Marek Kryniewski
Zespół Szkół Energetycznych w Gdańsku

Klasyczne i probabilistyczne miary jakości zadania testowego – nowe możliwości

Omówienie klasycznej teorii testu

Zalety klasycznej teorii testu

1. W celu określenia charakterystycznych parametrów testu potrzebna jest stosunkowo niewielka, ale reprezentatywna próba uczniów.
2. Oszacowanie podstawowych parametrów dla zadań i testu, takich jak: łatwość i moc różnicująca nie wymaga zaawansowanego aparatu matematycznego.
3. Model teoretyczny leżący u podstaw szacowania charakterystycznych parametrów dla testu i zadań jest łatwy do przyswojenia przez konstruktorów zadań, jak i osób interpretujących wyniki.

Ograniczenia klasycznej teorii testów

1. Podstawowe parametry, takie jak:
   - moc różnicująca zadań
   - łatwość całego testu
   - łatwość poszczególnych zadań

   zależą bardzo mocno od próby uczniów, którzy brali udział w testowaniu. Rzadko jednak dysponujemy próbą, która by dobrze reprezentowała całą populację generalną.
2. Ponadto, łatwość zadania, która ma charakter probabilistyczny, jest jednakowa dla całej próby uczniów niezależnie od zdolności uczniów i poziomu opanowania danej umiejętności.
3. Błąd pomiaru oraz estymacji parametrów jest taki sam dla całej badanej populacji.

Klasyczne podejście do konstruowania testów

Konstruktor egzaminu, kierując się klasyczną teorią testu, powinien zapewnić:
- trafność wewnętrzną testu;
- wysoki współczynnik rzetelności (od jego wielkości zależy błąd pomiarowy);

- odpowiedni współczynnik łatwości zadań i testu dostosowany do rodzaju egzaminu;
- wysoką moc różnicującą zadań.

Przy takim podejściu do konstruowania egzaminów konstruktor napotykał przeszkody:

- brak banku zadań z opisem klasycznych współczynników;
- przy korzystaniu z banku zadań z opisem klasycznych współczynników konstruktor musi pamiętać, że łatwość oraz moc różnicująca zależą od populacji uczniów, gdzie przeprowadzony był egzamin;
- brak wygodnych narzędzi do symulacji parametrów całego egzaminu.

Dlatego konstruktorzy trzymający się klasycznej teorii testu opierali się na planie testu, własnym doświadczeniu oraz intuicji.

Omówienie probabilistycznej teorii testu

Przyczyny wzrostu zainteresowania probabilistyczną teorią testu w Polsce

Obecnie w Polsce teoretycy pomiaru dydaktycznego oraz konstruktorzy testów wykazują coraz większe zainteresowanie probabilistyczną teorią wyniku zadania testowego (*Item Response Theory* – IRT). Wzrost zainteresowania wiąże się z:

- rozwojem technik komputerowych (odpowiednio szybkie komputery oraz możliwość zakupu specjalistycznego oprogramowania);
- dostęp do dużych zbiorów danych tworzonych przez system egzaminów zewnętrznych;
- konieczność tworzenia baz danych zadań testowych o określonych parametrach niezależnych od populacji uczniów; bazy te pozwolą na konstruowanie arkuszy egzaminacyjnych o założonych parametrach (dostosowanych do celu egzaminu);
- konieczność tworzenia równoległych wersji egzaminów o tych samych parametrach;
- ze względów politycznych oraz społecznych konieczne jest również porównywanie wyników egzaminów w kolejnych latach. IRT może być pomocne przy określeniu, czy egzaminy w kolejnych latach były egzaminami równoległymi oraz czy egzaminy miały ten sam stopień trudności?

Założenia IRT

Probabilistyczna teoria wyniku zadania testowego opiera się na trzech podstawowych założeniach:

1. o wymiarach przestrzeni cechy nioebserwowalnej (latentnej – *dimentionality oflatent space*). Wszystkie istniejące zależności statystyczne między zadaniami testowymi są wyjaśniane przez odwołanie się do jednej cechy latentnej (dla testów to: wiedza–umiejętności);

2. o lokalnej niezależności zadań testowych. W założeniu tym przyjmuje się, że odpowiedzi każdej osoby badanej na jedno zadanie testowe nie zależą od jej odpowiedzi na jakiekolwiek inne zadanie tego testu;
3. o krzywej charakterystycznej zadania testowego (*Item Characteristic Curve* – ICC). Krzywa ta opisuje związek pomiędzy ukrytą cechą (latentną), np. stanem wiedzy (umiejętności) ucznia, którą chcemy poznać za pośrednictwem wyniku rozwiązania zadania a prawdopodobieństwem prawidłowej odpowiedzi na to zadanie.

W teorii IRT poziom wiedzy (umiejętności) opisany jest przez theta „Θ" (wynik prawdziwy), która ma związek z mierzoną wiedzą (umiejętnością). Prawdopodobieństwo, że egzaminowany udzieli poprawnej odpowiedzi na zadanie, jest opisane przez funkcję P(Θ).

