poprzez cel arkusza egzaminacyjnego i przewidywany rozkład badanej umiejętności w egzaminowanej populacji. Jeżeli chodzi o moc różnicującą, to generalnie w egzaminach, których wynik wykorzystywany jest jako jeden z czynników selekcyjnych do szkoły wyższego szczebla, bardziej pożądane są zadania, które pozwalają silniej zróżnicować uczniów ze względu na poziom badanej umiejętności niż zadania słabo różnicujące.

Obliczone na podstawie próby standaryzacyjnej charakterystyki zadań zgodnie z klasyczną teorią testu nie są inwariantami (nie są niezmiennikami, gdy wybierzemy inną próbę otrzymamy inne wartości) w populacji, w której poziom badanej umiejętności zmienia się i to w szerokim zakresie. Tak więc, sukces właściwego doboru zadań do arkusza zależy od tego, w jakim stopniu próba standaryzacyjna arkusza egzaminacyjnego jest podobna do populacji pod względem badanej umiejętności.

W praktyce rzadko udaje się zapewnić takie warunki, aby grupa uczniów, na której standaryzujemy arkusze egzaminacyjne, była podobna do populacji, w której przeprowadzany jest egzamin. Jeżeli próba standaryzacyjna nie najlepiej reprezentuje populację, to określone na niej parametry zadań będą dalekie od tego, jakie później otrzymamy podczas egzaminu. W takim przypadku mamy małe prawdopodobieństwo, aby arkusz egzaminacyjny, wykorzystujący taką technikę wyboru zadań, miał zamierzone i określone na etapie planowania egzaminu własności. Inny aspekt tego samego problemu to czas przeprowadzania standaryzacji arkusza egzaminacyjnego. Jeżeli badania standaryzacyjne zadań przeprowadzimy na przykład we wrześniu, a egzamin jest w maju, to musimy pamiętać o tym, że poziom umiejętności przewidzianych na egzamin jest całkiem inny w populacji i próbie standaryzacyjnej. Podobnie wygląda sytuacja, gdy testujemy zadania w okresie sesji egzaminacyjnej, ale na populacji o rok młodszych uczniów.

Innym problemem, który trudno przezwyciężyć, jest fakt, że parametry zadań, takie jak moc różnicująca i ich łatwość, oszacowane zgodnie z klasyczną teorią testu, zależą istotnie od charakterystyk innych zadań, które stanowiły arkusz egzaminacyjny. W konsekwencji, dysponując zadaniami opisanymi zgodnie z klasyczną teorią testu, trudno jest stworzyć bank zadań, który byłby pożądanym zapleczem dla autorów arkuszy egzaminacyjnych. Parametry zadań przekazanych do banku są określone na podstawie ich zastosowania w kontekście innych zadań, które tworzyły arkusz egzaminacyjny. Zadania te w innym arkuszu będą miały prawdopodobnie inne parametry.

Kolejny aspekt tego samego problemu wiąże się z tym, że do naszego banku zadań, którym posługują się autorzy arkuszy egzaminacyjnych, trafiają czasem zadania, które były zastosowane w testach wykorzystywanych do innych celów niż egzamin, który planujemy przeprowadzić. Możemy się spodziewać, że określone na podstawie tych zastosowań parametry zadań będą znacznie odbiegać od tych, które przyjmą zadania podczas egzaminu.

Dalszy problem wiąże się z zapewnieniem pożądanej rzetelności dla arkusza egzaminacyjnego. Na podstawie parametrów zadań, określonych tylko z zastosowaniem klasycznej teorii testu, nie możemy przewidzieć stopnia precyzji oszacowania umiejętności mierzonej arkuszem egzaminacyjnym przed jego zastosowaniem. Mając do dyspozycji kilka arkuszy do wyboru, nie jesteśmy w stanie określić, który z nich pozwoli na dokładniejszy pomiar osiągnięć egzaminowanych uczniów.

W budowanym od kilku lat polskim systemie egzaminacyjnym nie dysponujemy jeszcze profesjonalnie opracowanymi bankami zadań, np. takimi, jakimi dysponują wyższe uczelnie medyczne. Najczęściej do każdego egzaminu autorzy próbują tworzyć nowe autorskie zadania, których parametry mogą być tylko intuicyjnie przewidywane. Ze względu na konieczność zachowania niejawności zadań niemożliwa jest także standaryzacja zadań na reprezentacyjnych losowych próbach.

Rozwijająca się od pół wieku teoria odpowiedzi na zadania testowe (ang. *Item response theory* - IRT), która w ostatnich latach cieszy się coraz większym zainteresowaniem konstruktorów testów, zdaje się oferować skuteczniejsze metody wyboru zadań do arkusza egzaminacyjnego niż ma to miejsce w przypadku klasycznej teorii.

PODSTAWOWE ZAŁOŻENIE IRT

U podstaw teorii odpowiedzi na zadanie testowe leżą trzy założenia:

1. wymiarach przestrzeni badanej cechy (cech) za pomocą zadań tworzących test,
2. lokalnej niezależności zadań,
3. krzywej charakterystycznej zadania testowego.

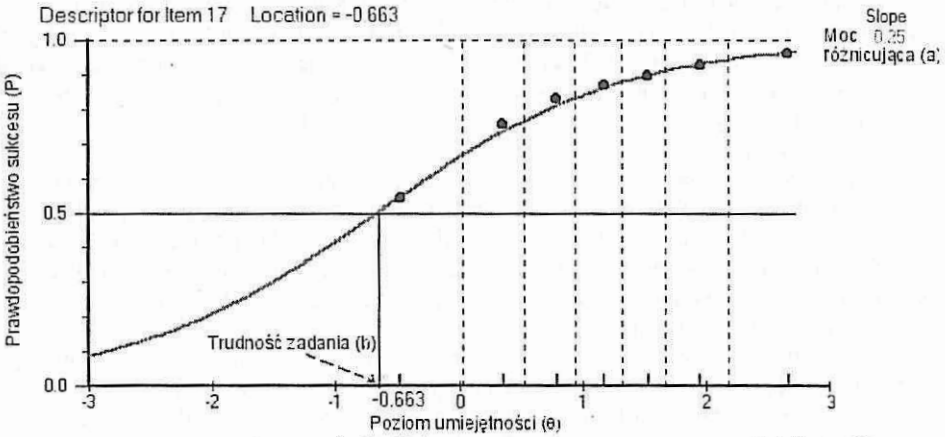
Zgodnie z pierwszym założeniem, otrzymany na egzaminie wynik może być wyjaśniony lub przewidziany na podstawie szeregu czynników nazywanych cechą lub ukrytą cechą – w naszym przypadku są to umiejętności egzaminowanego ucznia.

