Teoria i praktyka oceniania zewnętrznego
28–30 V 2001

Henryk SZALENIEC
Okręgowa Komisja Egzaminacyjna
w Krakowie

ZASTOSOWANIE TEORII ANALIZY ZADANIA TESTOWEGO (IRT) W PROCESIE OCENIANIA ZEWNĘTRZNEGO

Tworzenie pojedynczych zadań, jak i całych testów, w tym arkuszy egzaminacyjnych, oraz analiza ilościowa poszczególnych zadań, a także całych arkuszy egzaminacyjnych opiera się głównie na klasycznej teorii testu ugruntowanej w Polsce przez Bolesława Niemierkę.

Klasyczna teoria testu ma wiele istotnych zalet. Po pierwsze, aby określić charakterystyczne parametry testu potrzebna jest stosunkowo niewielka, ale reprezentatywna próba uczniów. Po drugie, oszacowanie podstawowych parametrów dla zadań i testu, takich jak łatwość i moc różnicująca, nie wymaga zaawansowanego aparatu matematycznego. Po trzecie, model teoretyczny leżący u podstaw szacowania charakterystycznych parametrów dla testu i zadań jest łatwy do przyswojenia zarówno przez konstruktorów zadań, jak i przez osoby interpretujące wyniki.

Klasyczna teoria testu ma jednak szereg ograniczeń. Dwa podstawowe parametry, takie jak moc różnicująca zadań, łatwość całego testu, jak i poszczególnych zadań, zależą bardzo mocno od próby uczniów, którzy brali udział w testowaniu. Jeżeli w próbie przeważali uczniowie o wysokim poziomie umiejętności, to w rezultacie analizy wyników otrzymamy informacje, że test był bardzo łatwy. I odwrotnie, jeżeli w próbie przeważali uczniowie o niskim poziomie badanych umiejętności, to test okazał się bardzo trudny. Możemy więc powiedzieć, że zgodnie z klasyczną teorią testu łatwość zarówno całego testu, jak i poszczególnych zadań jest ściśle uzależniona od próby. Jeżeli więc dokonamy standaryzacji zadań przeznaczonych do arkusza egzaminacyjnego na próbie uczniów, opierając się na klasycznej teorii testu, to istnieje duże prawdopodobieństwo, że podczas egzaminów zadania będą miały całkiem inne parametry niż te, które oszacowano na podstawie badań standaryzacyjnych. Jeżeli chodzi o moc różnicującą zadań, to im bardziej heterogeniczna (zróżnicowana pod względem badanych umiejętności) jest próba uczniów, tym wyższy jest współczynnik korelacji punktowo-dwuseryjnej opisujący moc różnicującą zadania. I odwrotnie — im grupa bardziej homogeniczna, tym mniejsza moc różnicująca zadań (Szaleniec, Szmigel 2001).

Ponadto łatwość zadania, która ma charakter probabilistyczny, jest jednakowa dla całej próby, niezależnie od zdolności uczniów i poziomu opanowania danej umiejętności. Jeżeli w wyniku badań standaryzacyjnych określimy, że łatwość zadania wynosi $p = 0{,}75$, to oznacza to także, że dla ucznia losowo wybranego z populacji prawdopodobieństwo rozwiązania tego zadania wynosi 0,75. Nawet jeżeli próba standaryzacyjna była reprezentatywna dla populacji, to musimy przyznać, że nie braliśmy w tym opisie pod uwagę indywidualnych różnic uczniów w zakresie poziomu umiejętności badanych danym testem. Dla uczniów bardzo zdolnych rzeczywiste prawdopodobieństwo poprawnego rozwiązania zadania jest wyższe niż 0,75,

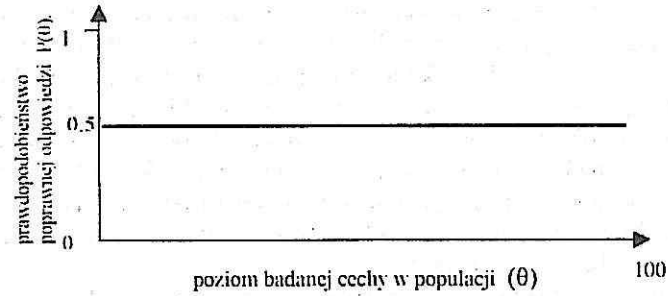
natomiast dla uczniów mniej zdolnych ma ono wartość poniżej 0,75. Posługując się klasyczną teorią testu możemy powiedzieć tylko, że p określa średnią wartość łatwości zadania w próbie wszystkich uczniów zarówno mniej, jak i bardziej zdolnych.