Przyjmuje, że funkcja P(Θ) jest ciągła, a jej wartości zawierają się w przedziale od $-\infty$ do $+\infty$. Ponieważ skala jest najczęściej wyrażana w postaci konwencjonalnych wartości z, to w praktyce wszystkie wyniki mieszczą się w przedziale od –4z do +4z.

Model trójparametryczny

Najlepszym modelem do zastosowania w pomiarze dydaktycznym jest model trójparametryczny. Jest on modelem najogólniejszym, sformułowanym po raz pierwszy przez Birnbauma (1968). W modelu tym przyjmuje się, że prawdopodobieństwo udzielenia odpowiedzi prawidłowej na zadanie testowe zależy od trzech parametrów charakteryzujących zadanie testowe:

- trudności zadania testowego — parametr **a**
- jego mocy różnicującej — parametr **b**
- współczynnika zgadywania — parametr **c**.

Funkcję opisującą związek między prawdopodobieństwem $P_i(\Theta)$ udzielenia odpowiedzi prawidłowej na i-te zadanie, a poziomem wiedzy (umiejętności) w znormalizowanej skali można przedstawić jako:

$$P_i(\Theta) = c_i + \frac{1 - c_i}{1 + e - D \times a_i \times (\Theta - b_i)} \qquad i = 1, 2, \ldots n$$

gdzie:
n – ilość zadań w teście
c_i – to współczynnik zgadywania dla i-tego zadania
b_i – współczynnik trudności dla i-tego zadania
a_i – to współczynnik mocy różnicującej dla i-tego zadania
e – liczba Eulera w przybliżeniu e = 2,718... jest używana jako podstawa logarytmu naturalnego
D – stała maksymalizująca dopasowanie krzywej logistycznej do *ogivy* rozkładu normalnego; D = 1,7.

Najczęściej bowiem, aby rzetelność obliczeń podczas estymacji współczynników **a, b, c** była zadowalająca, trzeba przetworzyć dane uzyskane z przebadania przynajmniej 1000 osób.

Funkcja informacyjna oraz błąd standardowy pomiaru

Funkcja informacyjna podaje rozkład wielkości informacji niesionej przez zadanie dla całego zakresu zmiennej Θ. Wykres funkcji informacji $I_i(\Theta)$ zmiennej Θ lub znajomość jej wzoru pozwala na określenie w jakim przedziale zmiennej Θ zadanie niosło najwięcej informacji, czyli w jakim przedziale zmiennej Θ zadanie najlepiej opisywało zasób wiedzy badanych.

Dla trójparametrycznego modelu logistycznego Birnbaum w roku 1968 wyprowadził wzór na opisujący funkcję informacji dla pojedynczego zadania.

$$I_i(\Theta) = \frac{2{,}89 \times a_i^2 \times (1 - c_i)}{\left[c_i + e^{1{,}7 \times ai \times (\Theta - b_i)}\right] \times \left[c_i + e^{-1{,}7 \times ai \times (\Theta - b_i)}\right]^2}$$

Funkcja ta posiada jedno maksimum, którego wartość zmiennej Θ_{max} obliczył Birnbaum:

$$\Theta_{max} = b_i + \frac{1}{1{,}7 \times a_i} \ln[o{,}5 \times (1 + \sqrt{1 + 8 \times c_i})]$$

Można zdefiniować funkcję informacyjną dla całego testu jako sumę funkcji informacyjnych dla poszczególnych zadań.

Funkcja ta może mieć więcej niż jedno maksimum. Można modelować przebieg tej funkcji w zależności od charakteru testu poprzez dobór zadań składowych o znanych parametrach **a, b, c.**

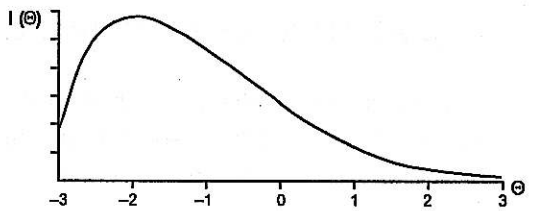
Znajomość funkcji informacyjnej dla testu pozwala na obliczenie błędu pomiarowego czyli błędu standardowego estymacji wartości Θ.

Błąd standardowy estymacji jest funkcją zmiennej Θ, tak więc nie jest wartością stałą dla wszystkich badanych tak, jak to było w klasycznej teorii testu, ale zmienia się w zależności od Θ. Przyjmuje swoje minimum w miejscu maksimum funkcji informacji.

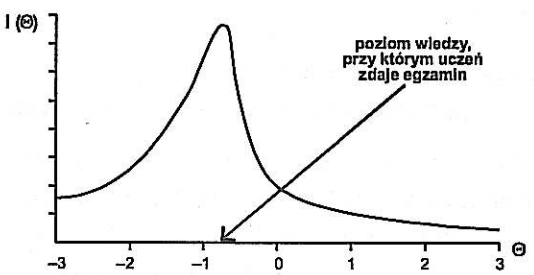
Pożądany przebiegu funkcji informacyjnej w zależności od zastosowania testu

Poprzez modelowanie funkcji informacyjnej, czyli odpowiedni dobór zadań wchodzących w skład testu, możemy otrzymać różne przebiegi funkcji informacyjnej. Możliwość otrzymania maksimum funkcji informacyjnej dla określonej wartości Θ stwarza możliwość otrzymania funkcji informacyjnej dostosowanej do charakteru testu.