W skrócie można powiedzieć, że pierwsze założenie jest spełnione, gdy odpowiedzi ucznia zależą jedynie od pojedynczej lub złożonej cechy ucznia (umiejętności badanej arkuszem egzaminacyjnym). W praktyce wymaga to wyspecyfikowania wszystkich umiejętności, od których zależy powodzenie w rozwiązaniu danego zadania.

Założenie drugie, dotyczące lokalnej niezależności zadań, jest spełnione, gdy odpowiedzi ucznia na wybrane zadanie nie zależą od jego odpowiedzi na inne zadania. Założenie to wydaje się niezgodne z naszą pomiarową intuicją, gdyż zadania mierzące tę samą umiejętność będą korelowały ze sobą. Założenie to generalnie nie jest spełnione, gdy np. poziom rozwiązania zadania o bogatej fabule badającego umiejętności matematyczne zależy istotnie od umiejętności czytania tekstu a rozkład tej umiejętności nie jest znany w populacji. Uczniowie o wysokich umiejętnościach matematycznych, ale kiepsko czytający, słabo roz-

wiążą takie zadania. Wyniki tych zadań będą ze sobą skorelowane, gdyż mierzą tę samą cechę, która nie jest wyszczególniona w przestrzeni badanych umiejętności.

Trzecie założenie dotyczy związku pomiędzy jakością rozwiązania zadnia przez egzaminowanego a umiejętnościami koniecznymi do osiągnięcia sukcesu. Związek ten może być opisany monotonicznie rosnącą funkcją nazywaną funkcją charakterystyczną zadania (ang. *item characteristic curve* - ICC). Zgodnie z funkcją charakterystyczną zadania, wraz ze wzrostem poziomu umiejętności egzaminowanego rośnie prawdopodobieństwo poprawnej odpowiedzi na zadanie. Rysunek 1 przedstawia funkcję charakterystyczną zadania wraz z empirycznym rozkładem statystycznym badanej umiejętności dla siedmiu grup egzaminowanych o wzrastającym poziomie umiejętności. Wybór liczby grup należy do przeprowadzającego analizę i zależy najczęściej od wielkości populacji i przedziału zmienności poziomu umiejętności egzaminowanych.



Rysunek 1. Krzywa charakterystyczna zadania. Siedem punktów na wykresie to wynik średni dla uczniów należących do poszczególnych grup

Na osi pionowej przedstawione jest prawdopodobieństwo sukcesu uczniów rozwiązujących to zadanie. Oś pozioma wykresu przedstawia jednocześnie poziom umiejętności (θ) i trudność zadania (b). Poziom umiejętności i trudność zadań mogą przyjmować zarówno ujemne, jak i dodatnie wartości. Teoretycznie poziom umiejętności może się rozciągać od – nieskończoności do +nieskończoności. W praktyce poziom badanych umiejętności zwykle mieści się w przedziale od –3 do +3. Podobnie jest, jeżeli chodzi o trudność zadań. Zadania o trudności -3 to zadania bardzo łatwe. Zadania, dla których parametr trudności przyjmuje wartość b=+3, są zadaniami bardzo trudnymi. Zwykle trudność zadań jest tak skalowana, aby średnia była równa 0 a odchylenie standardowe 1. Tak więc, zadania dla których parametr b (trudność) ma wartość 0 są zadaniami o średniej trudności.

Jak można zauważyć na rysunku 1, egzaminowani o wyższym poziomie umiejętności badanej zadaniem mają wyższe prawdopodobieństwo poprawnej odpowiedzi niż uczniowie o niższym poziomie umiejętności niezależnie, do której grupy należą. Zależność ta nie jest zależnością prostoliniową. Przedstawia ją krzywa logistyczna kształtem przypominająca literę S.

Istnieje wiele modeli w ramach teorii odpowiedzi na zadanie testowe (IRT), różniących się matematycznym przedstawieniem funkcji charakterystycznej zadania czy liczbą parametrów tworzących model. Każdy z modeli stosuje jeden lub więcej parametrów opisujących zadanie oraz jeden lub więcej parametrów opisujących egzaminowanego. Analizy

prowadzone w Okręgowej Komisji Egzaminacyjnej z zastosowaniem programu komputerowego Rumm wykorzystują trójparametryczny model logistyczny sformułowany po raz pierwszy przez A. Birnbauma. Zgodnie z modelem prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie egzaminacyjne przez ucznia o poziomie umiejętności θ może być przedstawione funkcją logistyczną:

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+e^{-Da_i\,(\theta-b_i)}} \qquad (1),$$

gdzie:

b – parametr trudności zadania,
a – moc różnicująca,
c – współczynnik zgadywania odpowiedzi,
D – stała skalowania równa 1,7,
e – podstawa logarytmów naturalnych,
θ - poziom badanej cechy (umiejętności).

Jeszcze raz podkreślmy, że inaczej niż w klasycznej teorii pomiaru trudność zadań wyrażona jest w tej samej skali, co poziom badanej umiejętności i może przyjmować wartości zarówno ujemne jak i dodatnie.

Pierwszym krokiem analizy niezależnie od wybranego modelu jest oszacowanie parametrów zadania (a,b,c).

Wybrany do analizy model może być użyty lub nie do zastosowania względem zbioru danych empirycznych. Oznacza to, że model może niewłaściwie przewidywać i wyjaśniać wyniki egzaminu. Dlatego jednym z najważniejszych kroków podczas stosowania teorii analizy zadania testowego do oceny wyników egzaminów jest oszacowanie, czy wybraliśmy właściwy model i czy w ogóle możemy zastosować IRT do analizy naszych danych. Procedury szacowania czy dane empiryczne spełniają wymagania danego modelu, zwykle są integralną częścią programów komputerowych umożliwiających praktycznie stosowanie IRT.

Kiedy model właściwie opisuje dane empiryczne, uzyskujemy opis szeregu istotnych cech pojedynczych zadań, jak i egzaminowanych uczniów, które są pożądane dla pomiaru dydaktycznego.

Jeżeli arkusz egzaminacyjny został trafnie przygotowany do danego egzaminu oraz gdy wyniki egzaminu spełniają założenia wybranego modelu, to po pierwsze, oszacowany poziom umiejętności egzaminowanego jest niezależny od zastosowanego arkusza egzaminacyjnego oraz oszacowane parametry zadania są niezależne od grupy egzaminowanych danym arkuszem egzaminacyjnym. Innymi słowy poziom umiejętności oszacowany na podstawie różnych zbiorów zadań mierzących tę samą umiejętność jest w granicach błędu pomiarowego taki sam. Po drugie, parametry zadania oszacowane na postawie różnych grup egzaminowanych są takie same w granicach błędu związanego z wyborem próby. Możemy więc powiedzieć, że dla IRT parametry opisujące zadanie, jak i parametry opisujące poziom osiągnięć ucznia są inwariantem (niezmiennikiem). Niezależność parametrów osiągnięć egzaminowanego od parametrów zastosowanych zadań oraz niezależność parametrów zadań od wyboru próby egzaminowanych uczniów jest koronnym atutem teorii odpowiedzi na zadanie testowe. Ta niezależność osiągana jest poprzez wykorzystanie informacji o zadaniach do oszacowania poziomu osiągnięć uczniów, a także poprzez symultaniczne wykorzystanie informacji o osiągnięciach uczniów do oszacowania parametrów zadań.