Dzisiaj, kiedy trwają prace nad przygotowaniem egzaminów zewnętrznych, i wiążą się z nimi ogromne oczekiwania zarówno zdających, jak i nauczycieli, istnieje konieczność tworzenia banków zadań, wzbogacenia arsenału narzędzi wspomagających zarówno tworzenie zadań, jak i ich analizę. Oprócz klasycznej teorii testu potrzebna jest nam teoria, która dostarczy modeli umożliwiających określenie parametrów zadań sprawdzających, niezależnych od próby standaryzacyjnej i od własności testu, w którego skład wchodzą. Takie nadzieje możemy wiązać z probabilistyczną teorią odpowiedzi na zadanie IRT (*Item Response Theory*). Podwaliny obecnie stosowanej teorii IRT dali niezależnie od siebie Rasch i Lord. Nie utrwaliła się jeszcze w języku polskim odpowiednia nazwa dla IRT. Dobrą propozycją wydaje się być zaproponowane przez B. Niemierkę sformułowanie „teoria odpowiedzi na zadanie" (TOZ) lub „probabilistyczna (stochastyczna) teoria odpowiedzi na zadanie". Agata Bieniek w wykładzie internetowego kursu z oceniania, zorganizowanego przez Centrum Otwartej i Multimedialnej Edukacji Uniwersytetu Warszawskiego, używa nazwy „teoria analizy odpowiedzi" (TAO) (Bieniek 2000). W niniejszym artykule będziemy posługiwać się angielskim akronimem IRT dla nazywania analizowanej teorii.

Zaawansowana matematyka, na której opiera się teoria, długo uniemożliwiała jej szersze praktyczne zastosowanie. Dzięki rozwojowi komputerów osobistych i profesjonalnego oprogramowania IRT w ostatnim dwudziestoleciu znalazła praktyczne zastosowanie w psychometrii i pomiarze dydaktycznym w wielu krajach. Osią tej teorii jest krzywa charakteryzująca zadanie testowe. Krzywa ta opisuje związek pomiędzy gotowością do rozwiązania zadania przez danego ucznia, wynikającą z jego poziomu umiejętności (czyli ukrytą cechą, którą mamy nadzieję poznać za pośrednictwem wyniku rozwiązania zadania, a która jest przedmiotem pomiaru) a prawdopodobieństwem prawidłowego rozwiązania tego zadania. Taki związek jest obserwowany w każdej procedurze pomiaru osiągnięć. Brak związku może sugerować, że zadanie nie jest w stanie różnicować uczniów o różnym nasileniu badanej cechy. W IRT poziom umiejętności opisany jest przez θ, która ma związek z mierzoną umiejętnością. Prawdopodobieństwo, że egzaminowany udzieli poprawnej odpowiedzi na zadanie, jest opisane przez funkcję $P(\theta)$.

Podstawowe założenie IRT głosi, że jeżeli znamy związek pomiędzy θ i $P(\theta)$ dla każdego zadania w teście, to charakterystyka każdego zadania, poziom umiejętności każdego egzaminowanego oraz błąd pomiaru związany z wynikiem mogą być oszacowane matematycznie. Ponadto charakterystyki każdego zadania, z których zbudowany jest test, także mogą być wyprowadzone z modelu matematycznego. Znajomość charakterystyk poszczególnych zadań daje możliwość, na poziomie tworzenia testu, zapobiegania dyskryminacji różnych warstw badanych wyodrębnionych z populacji.