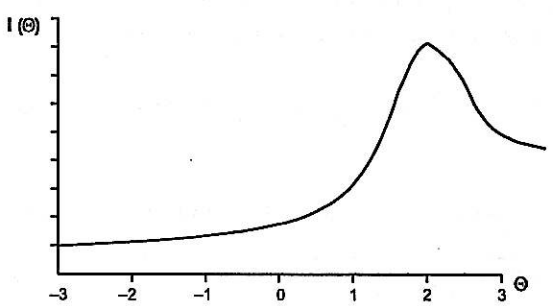
Rysunek 1.
Przebieg funkcji informacyjnej dla testu, którego zadaniem jest diagnoza wiadomości uczniów w zakresie podstawowym, np: test z zakresu podstawowych umiejętności czytania ze zrozumieniem



Rysunek 2.
Przebieg funkcji informacyjnej dla testu, w którym podany jest punkt odcięcia (zaliczył lub nie), np: egzamin maturalny



Rysunek 3.
Przebieg funkcji informacyjnej dla testu, w którym podany jest punkt odcięcia, lecz jest on ustawiony dla wysokich wiadomości, np: konkurs przedmiotowy



Można budować funkcję informacyjną posiadającą dwa maksima, np: dla testu, gdzie podane są dwa punkty odcięcia: pierwszy punkt to – zdał lub nie, i drugi – zdał na poziomie podstawowym lub rozszerzonym.

Probabilistyczne podejście do konstruowania testów

1. Korzystamy z banku zadań testowych o znanych parametrach a, b, c. Należy zwrócić uwagę na to, czy zdanie dobrze pasowało do stosowanego modelu. Informacja o dopasowaniu powinna być podana w banku zadań obok parametrów a, b, c.
2. Wybieramy zadania pasujące treściowo do planu testu.
3. Wykonujemy symulację przebiegu funkcji informacyjnej (o przebiegu dostosowanym do zastosowania testu), korzystając np. z arkusza kalkulacyjnego Excel.
4. Obliczamy parametry a, b, c dla całego testu, sprawdzając ich poziom jako dodatkowe kryterium.

Nieformalne związki między klasycznymi a probabilistycznymi parametrami opisu zadań testowych

Chcąc zastosować probabilistyczne podejście do konstruowania testów, należy zastosować zadania umieszczone w banku zadań testowych o znanych parametrach a, b, c. Dotarcie do takich informacji może być trudne. Zadania opisane klasycznymi parametrami mogą być łatwiejsze do uzyskania. Interesujące jest więc udzielenie, jakie istnieją nieformalne związki między klasycznymi a probabilistycznymi parametrami opisu zadań testowych.

Poniżej przedstawione zostało pięć charakterystycznych zadań pochodzących z Rejonowego Konkursu Informatycznego dla Województwa Pomorskiego.

Parametry klasyczne zostały obliczone przez program Iteman 3.5, a parametry probabilistyczne przez program Multilog 6.30. Wykresy funkcji charakterystycznych ICC zadań oraz funkcji informacyjnych wykonane zostały z użyciem arkusza kalkulacyjnego Excel.

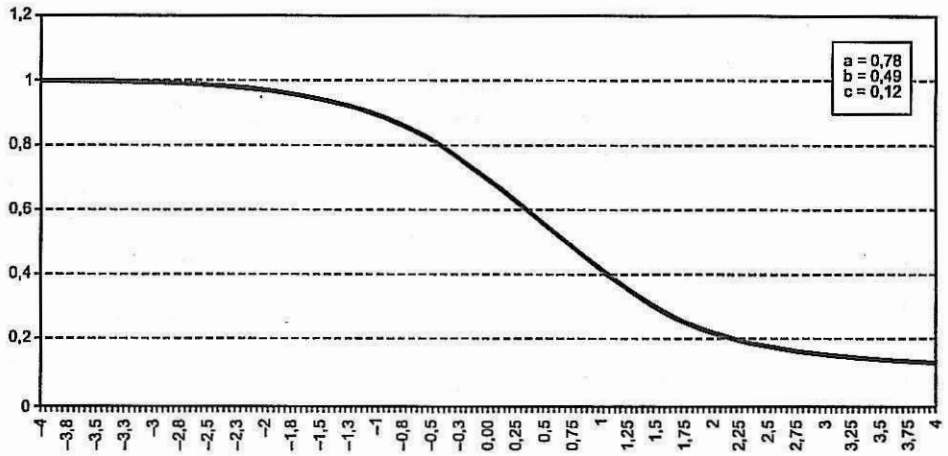
Uwaga: na wykresach funkcji informacyjnej nie zostało zachowane skalowanie osi Y.