Najbardziej obiecującą dziedziną teorii odpowiedzi na zadanie testowe (IRT) do praktycznego wykorzystania przez okręgowe komisje egzaminacyjne wydaje się możliwość selekcji zadań do arkusza egzaminacyjnego w zależności od ilości informacji, jaką dane zadanie dostarcza do oszacowania poziomu badanej umiejętności w różnych przedziałach skali. Tak

więc, dysponując dostępem do banku zadań stworzonym na bazie IRT, można zbudować arkusz o zadanej charakterystyce funkcji informacyjnej.

INFORMACJA, JAKIEJ MOŻE DOSTARCZYĆ ZADANIE EGZAMINACYJNE

Kiedy mówimy, że dysponujemy informacją, to mamy na myśli, iż wiemy coś szczególnego o danym obiekcie lub na dany temat. Mówimy, że wiemy bardziej lub mniej dokładnie, jakie umiejętności posiada uczeń i jakie zadania z danej dziedziny sprawią mu trudność, a jakie - nie. Ten jakościowy aspekt informacji ma kolosalne znaczenie w pomiarze dydaktycznym i stanowi szerokie pole dociekań egzaminatorów zatrudnionych w okręgowych komisjach egzaminacyjnych.

W tym artykule szczególny nacisk chciałbym położyć jednak na ilościowy aspekt informacji. Zajmiemy się poszukiwaniem odpowiedzi na pytanie, jak zmierzyć ilość informacji dostarczanej, np. przez poszczególne zadania arkusza egzaminacyjnego i jak wykorzystać tę wiedzę do tworzenia lepszych arkuszy egzaminacyjnych. Podstawy ilościowej teorii informacji stworzył C. Shannon już w 1948 roku. Za punkt wyjścia do sformułowania miary informacji przyjął miarę niepewności.

Jeżeli potrafimy zapisać ilościowo niepewność o jakimś wydarzeniu, to otrzymanie komunikatu o tym wydarzeniu zmniejsza naszą niepewność o nim. Można więc powiedzieć, że ilość informacji I zawarta w komunikacie K o jakimś zdarzeniu Z równa jest różnicy pomiędzy początkową niepewnością na temat zdarzenia Z, a niepewnością, jaka nadal pozostaje po nadejściu komunikatu K.

Odnieśmy to do dziedziny, którą zajmujemy się w Okręgowej Komisji Egzaminacyjnej. Komunikatem jest rozwiązanie przez ucznia zadania zawartego w arkuszu egzaminacyjnym. Niepewność początkowa to założony lub domniemany stan umiejętności ucznia z dziedziny, której dotyczy pomiar. Niepewność końcowa to niepewność, która nadal pozostaje po sprawdzeniu oraz ocenie zadania i rozwiązywanego przez ucznia na egzaminie. Bardziej formalnie tę myśl możemy zapisać następująco:

$$I(Z \mid K) = H(Z) - H(Z \mid K) \qquad (2),$$

gdzie $H(Z)$ oznacza pierwotną (początkową) niepewność dotyczącą zdarzenia Z. Natomiast $H(Z \mid K)$ jest niepewnością, jaka pozostaje nadal mimo odebrania komunikatu K.

Tworząc arkusz egzaminacyjny, wiemy ogólnie, dla jakich uczniów go tworzymy. Wiemy, jaki rodzaj edukacji mają za sobą, jak zróżnicowani są uczniowie w badanej populacji, ale nie wiemy, jaki poziom umiejętności z danej dziedziny posiada każdy z nich. Czyli niepewność $H(Z)$, to przedegzaminacyjna niepewność dotycząca wyników uczniów, których czeka sesja egzaminacyjna. Po egzaminie znamy wyniki uczniów. Znamy wynik każdego ucznia, który przystąpił do egzaminu. Niepewność zmalała, chociaż nadal istnieje na poziomie $H(Z \mid K)$. Różnica pomiędzy niepewnością co do umiejętności ucznia przed aktem rozwiązania zadania i po otrzymaniu wyniku oceny rozwiązania tego zadania stanowi ilość informacji $I(Z \mid K)$, jaką dostarczyło to zadanie o poziomie umiejętności egzaminowanego.

C. Shannon założył, że miara niepewności dotycząca zdarzenia wiąże się z prawdopodobieństwem tego zdarzenia $P(Z)$, formułując trzy następujące postulaty:

1. Niepewność zdarzenia wynosi 0 $H(Z)=0$, jeżeli prawdopodobieństwo danego zdarzenia (stan) wynosi 1 $P(Z)=1$.
2. Im mniejsze prawdopodobieństwo danego zdarzenia czy stanu, tym większa jego niepewność. Jeżeli $P(Z_1) < P(Z_2)$ to $H(Z_1) > H(Z_2)$.
3. Jeżeli zdarzenie Z jest złożeniem dwóch niezależnych zdarzeń Z_a i Z_b, to niepewność takiego zdarzenia jest równa sumie niepewności zdarzeń składowych.

Na podstawie tych postulatów można przyjąć, że matematyczna formuła opisująca niepewność przyjmuje następującą postać:

$$H(Z) = -\log P(Z) \qquad (3)$$

Do rozstrzygnięcia została nam kwestia jednostek niepewności. Jednostki te będą zależały od podstawy logarytmu, jaką przyjmiemy w równaniu. Jeżeli przyjmiemy logarytmy przy podstawie 2, będziemy mieli do czynienia ze znaną powszechnie jednostką **bit**, która jest skrótem od angielskiej nazwy *binary information unit*. Jeżeli do definiowania informacji skorzystamy z logarytmów dziesiętnych, to będziemy mówić o **ditach** informacji (*digital information unit*), jeżeli natomiast posłużymy się logarytmami naturalnymi, to informacje wyrażać będziemy w **nitach** (*natural information unit*).