W klasycznej teorii testu dysponujemy także krzywą charakterystyczną dla każdego zadania. Łatwość zdania p określa prawdopodobieństwo poprawnego rozwiązania danego zadania. Łatwość zadania w klasycznej teorii testu jest taka sama dla wszystkich biorących udział w egzaminie, niezależnie od tego, jaki poziom badanej cechy oni reprezentują. Ryc. 1. przedstawia krzywą charakterystyczną zadania wybranego z pilotażowego egzaminu maturalnego z geografii, przeprowadzonego przez OKE w Krakowie w marcu 2000 r.



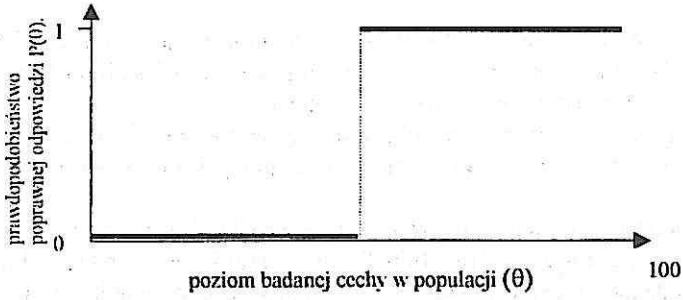
Ryc. 1. Krzywa charakterystyczna zadania zastosowanego na maturalnym egzaminie pilotażowym z geografii zgodnie z klasyczną teorią testu

Jak można odczytać z rysunku, prawdopodobieństwo poprawnego rozwiązania zadania wynosi 0,5 i jest niezależne od poziomu badanej cechy w całej populacji.

Podsumujmy. Jeżeli zadanie ma łatwość 0,5 i jest zadaniem trafnym, to prawdopodobieństwo poprawnego rozwiązania tego zadania P(θ) wynosi 0,5.

ANALIZA WYNIKÓW TESTOWANIA ZGODNIE Z MODELEM GUTTMANA

Posiłkując się ciągle klasyczną teorią testu, ale korzystając z modelu skalowania zaproponowanego przez L. Guttmana, charakterystykę zadania, dla którego prawdopodobieństwo rozwiązania równa się łatwości i wynosi $p = 0,5$, możemy opisać funkcją schodkową przedstawioną na ryc. 2.



Ryc. 2. Krzywa charakterystyczna zadania zastosowanego na maturalnym egzaminie pilotażowym z geografii zgodnie z klasyczną teorią testu, z zastosowaniem modelu skalowania Guttmana

Dla uczniów poniżej progu zaliczenia przyjętego dla poziomu umiejętności badanej przez zadanie prawdopodobieństwo poprawnego rozwiązania wynosi $P(\theta) = 0$, a powyżej tego progu $P(\theta) = 1$. Zgodnie z modelem zadanie dobrze różnicuje uczniów, którzy nie osiągnęli wymaganego progu od uczniów powyżej progu. Nie ma jednak możliwości

przewidzenia, jak różnią się prawdopodobieństwa poprawnej odpowiedzi dwóch uczniów znajdujących się poniżej progu, czy też dwóch uczniów znajdujących się powyżej progu.

Aby lepiej uzmysłowić sobie, jak funkcjonuje test, którego zadania spełniają założenia struktury Guttmana, rozważmy teoretyczny przykład przedstawiony w tab. 1. Nasz hipotetyczny test składa się ze zbioru 6 zadań punktowanych 0–1 i uszeregowanych według rosnącej trudności (malejącej łatwości). Gdy zsumujemy wyniki każdego ucznia, to suma punktów odzwierciedla względny poziom umiejętności uczniów sprawdzany danym testem. Możemy jednak zauważyć, że struktura odpowiedzi odzwierciedla coś więcej niż wyrażony jedną liczbą wynik punktowy.

Przypatrzmy się dokładniej tabeli wyników. Uczeń E otrzymał 4 punkty, a uczeń C tylko 2. Nie oznacza to jednak, że uczeń E rozwiązał o dwa dowolne zadania więcej od ucznia C. Uczeń E rozwiązał zadanie 1 i 2 oraz udzielił poprawnej odpowiedzi dodatkowo na zadanie 3 i 4.