ZADANIE 2. (błędnie skonstruowane, nie różnicujące, bardzo trudne)

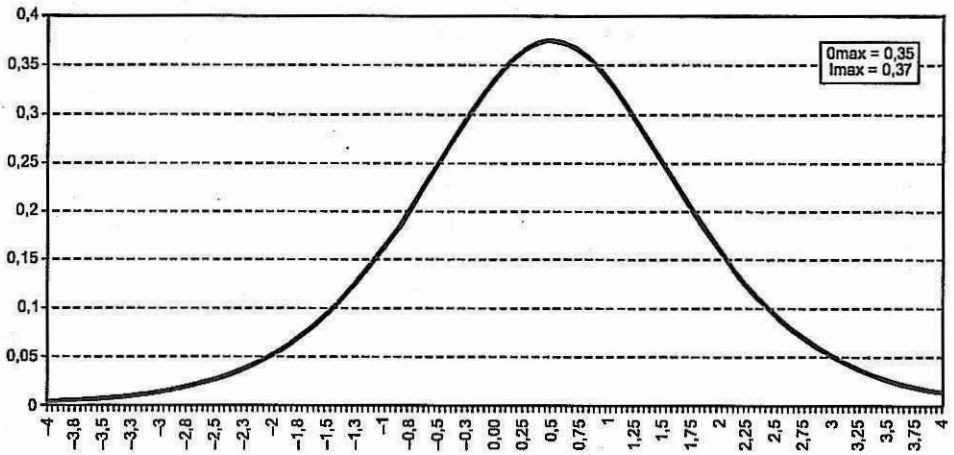
Pytanie 2: W programie MS Excel po wpisaniu wyrażenia +B12^0,5
A) jest niepoprawne
B) spowoduje wyświetlenie wyrażenia jako tekstu
C) oznacza pierwiastek z zawartości komórki B12
D) oznacza podniesienie do drugiej potęgi zawartości komórki B12

klasyczne miary oceny			
	miary mocy różnicującej		
łatwość	D_{27}	punktowo-dwuseryjny (*point biserial*)	dwuseryjny (*biserial*)
0,17	–0,03	0,00	0,00

Rysunek 4.
Krzywa charakterystyczna dla zadania 2



Rysunek 5.
Funkcja informacyjna dla zadania 2

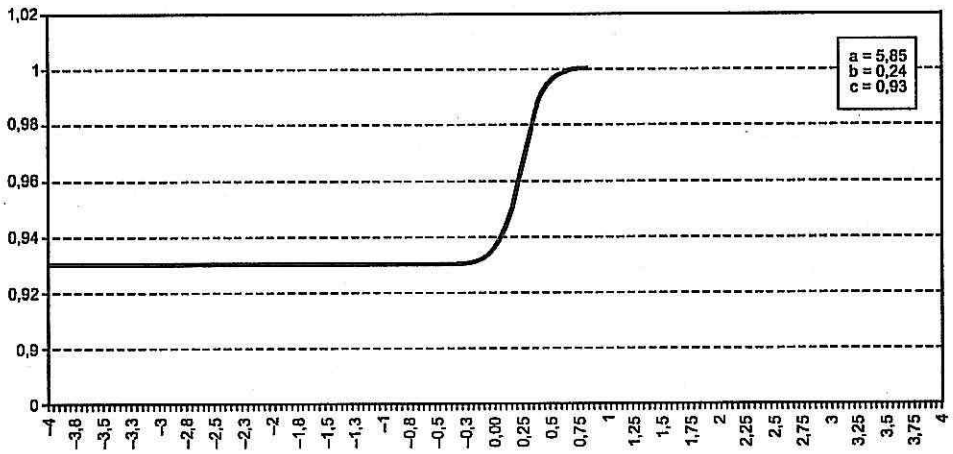


ZADANIE 3 (bardzo łatwe, słabo różnicujące)

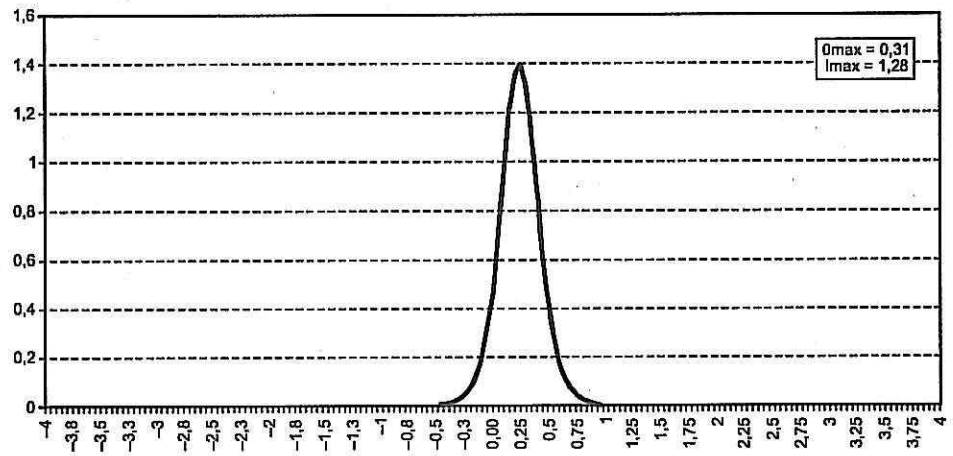
Pytanie 3: Której ikony nie usuniesz z pulpitu klawiszem Delete:			
A) CorelCity	B) Kosz	C) Micorsoft Outlook	D) MS Word

klasyczne miary oceny			
	miary mocy różnicującej		
łatwość	D_{27}	punktowo-dwuseryjny (*point biserial*)	dwuseryjny (*biserial*)
0,96	0,09	0,15	0,35