W statystyce i w psychometrii pojęcie informacji ma podobny zakres semantyczny. F.A. Fisher definiuje informację jako odwrotność precyzji, z jaką oszacowany jest jakiś parametr. Tak więc, jeżeli dany parametr możemy z dużą dokładnością oszacować, to wiemy więcej o wartości tego parametru niż w przypadku mniej dokładnego oszacowania. Statystycznie, dokładność, z jaką dany parametr jest oszacowany, jest miarą zmienności w otoczeniu tego parametru. Stąd, miarą precyzji oszacowania jest wariancja estymatora, który możemy oznaczyć σ^2. Ilość informacji o poziomie osiągnięć będziemy oznaczać poprzez ***I***. Tak więc możemy zapisać równanie:

$$I(\theta) = \frac{1}{\sigma^2} \qquad (4)$$

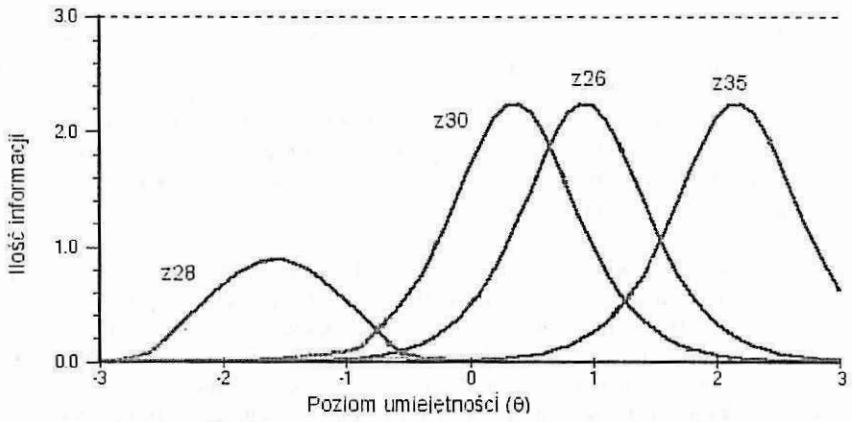
Zadaniem okręgowej komisji egzaminacyjnej jest oszacowanie, najdokładniej jak to tylko jest możliwe, umiejętności ucznia badanego arkuszem egzaminacyjnym. Badana umiejętność nie jest obserwowalna bezpośrednio. Z tego powodu często nazywamy ją zmienną nieobserwowalną bezpośrednio. Bezpośrednio obserwujemy zapisy rozwiązań zadań egzaminacyjnych, przy których dana umiejętność w większym lub mniejszym stopniu była konieczna. Jak już wcześniej wspominano, parametr, reprezentujący umiejętność, oznaczamy grecką literą θ Parametr ten chcemy z jak największą precyzją określić. Wielkość, która opisuje, czyli przybliża wartość badanej umiejętności θ na podstawie wyników egzaminu nazywana jest estymatorem danej umiejętności. Oznaczymy ją $\hat{\theta}$ (theta z daszkiem). Jak widać z równania *(4)*, ilość informacji dla danej wartości umiejętności θ jest równa odwrotności wariancji błędu oszacowania badanej cechy (umiejętności). Równanie (4) możemy zapisać także w postaci:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad (4a)$$

SE(θ) oznacza błąd standardowy oszacowania.

Jeżeli ilość dostarczonej przez zadania informacji dla danego poziomu umiejętności θ jest duża, znaczy to, że prawdziwa wartość umiejętności ucznia może być oszacowana z dużą precyzją (mały błąd standardowy oszacowania). Oznacza to także, że wszystkie estymatory umiejętności danego ucznia będą bardzo blisko jego prawdziwego poziomu umiejętności. Jeżeli ilość informacji dostarczana przez wynik egzaminowania jest mała, to estymatory rozrzucone są dosyć daleko wokół prawdziwej wartości mierzonej. Korzystając z przedstawionego w dalszej części tekstu równania (7 lub 8) możemy oszacować ilość informacji, jaką dostarcza dane zadanie w całym zakresie umiejętności począwszy od $-\infty$ do $+\infty$ (w praktyce jest to najczęściej przedział od –3 do +3). Ponieważ poziom umiejętności jest

zmienną ciągłą, także informacja, jaką dostarcza zadanie, będzie wielkością ciągłą. Rysunek 2. ilustruje funkcje informacyjne dla zadań wybranych z matematyczno-przyrodniczego arkusza egzaminacyjnego zastosowanego w wiosennej sesji egzaminacyjnej 2002.



Rysunek 2. Funkcje informacyjne czterech zadań wybranych z matematyczno-przyrodniczego arkusza gimnazjalnego

Jak można odczytać z rysunku, każde z zadań dostarcza maksimum informacji dla innego poziomu umiejętności.

Wszystkie cztery zadania to zadanie otwarte. Za zadanie 28 egzaminowany mógł maksymalnie otrzymać 2 punkty a za pozostałe trzy zadania 3 punkty. Zadania te różnią się trudnością **b**, mocą różnicującą **a** oraz ilością informacji, którą mogą dostarczyć do pomiaru umiejętności uczniów o różnym poziomie osiągnięć. Parametry poszczególnych zadań przedstawione są w tabeli 1.

Tabela 1. Parametry wybranych zadań z matematyczno-przyrodniczego arkusza gimnazjalnego

Zadanie	Maksymalna liczba punktów za zadanie	Parametry zadań zgodnie z IRT			Parametry zadań zgodnie z klasyczną teorią testu	
		Trudność (logits)	Moc różnicująca	Maksimum funkcji informacyjnej	Łatwość (p)	Moc różnicująca (r_{pb})
26	3	+0,934	1,32	2,24	0,53	0,72
28	2	-1,576	0,74	0,95	0,95	0,27
30	3	+0364	0,54	2,24	0,58	0,43
35	3	+2.104	0,142	2,25	0,22	0,60

Analizując rysunek 2 widzimy, że funkcje informacyjne dla zadań mają swoje maksimum dla poziomu umiejętności odpowiadającemu trudności zadania.

W ogólnym przypadku dla modelu trójparametrycznego poziom umiejętności, dla którego przypada maksimum informacji θ_{max} wyraża się wzorem:

$$\theta_{max} = b_i + \frac{1}{Da_i} \ln[0{,}5(1+\sqrt{1+8c_i})] \qquad (5)$$

Jeżeli parametr zgadywania c_i równy jest 0, to $\theta_{max} = b_i$, czyli tak, jak to ma miejsce w przypadku analizowanych zadań.