TABELA 1

Struktura odpowiedzi uczniów w teście spełniającym założenia Guttmana

Zadanie/uczeń	1	2	3	4	5	6	Wynik
A	0	0	0	0	0	0	0
B	1	0	0	0	0	0	1
C	1	1	0	0	0	0	2
D	1	1	1	0	0	0	3
E	1	1	1	1	0	0	4
F	1	1	1	1	1	0	5

To samo możemy powiedzieć o dowolnie wybranej parze uczniów. Nie wystarczy stwierdzić, że dany uczeń otrzymał więcej punktów, ale jest ważne, które zadania rozwiązał poza rozwiązanym przez ucznia z niższym wynikiem.

Patrząc na tab. 1 możemy wnioskować, że im wyższy wynik otrzymał uczeń za rozwiązanie testu, tym wyższy jest poziom jego umiejętności, które były przedmiotem badania.

Tak bywa tylko w idealnym, teoretycznym przypadku. Rzeczywiste wyniki bardziej lub mniej odbiegają od teoretycznego rozkładu. Przedstawiając jednak wyniki zgodnie ze strukturą Guttmana, uzyskujemy dodatkową informację na temat narzędzia pomiarowego, jakie stanowi zastosowany test. Jeżeli test jest w dużym stopniu zgodny ze strukturą Guttmana, to wynik końcowy ucznia determinuje jednocześnie teoretyczną strukturę jego odpowiedzi. Widzimy więc, że w tej analizie wynik końcowy ucznia ma kluczowe znaczenie. W teorii opartej na modelach Rascha wynik końcowy ucznia nie determinuje struktury jego odpowiedzi na poszczególne zadania, ale struktura Guttmana jest jedną z najbardziej prawdopodobnych.

Rozważmy wyniki jednej klasy (27 uczniów), rozwiązującej test z fizyki złożony z 12 zadań punktowanych 0–1.

TABELA 2

Wyniki uczniów w teście złożonym z dwunastu zadań punktowanych 0–1

Uczeń	Numery zadań uporządkowanych według rosnącej trudności (malejącej łatwości)												Wynik surowy	Wskaźnik umiejętności	Częstość w grupie o tej samej liczbie punktów	Opis dydaktyczny wyniku
	1	4	12	8	7	9	5	2	6	10	11	3				
1.	1	0	0	0	0	0	0	0	0	0	0	0	1	B_1	1	niski $p_9 = 0{,}13$ $p_6 = 0{,}00$
2.	1	1	0	0	0	0	0	0	0	0	0	0	2	B_2	3	
3.	1	1	0	0	0	0	0	0	0	0	0	0	2	B_3		
4.	1	0	1	0	0	0	0	0	0	0	0	0	2	B_4		
5.	1	1	1	0	0	0	0	0	0	0	0	0	3	B_5	2	
6.	1	1	0	1	0	0	0	0	0	0	0	0	3	B_6		
7.	1	1	0	1	1	0	0	0	0	0	0	0	4	B_7	2	
8.	1	0	1	1	0	1	0	0	0	0	0	0	4	B_8		
9.	1	1	1	1	1	0	0	0	0	0	0	0	5	B_9	2	średni $p_9 = 0{,}67$ $p_6 = 0{,}25$
10.	1	0	1	1	1	1	0	0	0	0	0	0	5	B_{10}		
11.	1	1	1	1	1	1	0	0	0	0	0	0	6	B_{11}	3	
12.	1	1	1	1	1	1	0	0	0	0	0	0	6	B_{12}		
13.	1	1	1	1	1	0	0	1	0	0	0	0	6	B_{13}		
14.	1	1	1	1	1	0	1	1	0	0	0	0	7	B_{14}	5	
15.	1	1	1	1	1	1	0	1	0	0	0	0	7	B_{15}		
16.	1	1	1	1	1	0	1	1	0	0	0	0	7	B_{16}		
17.	1	1	1	1	0	1	1	0	1	0	0	0	7	B_{17}		
18.	1	1	1	1	0	1	1	0	1	0	0	0	7	B_{18}		
19.	1	1	1	1	1	1	1	0	1	0	0	0	8	B_{19}	2	
20.	1	1	1	1	1	1	1	0	0	0	1	0	8	B_{20}		
21.	1	1	1	1	1	1	1	1	1	0	0	0	9	B_{21}	2	wysoki $p_9 = 1{,}00$ $p_6 = 0{,}86$
22.	1	1	1	1	1	1	1	0	1	0	1	0	9	B_{22}		
23.	1	1	1	1	1	1	1	0	1	1	1	0	10	B_{23}	2	
24.	1	1	1	1	1	1	1	1	0	1	0	1	10	B_{24}		