Rysunek 6.
Krzywa charakterystyczna dla zadania 3



Rysunek 7.
Funkcja informacyjna dla zadania 3

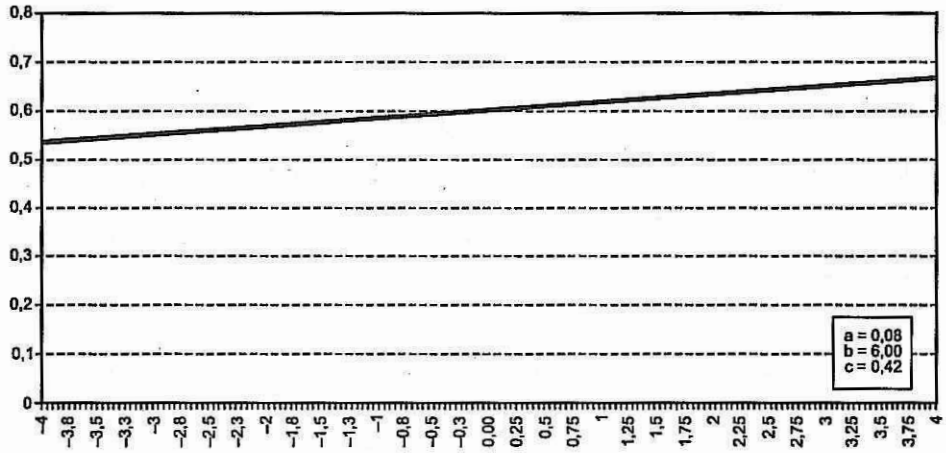


ZADANIE 9 (umiarkowanie trudne, słabo różnicujące)

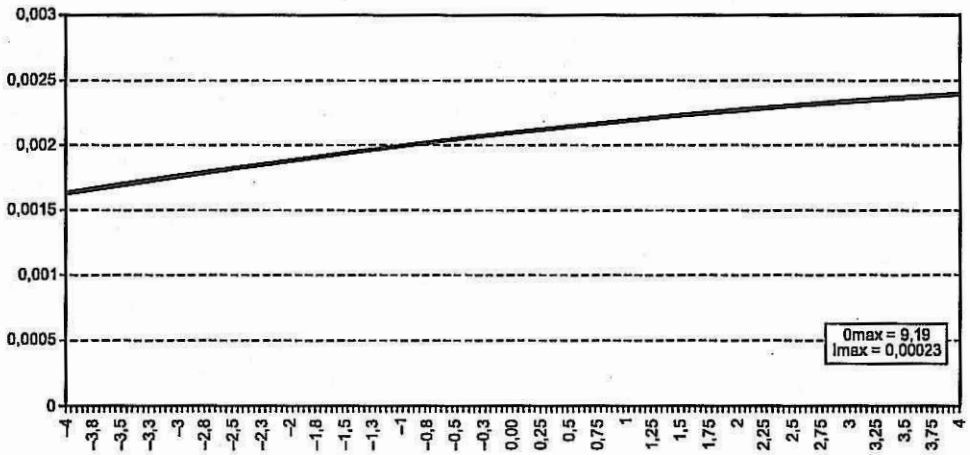
Pytanie 9: System operacyjny ładowany jest przez:			
A) dysk twardy	B) procedury POST-owe	C) użytkownika	D) BIOS

klasyczne miary oceny			
	miary mocy różnicującej		
łatwość	D_{27}	punktowo-dwuseryjny (*point biserial*)	dwuseryjny (*biserial*)
0,59	0,13	0,13	0,17

Rysunek 8.
Krzywa charakterystyczna dla zadania 9



Rysunek 9.
Funkcja informacyjna dla zadania 9

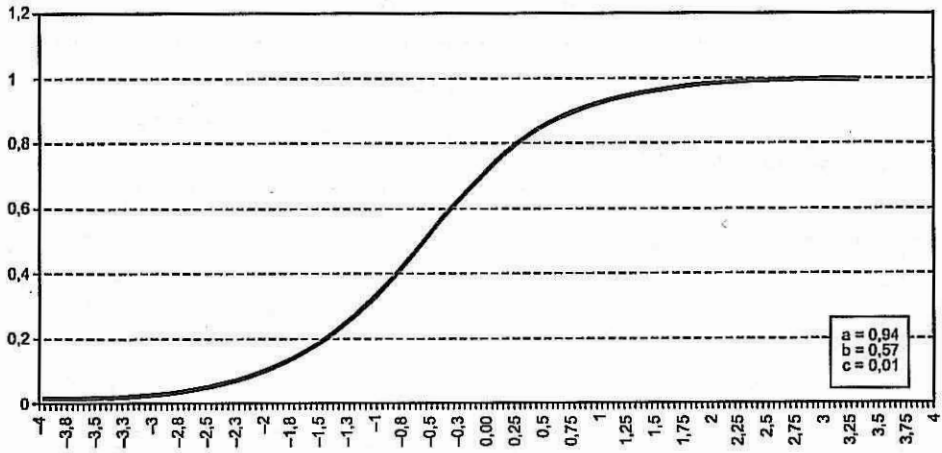


ZADANIE 10 (umiarkowanie trudne, dobrze różnicujące)