Dla zadania 26. poziom trudności wynosi +0,934 logits (location = +0,934). Maksimum funkcji informacyjnej wynosi 2,24.

$$I(\theta_{max}=+0{,}934) = 2{,}24 \qquad (6)$$

Jeżeli oddalamy się od lokalizacji +0,934 w lewo lub w prawo, to obserwujemy spadek wartości funkcji informacyjnej. W przedziale umiejętności $+0{,}2<\theta<+1{,}8$ wartość funkcji informacyjnej jest większa od 1. W tym przedziale, jak możemy obliczyć, poziom umiejętności mierzony tym zadaniem może być oszacowany ze stosunkowo dużą precyzją. Poza tym przedziałem precyzja oszacowania poziomu umiejętności uczniów szybko maleje. Tak więc, funkcja informacyjna mówi nam, jak dokładnie może być oszacowany określony poziom umiejętności badanych danym zadaniem. Jeżeli poziom umiejętności ucznia pokrywa się z poziomem trudności zadania lub nieznacznie od niego odbiega, to oszacowanie umiejętności ucznia badanej tym zadaniem jest najbardziej precyzyjne. Warto tutaj pamiętać, że funkcja informacyjna nie zależy od rozkładu wyników uczniów badanych tym zadaniem, a jedynie od parametrów charakteryzujących dane zadanie takich, jak jego trudność i moc różnicująca czy parametr zgadywania odpowiedzi.

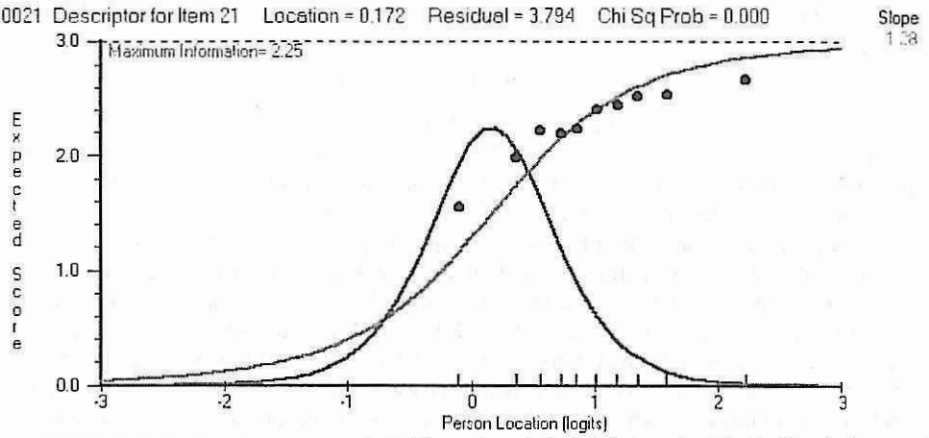
Przyjrzyjmy się teraz funkcji informacyjnej zadania 28. Za zadanie to egzaminowany mógł otrzymać maksymalnie 2 punkty. Jest to zadanie łatwe o poziomie trudności −1,576 i mocy różnicującej 0,74.

Zadanie to najwięcej informacji dostarcza o uczniach, dla których poziom umiejętności lokalizuje się na skali w punkcie $\theta = -1{,}576$ logits, a więc w obszarze gdzie funkcja informacyjna zadania 26 przyjmuje wartość bliską 0 (por. rysunek 2). Zadanie punktowane od 0 do 2 i mocy różnicującej 0,74 dostarcza znacznie mniej informacji w punkcie swojego maksimum niż poprzednie zadanie $I(\theta_{max}=-1{,}576) = 0{,}95$.

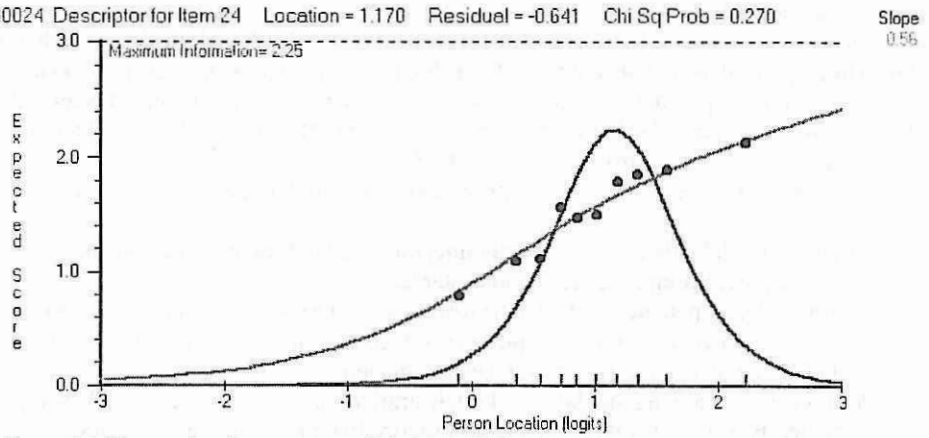
Na podstawie przeanalizowanego przykładu możemy podsumować pierwszą część rozważań.

1. Po pierwsze, dokładność oszacowania mierzonej cechy θ przez dane zadanie istotnie zależy od ilości informacji, jakiej ono dostarcza.
2. Po drugie, jeżeli parametr zgadywania równy jest 0, każde zadanie najwięcej informacji (maksimum funkcji informacyjnej $I(\theta)$.) dostarcza dla wartości skali umiejętności dokładnie odpowiadającej trudności danego zadania.
3. Po trzecie, im bardziej smukły jest kształt graficznego przedstawienia funkcji informacyjnej, tym węższy zakres umiejętności może być precyzyjne oszacowany na podstawie tego zadania.
4. Po czwarte, im więcej punktów surowych mógł uzyskać uczeń za dane zadanie oraz im większa jest jego moc różnicująca, tym większa jest maksymalna wartość funkcji informacyjnej danego zadania.

Zanim przejdziemy do matematycznego opisu funkcji informacyjnej zadania, rozważmy jeszcze przykłady dwóch zadań z maturalnego arkusza egzaminu z geografii, za których rozwiązanie uczeń mógł uzyskać maksymalnie po 3 punkty.



Rysunek 3. Krzywa charakterystyczna zadania 21 oraz jego funkcja informacyjna. Trudność zadania wynosi +0,172 a ogólna moc różnicująca 1,28



Rysunek 4. Krzywa charakterystyczna zadania 24 oraz jego funkcja informacyjna. Trudność zadania wynosi +1,170 a ogólna moc różnicująca 0,56

Zadania te mają podobne funkcje informacyjne, ale w odmienny sposób różnicują uczniów. Z przedstawionej na rysunku 3 charakterystyki zadania można odczytać, że zadanie 21 silnie różnicuje uczniów. Parametr określający moc różnicującą a=1,28. Natomiast zadanie 24. o wiele słabiej różnicuje uczniów (a=0,56). Zwróćmy jednak uwagę, że zadanie 21 silnie różnicuje uczniów o poziomie umiejętności rozciągających się na skali od –0,8 do +0,8. Poza tym obszarem zadanie to różnicuje bardzo słabo. Inaczej jest w przypadku zadania 24, które, ma małą moc różnicującą (slope = 0,54), ale z taką mocą różnicuje ono uczniów w szerokim obszarze skali umiejętności od –1 do +2. W całym tym obszarze krzywa charakterystyczna zadania 24 z dobrym przybliżeniem przypomina linię prostą. Mamy tu do czynienia z pewnym paradoksem. Zadanie o dużym współczynniku mocy różnicującej różnicuje w wąskim zakresie poziomu badanej cechy (w naszym przypadku umiejętności geograficznych). Paradoks ten nazywany jest bandwidth paradox. Może się więc okazać, że zadanie o wysokim wskaźniku mocy różnicującej różnicuje uczniów w obszarze skali, który z punktu widzenia zadań egzaminu jest mało przydatny.