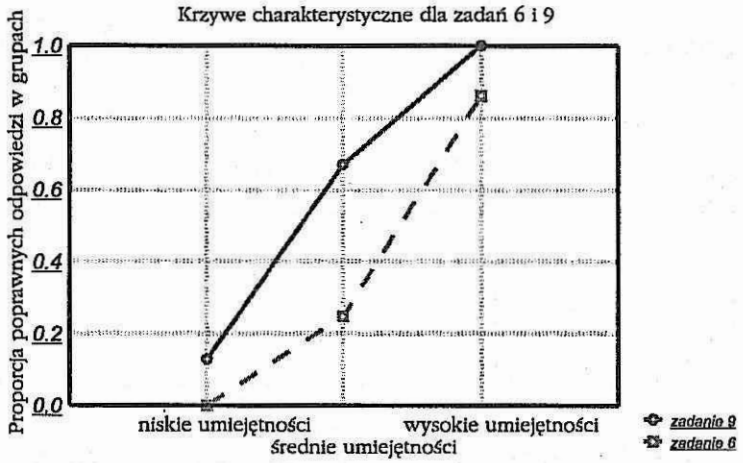
Uczeń	Numery zadań uporządkowanych według rosnącej trudności (malejącej łatwości)												Wynik surowy	Wskaźnik umiejętności	Częstość w grupie o tej samej liczbie punktów	Opis dydaktyczny wyniku
	1	4	12	8	7	9	5	2	6	10	11	3				
25.	1	1	1	1	1	1	1	1	1	1	0	1	11	B_{25}	2	wysoki $p_9 = 1{,}00$ $p_6 = 0{,}86$
26.	1	1	1	1	1	1	1	1	1	1	1	0	11	B_{26}		
27.	1	1	1	1	1	1	1	1	1	1	1	1	12	B_{27}	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...		
N																
Liczba poprawnych odpowiedzi	26	23	22	22	18	16	13	9	9	5	5	3				
	D_1	D_4	D_{12}	D_8	D_7	D_9	D_5	D_2	D_6	D_{10}	D_{11}	D_3				

Korzystając z propozycji, jaką przedstawił Dawid Andrich podczas kursu wykorzystania modelu Rascha (Andrich 2000), obliczymy proporcje poprawnych odpowiedzi w każdej z trzech grup np. dla zadania 6 i 9. Ponieważ zadania były punktowane 0–1, otrzymamy średnie łatwości zadań w grupach uczniów o różnym poziomie badanej testem umiejętności.

Poziom badanych umiejętności w trzech grupach uczniów o zróżnicowanym poziomie umiejętności badanych testem	Proporcja poprawnych odpowiedzi w poszczególnych grupach	
	Zadanie 9	Zadanie 6
Uczniowie o niskim poziomie umiejętności	0,13	0,00
Uczniowie o średnim poziomie umiejętności	0,67	0,25
Uczniowie o wysokim poziomie umiejętności	1,00	0,86

Jeżeli teraz przedstawimy graficznie zależność proporcji poprawnych odpowiedzi od grupy uczniów charakteryzowanej w populacji poziomem umiejętności badanych testem, to otrzymamy krzywą charakterystyczną zadania. Ryc. 3. przedstawia empiryczne krzywe charakterystyczne dla zadań 6 i 9.