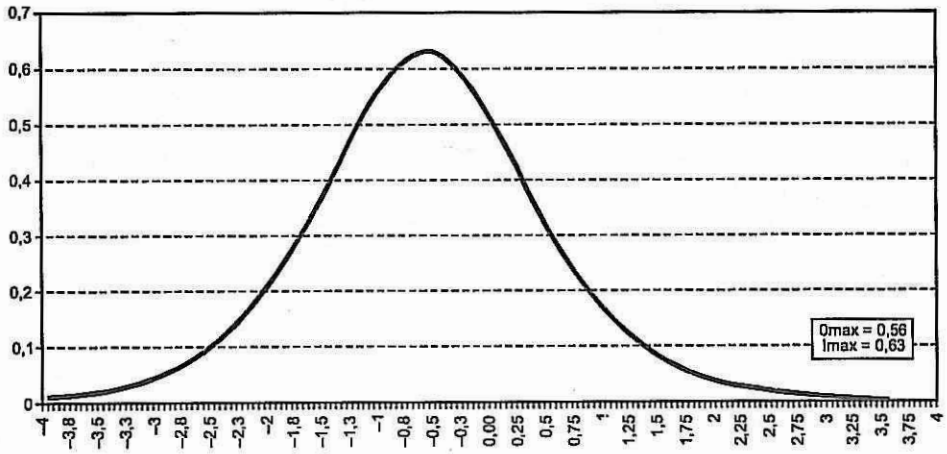
Pytanie 10: Programy wykonywalne mają rozszerzenia:			
A) bat exe run	B) exe sys com	C) bat com exe	D) bat ini exe

klasyczne miary oceny			
	miary mocy różnicującej		
łatwość	D_{27}	punktowo-dwuseryjny (*point biserial*)	dwuseryjny (*biserial*)
0,65	0,55	0,47	0,61

Rysunek 10.
Krzywa charakterystyczna dla zadania 10



Rysunek 11.
Funkcja informacyjna dla zadania 10

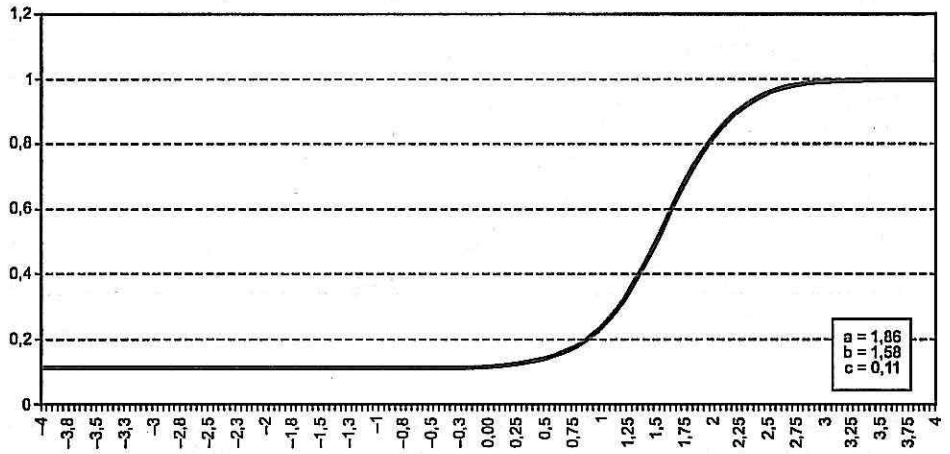


ZADANIE 12 (bardzo trudne, różnicujące)

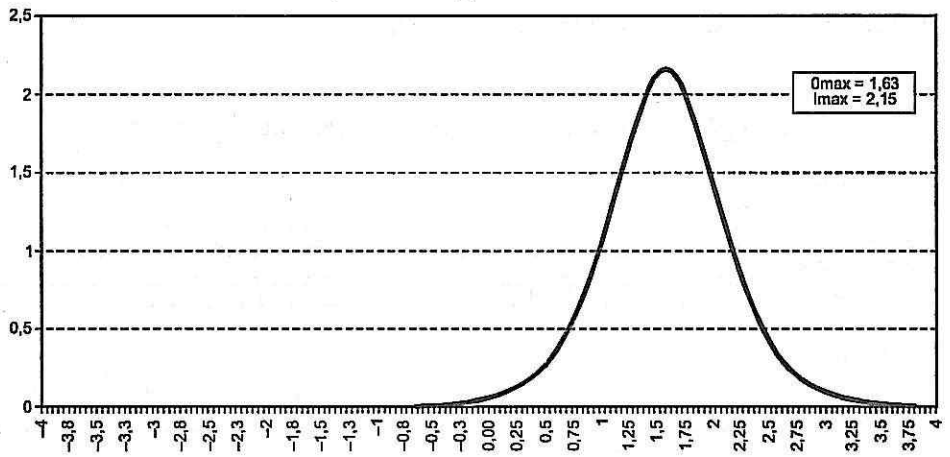
Pytanie 12: Zbiory konieczne do startu systemu z dyskietki systemowej DOS to:			
A) autoexec.bat config.sys command.com	B) io.sys command.com msdos.sys	C) autoexec.bat io.sys config.sys	D) msdos.sys autoexec.bat config.sys

klasyczne miary oceny			
	miary mocy różnicującej		
łatwość	D_{27}	punktowo-dwuseryjny (*point biserial*)	dwuseryjny (*biserial*)
0,19	0,24	0,33	0,48