Fakt ten musimy brać pod uwagę, kiedy kolekcjonujemy zadania do arkusza egzaminacyjnego. W wielu wypadkach konieczny będzie kompromis pomiędzy zadaniami o dużej

mocy różnicującej, ale w wąskim przedziale skali umiejętności, a zadaniami o niższym wskaźniku mocy, ale za to w bardzo szerokim przedziale skali.

JAK FUNKCJA INFORMACYJNA ZADANIA ZALEŻY OD INNYCH JEGO PARAMETRÓW?

Związek pomiędzy wartością informacyjną zadania a innymi jego parametrami może być opisany z różnych perspektyw. Można przyjąć, że jest to stosunek mocy różnicującej zadania w punkcie, w którym krzywa charakterystyczna zadania odpowiada jego trudności (slope) do przewidywanego teoretycznie błędu pomiaru umiejętności w tym punkcie (czyli w punkcie odpowiadającym trudności zadania). Matematycznie (za Hambleton i Swaminathan) funkcja informacyjna zadania *i*, za rozwiązanie którego egzaminowany może uzyskać 0 lub 1 może być opisana w następujący sposób:

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta) \bullet Q_i(\theta)} \qquad (7),$$

gdzie:

$I_i(\theta)$ – funkcja informacyjną zadania *i*,

$P_i'(\theta)$ – pierwsza pochodna (nachylenie) krzywej charakterystycznej zadania w punkcie skali umiejętności odpowiadającym trudności zadania,

$P_i(\theta)$ – prawdopodobieństwo sukcesu w zadaniu *i*,

$Q_i = 1 - P_i\theta)$ – prawdopodobieństwo błędnej odpowiedzi w zadaniu *i*.

W trójparametrycznym logistycznym modelu, uwzględniającym zgadywanie równanie przyjmuje postać:

$$I_i(\theta) = D^2 a^2{}_i \bullet \frac{[1-c_i]}{[c_i + e^{D(\theta - b_i)}][1 + e^{-Da_i(\theta - b_i)}]} \qquad (8)$$

Znając parametry zadania można obliczyć wartości funkcji informacyjnej dla interesującego nas przedziału poziomu umiejętności θ.

FUNKCJA INFORMACYJNA ARKUSZA EGZAMINACYJNEGO

Arkusz egzaminacyjny jest zbiorem celowo przygotowanych do tego egzaminu zadań. Zadania powinny być dobierane tak, aby arkusz egzaminacyjny posiadał pożądane właściwości pomiarowe. Innymi słowy chodzi o to, aby za jego pomocą można było oszacować umiejętności uczniów z danej dziedziny z dostateczną precyzją. Jeżeli znamy funkcje informacyjne zadań tworzących arkusz egzaminacyjny, to stosunkowo łatwo możemy oszacować funkcję informacyjną całego arkusza, gdyż jest ona prostą sumą funkcji informacyjnych poszczególnych zadań. Matematycznie możemy to wyrazić w następujący sposób:

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta) \qquad (9),$$

gdzie:

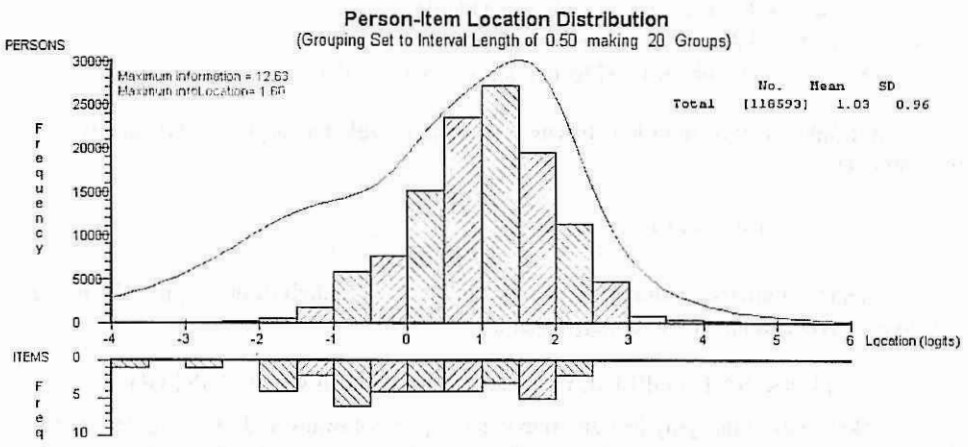
$I(\theta)$ – oznacza ilość informacji dostarczanej przez test o uczniu o poziomie umiejętności θ,

$I_i(\theta)$ – określa ilość informacji dostarczanej przez zadanie *i* o uczniu, którego poziom umiejętności wynosi θ

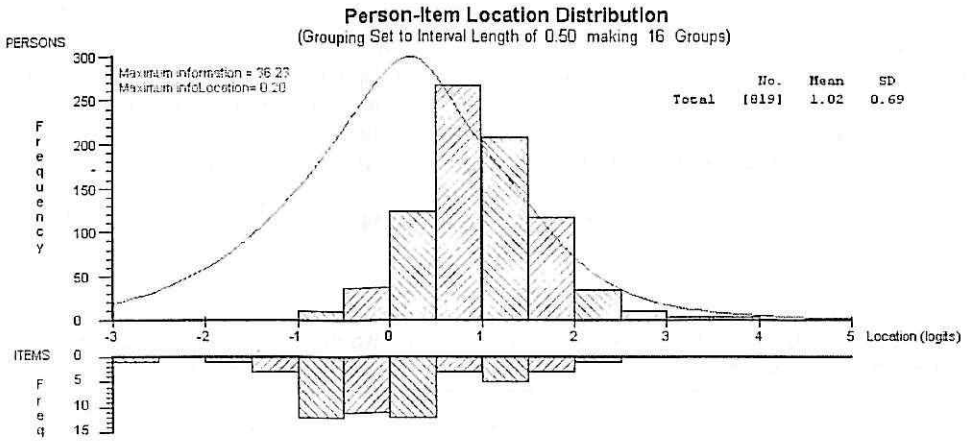
N jest liczbą zadań, z których składa się arkusz egzaminacyjny.