Jeżeli populację uczniów znacznie zwiększymy i podzielimy na więcej grup, to otrzymamy krzywą bardziej gładką. Przedstawione na rysunku charakterystyki informują, jak zmienia się trudność zadania w zależności od tego, w jakiej grupie przeprowadzony został pomiar.



Ryc. 3. Empiryczne krzywe charakterystyczne dla zadań 6 i 9 z testu, którego wyniki przedstawione są w tab. 2

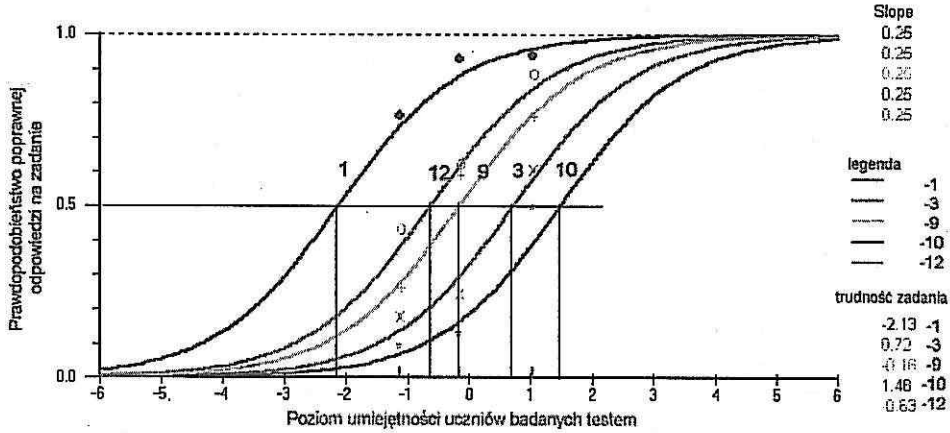
ZASTOSOWANIE „PROBABILISTYCZNEJ TEORII ODPOWIEDZI NA ZADANIE TESTOWE"

Zgodnie z założeniami „probabilistycznej teorii odpowiedzi na zadanie testowe" zarówno trudność zadania zastosowanego w teście, jak i poziom umiejętności (zdolność do wykonania zadania) dadzą się wyrazić za pomocą takiej samej liniowej skali, czyli w takich samych jednostkach.

Ponadto parametry określające i trudność zadania, i poziom umiejętności ucznia są od siebie niezależne. Założenia te sprowadzają się do przyjęcia, że znamy takie przekształcenie M, które pozwala na podstawie wyników testowania powiązać prawdopodobieństwo poprawnej odpowiedzi z parametrem określającym trudność zadania oraz z parametrem określającym poziom umiejętności badanych testem u danego ucznia, w którego skład wchodzi dane zadanie (Perline, Wright, Wainer 2000). Korzystając z tych założeń, dokonamy przeglądu kilku zastosowań teorii do analizy zadań.

Jak już wcześniej wspomniano, proporcja poprawnych odpowiedzi (łatwość dla zadań punktowanych 0–1) określa prawdopodobieństwo $P(x = 1)$ uzyskania za zadanie 1 punktu przez uczniów o różnym poziomie umiejętności badanych danym testem. Przyjrzyjmy się charakterystykom zadań, których wyniki dla jednej klasy przedstawiono w tab. 2.

Oś pionowa reprezentuje zarówno proporcje poprawnych odpowiedzi (punkty na wykresie), jak i teoretyczne prawdopodobieństwo rozwiązania przez uczniów zadania oszacowane jako funkcja trudności zadania i poziomu umiejętności uczniów w badanej testem dziedzinie. Na osi poziomej mamy wspólną znormalizowaną skalę, w której wyrażone są zarówno trudności zadań, jak i poziom umiejętności uczniów. Jest to skala liniowa o równych przedziałach odpowiadających odchyleniu standardowemu, dla której średnia wynosi 0. Ujemne wartości oznaczają poziom umiejętności i trudność zadań poniżej