Rysunek 12.
Krzywa charakterystyczna dla zadania 12



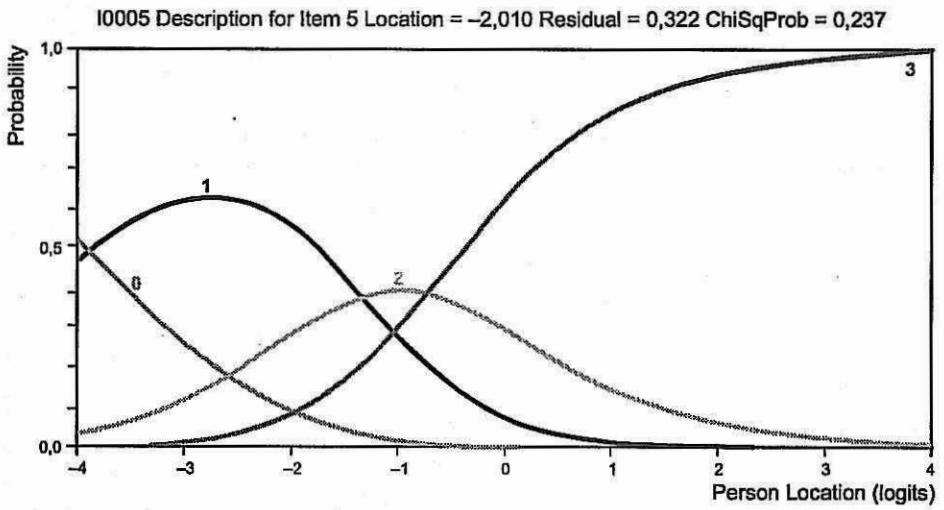
Rysunek 13.
Funkcja informacyjna dla zadania 12



Inne możliwości zastosowania probabilistycznej teorii zadania testowego

Analiza zadań wielopunktowych

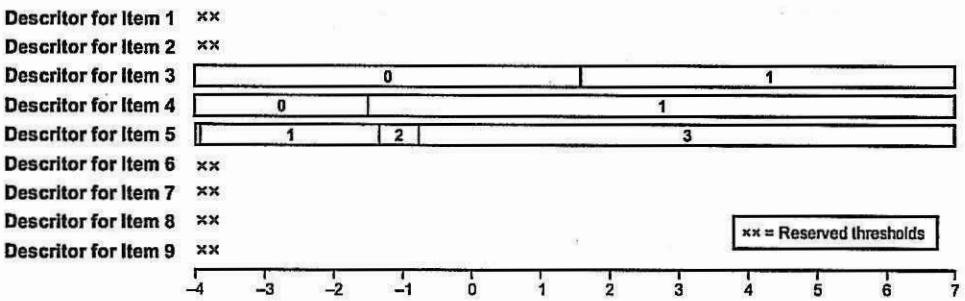
Rysunek 14.
Rozkłady prawdopodobieństwa dla części zadania 3b z arkusza I, punktowanego w skali 0–3 z Nowej Matury, maj 2002



Z wykresu rozkładu prawdopodobieństwa można uzyskać następujące informacje:
- przebiegi rozkładu prawdopodobieństwa dla wszystkich punktów w skali wielopunktowej;
- progi czyli odcięte punktów przecięcia krzywych rozkładu prawdopodobieństwa.

Analiza progów dla zadań wielopunktowych

Rysunek 15.
Mapa progów dla wybranych 9 zadań z Nowej Matury, maj 2002



Mapa progów przedstawia wartości dla wszystkich progów w całym teście (dla wybranych zadań). Podczas prezentacji progów zadań dla niektórych z nich nie są one podawane (są znaki xx), oznacza to, że obliczone progi nie są ustawione w narastającej formie, czyli np: próg między punktem 1 i 2 jest wyżej na skali unormowanej niż próg między punktem 2 i 3.

Brak monotoniczności ułożenia progów świadczy o:

- niezbyt trafnie skonstruowanym kluczu punktowania, przy zastosowaniu którego egzaminatorzy mają problemy z określeniem granicy między poszczególnymi punktami uzyskiwanymi przez ucznia;
- złym przygotowaniu egzaminatorów, braku szkolenia egzaminatorów lub też braku treningu przedegzaminacyjnego.