Maksymalna wartość funkcji informacyjnej testu jest istotnie większa niż pojedynczego zadania. Tak więc, za pomocą całego arkusza można znacznie precyzyjniej oszacować poziom umiejętności uczniów, niż to ma miejsce w przypadku pojedynczego zadania. Idealny arkusz egzaminacyjny powinien mieć krzywą informacyjną, która jest linią prostą równoległą do osi umiejętności (osi θ). Wtedy precyzja oszacowania umiejętności uczniów byłaby jednakowo wysoka dla całego przedziału umiejętności charakterystycznego dla danej populacji. Arkusze egzaminacyjne stosowane w trakcie próbnych i właściwych egzaminów w roku szkolnym 2001/2002 nie były budowane z wykorzystaniem wiedzy o funkcji informacyjnej zadań. Arkusze budowane były najczęściej z oryginalnych autorskich zadań o nieznanych charakterystykach funkcji informacyjnej, gdyż obecnie w tworzonym systemie egzaminacyjnym nie dysponujemy bankami zadań z oszacowanymi i wyskalowanymi parametrami.

Analizy przeprowadzone w Okręgowej Komisji Egzaminacyjnej w Krakowie po sesji egzaminacyjnej dostarczyły bogatego materiału na temat funkcji informacyjnych wszystkich zastosowanych arkuszy. Rysunki 5 i 6 ilustrują kształty funkcji informacyjnej matematyczno-przyrodniczego arkusza gimnazjalnego zbudowanego z 36 zadań i arkusza maturalnego z geografii złożonego z 52 zadań. Funkcje informacyjne tych arkuszy zostały otrzymane z wykorzystaniem programu komputerowego RUMM i wyników 118593 gimnazjalistów i 819 maturzystów zdających egzamin z geografii w szkołach w obszarze działania krakowskiej Okręgowej Komisji Egzaminacyjnej.

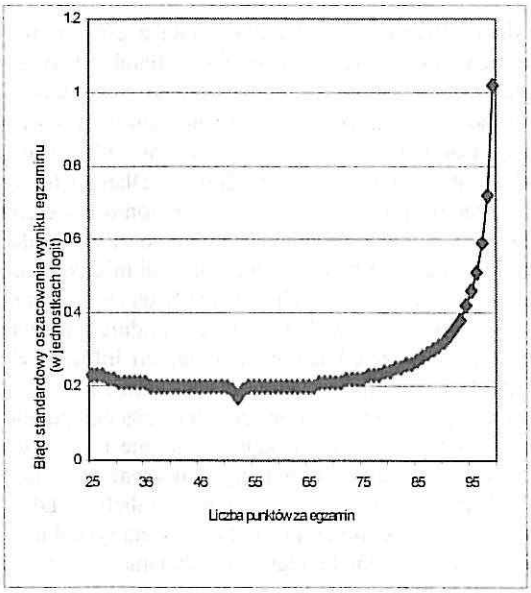


Rysunek 5. Funkcja informacyjna matematyczno-przyrodniczego arkusza egzaminacyjnego zastosowanego w sesji wiosennej 2002



Rysunek 6. Funkcja informacyjna maturalnego arkusza egzaminacyjnego z geografii zastosowanego w sesji wiosennej 2002

Jak można odczytać z rysunku 6, egzamin przeprowadzony z zastosowaniem arkusza geograficznego najwięcej informacji dostarczał o osiągnięciach uczniów, których wyniki plasują ich na skali w punkcie 0.20 i w jego bezpośrednim sąsiedztwie. Wartości skali równej 0,20 logits odpowiada wynik surowy równy prawie 52 punkty. Dla tej wartości skali wynik ucznia jest oszacowany z największą możliwą w tym egzaminie precyzją. Błąd standardowy wyniku ucznia w tym punkcie skali wynosi 0,194. Możemy więc przyjąć dla ucznia, który uzyskał surowy wynik 52 punkty, że w skali logits jego wynik prawdziwy zawiera się w przedziale $\theta = 0{,}200 \pm 2{*}0{,}194$.

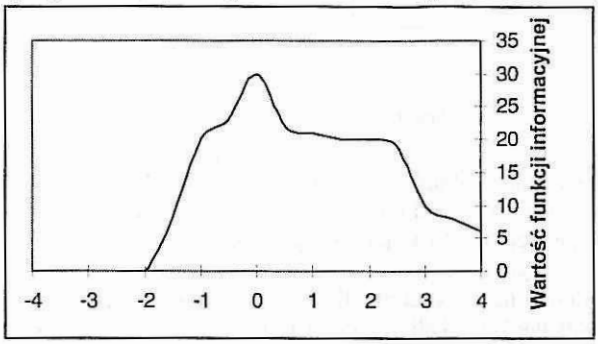


Rysunek 7. Zmieniająca się wartość błędu standardowego oszacowania wyniku w zależności od położenia na skali rezultatów egzaminu. Widoczne niewielkie minimum odpowiada wynikowi 52 punkty lub 0,20 logits w skali poziomu umiejętności

W egzaminie maturalnym z geografii najważniejszy był wynik 40 punktów (–0,28 logits), gdyż uzyskanie 40 i więcej punktów decydowało o zdaniu egzaminu maturalnego. W tym egzaminie wynik surowy 40 punktów jest oszacowany też z dobrą precyzją *se=0,200*, chociaż jest ona mniejsza niż dla wyniku 52 punkty. Konstruując arkusze do przyszłych egzaminów warto zadbać, aby dostarczał on możliwości maksymalnej precyzji oszacowania wyniku w punktach skali odpowiadających podejmowaniu decyzji o losach zdającego.

TWORZENIE ARKUSZA Z WYKORZYSTANIEM SZACOWANIA FUNKCJI INFORMACYJNEJ

Przyjmijmy, że tworzymy nowy arkusz egzaminacyjny na poziom maturalny z zadań o znanych parametrach. Możemy je tak dobrać, aby krzywa informacyjna arkusza egzaminacyjnego miała pożądany kształt, np. taki jak na zamieszczonym poniżej rysunku.