Zestawienie 1.
Wartości progów dla wybranych zadań (ITEM THRESHOLDS) z Nowej Matury, maj 2002

ITEM STATEMENT		THRESHOLDS			
Code	Statement	Mean	1	2	3
I0001	Descriptor for Item 1	,000	6,372	–4,739	–1,633
I0002	Descriptor for Item 2	,000	1,701	,966	–2,667
I0003	Descriptor for Item 3	,000	,000		
I0004	Descriptor for Item 4	,000	,000		
I0005	Descriptor for Item 5	,000	–1,928	,674	1,255
I0006	Descriptor for Item 6	,000	,258	–,102	–,157
I0007	Descriptor for Item 7	,000	,832	–,832	
I0008	Descriptor for Item 8	,000	,086	2,105	–2,191
I0009	Descriptor for Item 9	,000	–,860	1,113	–,253

Określenie parametrów dla badanej każdej osoby

Zestawienie 2.
Indywidualne określenie parametrów dla każdej osoby (INDIVIDUAL PERSON-FIT) – fragment z Nowej Matury, maj 2002

ID	Total	Max	Miss Extreme	Locn	SE	Residual	DegFree	DataPts	numer
1	3	22	9	-1,598	0,74	-0,212	7,7	9	1
2	4	22	9	-1,135	0,62	0,070	7,7	9	2
3	3	22	9	-1,598	0,74	0,314	7,7	9	3
4	9	22	9	-0,065	0,36	-0,810	7,7	9	4
77	14	22	9	0,514	0,34	-0,259	7,7	9	77
78	20	22	9	1,658	0,65	-0,942	7,7	9	78
79	14	22	9	0,514	0,34	1,411	7,7	9	79
80	17	22	9	0,912	0,40	-0,408	7,7	9	80
81	20	22	9	1,658	0,65	0,263	7,7	9	81
82	18	22	9	1,091	0,45	-0,626	7,7	9	82

Zestawienie to pozwala na określenie:

- miejsca każdego ucznia w unormowanej skali wiedzy, w praktyce od –3 do 3 (Locn);
- błąd standardowy szacowania miejsca ucznia w unormowanej skali (SE);
- Residual – dopasowanie każdej osoby do modelu.

Uwaga: Analiza IRT wybranych zadań matury z informatyki została wykonana z użyciem programu RUMM2010 produktu australijskiej firmy RUMM Laboratory Pty Ltd.. Adres internetowy www.rummlab.com.au

Sprawdzanie równoległości oraz poziomu trudności wersji testów

Wersje równoległe według klasycznej teorii testów są to testy mierzące „to samo w ten sam sposób", co oznacza narzędzia budowane według jednego planu, dające wyniki o jednakowych średnich, wariancjach i korelacjach z dowolną zmienną oraz o równej rzetelności (Niemierko, 2003).

Wersje równoległe według probabilistycznej teorii zadania testowego to testy budowane według jednego planu, posiadające takie same lub zbliżone krzywe charakterystyczne.

Ocena równoległości może być dokonana na podstawie:

- wielkość przestrzeni między krzywymi charakterystycznymi dla dwóch wersji testu;
- test hipotezy o równości trzech parametrów a, b, c dla dwóch wersji testu.

Porównanie stopnia trudności testów zbudowanych według jednego planu, może być oceniony na podstawie oceny różnicy wielkości parametru **b**.

Sprawdzanie poprawności tłumaczenia zadań testowych

Korzystając z zagranicznych banków danych zadań testowych konieczne będzie tłumaczenie zadań na język polski. Zadania w banku mogą mieć podane parametry probabilistyczne. Możemy poddać analizie IRT zadania po tłumaczeniu. Analiza porównawcza parametrów a, b,c przed tłumaczeniem i po tłumaczeniu pozwoli ocenić jakość przekładu. Ocenę identyczności parametrów a, b, c wykonujemy tak samo, jak sprawdzanie równoległości testów.

Testy adaptacyjne

Procedura testowania adaptacyjnego z użyciem IRT jest możliwa dzięki zastosowaniu komputerów ze specjalnym oprogramowaniem.

Jeden z możliwych algorytmów postępowania jest następujący:

1. W pierwszej kolejności badany rozwiązuje test wstępny. Na tej podstawie komputer określa poziom wiedzy ucznia z użyciem IRT. Wiedza ucznia Θ jest wyrażona w jednostce **z**.
2. Wyboru zadań w procedurze testowania adaptacyjnej dokonuje się na podstawie analizy funkcji informacyjnych zadań zgromadzonych w banku zadań.

Każde kolejne zadanie wybrane z banku zadań powinno dostarczać maksimum informacji.

3. Decyzje dotyczące tego, które zadania z banku zadań będą prezentowane egzaminowanemu, są podejmowane w czasie realizacji procedury testowania. Podjęcie decyzji wyboru każdego następnego zadania jest oparte na oszacowaniu wartości poziomu osiągnięć egzaminowanego Θ wynikającej z jego odpowiedzi na wcześniej rozwiązywane zadania (Sztejnberg, Hurek, 2003).

Zastosowanie testowania adaptacyjnego pozwoli na:
- określenie wiedzy ucznia z jak najmniejszym błędem pomiarowym;
- skrócenie czasu testowania poprzez zmniejszenie ilości zadań podczas procesu testowania jednego ucznia.

Bibliografia

Niemierko B. (2003), Materiały z sympozjum: „Zrównywanie wyników sprawdzianu 2003 do wyników sprawdzianu 2002", Sopot.

Sztejnberg A., Hurek J. (2003), *Zastosowanie osiągnięć technologii komputerowej w pomiarze edukacyjnym. Komputerowe testowanie w pełni adaptacyjne*, Uniwersytet Opolski.