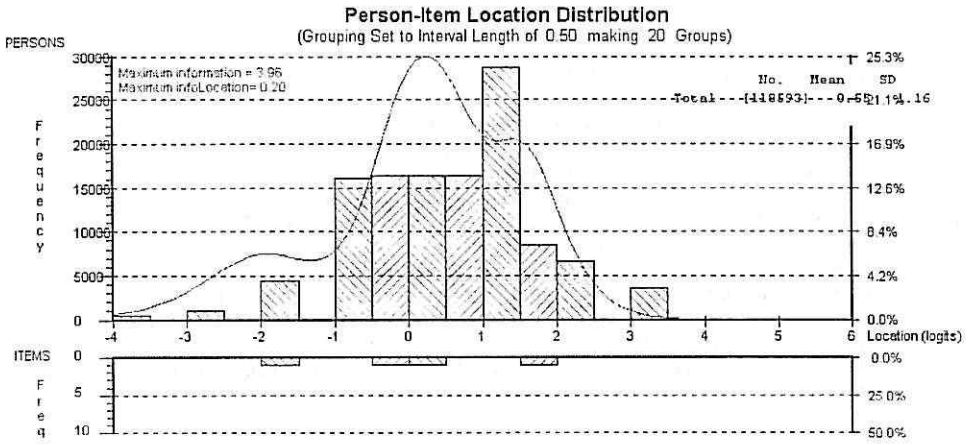


Rysunek 8. Planowany kształt funkcji informacyjnej maturalnego arkusza egzaminacyjnego z przedmiotu do wyboru

Korzystając z doświadczeń sesji egzaminacyjnej z 2002 r. przyjmijmy, że chcemy zbudować arkusz egzaminacyjny z geografii o podobnej charakterystyce.

Maksimum planowanej funkcji informacyjnej przypada dla pozycji skali $\theta = -0{,}19$, ponieważ tej pozycji skali odpowiada wynik surowy równy 40 punktów. Zgodnie z obowiązującymi przepisami był to wynik graniczny zdania egzaminu maturalnego z przedmiotów do wyboru w sesji wiosennej w 2002 r. Dlatego też arkusz egzaminacyjny powinien pozwolić na oszacowanie w tym obszarze poziomu umiejętności z możliwie najwyższą precyzją. Zakładając, że poziom badanej umiejętności w populacji zdających nie ulega gwałtownym wahaniom, można przyjąć, że umiejętności mierzone arkuszem egzaminacyjnym z geografii w populacji egzaminowanych również będą się rozciągać od –1 do +3 logits. Następny krok to wybór zadań. Wybór zadań nie tylko zgodnie z planem ze względu na mierzone osiągnięcia, ale także z uwzględnieniem ich funkcji informacyjnych tak, aby arkusz zapewnił pożądany kształt krzywej informacyjnej całego arkusza.

Dla egzaminu gimnazjalnego nie ma progu oznaczającego zdanie egzaminu. Planowany kształt funkcji informacyjnej mógłby wyglądać podobnie jak dla geografii tylko bez widocznego maksimum. Na zakończenie spróbujemy zbudować test złożony z czterech analizowanych zadań, których parametry opisane zostały w tabeli 1. Dla tych zadań graficzne przedstawienie ich funkcji informacyjnych przedstawiono na rysunku 2. Możemy teraz zobaczyć, jak wygląda funkcja informacyjna takiego czterozadaniowego testu.



Rysunek 9. Przedstawienie funkcji informacyjnej dla testu złożonego z czterech zadań: 26, 28, 30 i 35

Dokładając kolejne zadania możemy śledzić, jak zmienia się graficzne przedstawienie funkcji informacyjnej aż osiągniemy kształt zbliżony do zaplanowanego.

PODSUMOWANIE

Przeprowadzone po raz pierwszy w Polsce w masowej skali egzaminy zewnętrzne dostarczyły szeregu przesłanek do doskonalenia arkuszy egzaminacyjnych. Praca nad arkuszami może być obecnie wspomagana zarówno przez klasyczną teorię pomiaru dydaktycznego, jak i teorię odpowiedzi na zadanie testowe IRT. Bazująca na IRT skrupulatna analiza zadań po ich próbnym zastosowaniu może pomóc wykluczyć ukryte usterki zadań, które mogłyby ujawnić się dopiero na egzaminach właściwych. Uwzględnienie w pracach nad arkuszem wiedzy o funkcji informacyjnej zadań daje szanse na zbudowanie arkuszy, za pomocą których precyzja oszacowania wyników byłaby taka sama w całym przedziale umiejętności od najniższych do najwyższych, a więc dla całej populacji uczniów.

Warto już dzisiaj rozpocząć budowę banku zadań stosując IRT, aby w przyszłości dysponować parametrami zadań, które są niezależne od arkusza egzaminacyjnego, w którym były zastosowane i od rozkładu umiejętności w populacji, w której zastosowany był arkusz. Ponadto, trudność zadań i poziom umiejętności egzaminowanych uczniów badany tymi zadaniami może być określany w tej samej skali, co daje możliwość selekcji zadań najbardziej użytecznych do danego egzaminu. Jednak zadania pochodzące z różnych arkuszy egzaminacyjnych przed przekazaniem do banku wymagają skalowania, aby można było porównywać ich parametry. Jedna z najpopularniejszych metod skalowania opiera się na zadaniach kotwiczących. Może już dziś warto zadbać o to, aby arkusze przygotowywane na sesję egzaminacyjną w 2003 roku zawierały zadania kotwiczące stanowiące pomost pomiędzy arkuszami stosowanymi poprzez kolejne lata.

Henryk Szaleniec

LITERATURA

Hambleton R.K., Swaminathan H., Rogers H.J., *Fundamentals of Item Response Theor,*. Sage Publications, Inc. London 1991.

Hornowska E., *Testy psychologiczne. Teoria i praktyka*, Wydawnictwo Naukowe Scholar, Warszawa 2001.

B.Niemierko, M. K. Szmigel [red], *Teoria i praktyka oceniania zewnętrznego*. IV ogólnopolska konferencja z cyklu „Diagnostyka Edukacyjna", Pandit, Kraków 2001.

Szaleniec H., *Zastosowanie teorii analizy zadania testowego (IRT) w procesie oceniania zewnętrznego*, [w:]

Szaleniec H., Szmigel M. K., *Egzaminy zewnętrzne. Podnoszenie kompetencji nauczycieli w zakresie oceniania zewnętrznego*, Wydawnictwo „Zamiast Korepetycji", Kraków 2001.

Szaleniec H., *Probabilistyczne modele wyniku zadania testowego*, [w:] *Ocenianie szkolne ekonomika i polityka oświatowa probabilistyczne modele pomiaru*, Skrypt nr 3 dla uczestników III Podyplomowego Studium Ewaluacji Dydaktycznej na Uniwersytecie Gdańskim, Międzywydziałowe Studium Pedagogiczne Uniwersytetu Gdańskiego, Gdańsk 2002.

Wright B. D., Mok M., *Rasch models overview*, "Journal of Applied Measurement Constructing Variables" 1,1. 2000.