Ryc. 4. Krzywe charakterystyczne dla pięciu zadań o różnej trudności (wyniki zadań *z*1, *z*12, *z*9, *z*3, *z*10 zaczerpnięto z testu badań osiągnięć uczniów klas siódmych z fizyki prowadzonych w 1989 r. przez Henryka Szaleńca w szkołach województwa krakowskiego)

średniej, a dodatnie — powyżej średniej. Punkty zaznaczone na wykresie oznaczają dane empiryczne dla trzech grup badanych uczniów o niskim, średnim i wysokim poziomie osiągnięć w badanej testem dziedzinie (por. opis w ostatniej kolumnie tab. 2). Jeżeli poziom umiejętności ucznia z danej dziedziny jest równy trudności zadania, to prawdopodobieństwo rozwiązania zadania przez danego ucznia wynosi 0,5. Rysując prostą równoległą do osi poziomej układu współrzędnych na poziomie prawdopodobieństwa 0,5 możemy wyznaczyć graficznie trudności zadań. I tak bardzo łatwe zadanie 1, które rozwiązali prawie wszyscy uczniowie, ma w nowej skali trudność –2,13. Najtrudniejsze w teście zadanie 10 ma trudność +1,48.

Jak można zaobserwować na ryc. 4, trudność danego zadania odpowiada także wartości określonej na osi poziomej dla punktu przegięcia krzywej charakterystycznej zadania[1]. Parametr, który odpowiada za stromość nachylenia krzywej charakterystycznej zadania w jej punkcie przegięcia, określa moc różnicującą zadania (w tym tekście ten parametr będziemy oznaczać literką *a*). Im bardziej krzywa jest stroma, tym większa moc różnicująca zadania. Na ryc. 4 parametr *a* jest taki sam dla wszystkich zadań i wynosi 0,25 (*slope* 0,25). Zadania wybrane do analizy tworzą podtest zadań o rosnącej trudności od –2,13 do 1,48 i stałej mocy różnicującej. Za każde zadanie uczeń mógł uzyskać 0 lub 1 punkt.

W takim przypadku krzywą charakterystyczną zadania *i*, opisującą prawdopodobieństwo uzyskania przez ucznia *n* jednego punktu za poprawną odpowiedź, opisuje następujące równanie:

$$P\{\theta_n\} = \frac{1}{1 + e^{-D(\theta_n - \delta_i)}}$$

[1] Punkt przegięcia wykresu jest to punkt na wykresie, w którym styczna do krzywej przechodzi z jednej strony wykresu na drugą.

gdzie:

— θ_n oznacza poziom umiejętności ucznia, oszacowany na podstawie wyniku surowego uzyskanego w całym teście;
— $P\{\theta_n\}$ określa prawdopodobieństwo poprawnego rozwiązania analizowanego zadania i przez ucznia n o poziomie umiejętności θ;
— δ_i jest względnym wskaźnikiem trudności zadania;
— D natomiast jest stałą skalowania i wynosi 1,7.

Widzimy więc, że kluczowym parametrem statystycznym w tej analizie jest uzyskany w całym teście surowy wynik ucznia, który stanowi podstawę oszacowania poziomu umiejętności każdego ucznia badanego danym testem. Ten najprostszy model opisany powyższym równaniem nazywamy modelem jednoparametrycznym. Jedynym szacowanym parametrem jest trudność zadania δ_i. Jeżeli test składa się z zadań o różnej mocy różnicującej, to równanie przyjmuje postać:

$$P\{\theta_n\} = \frac{1}{1 + e^{-Da_i(\theta_n - \delta_i)}}$$

gdzie a_i jest parametrem opisującym moc różnicującą zadania.

Równanie to przyjmuje bardziej złożoną postać, gdy przejdziemy do testu zbudowanego z zadań wielokrotnego wyboru. Kolejny parametr, który może być oszacowany, to poprawka na zgadywanie. Poprawka ta zależy od atrakcyjności poszczególnych dystraktorów, a także od poziomu umiejętności danego ucznia. Po wprowadzeniu trzeciego parametru równanie przyjmuje postać:

$$P\{\theta_n\} = c_{in} + \frac{1 - c_{in}}{1 + e^{-Da_i(\theta_n - \delta_i)}}$$

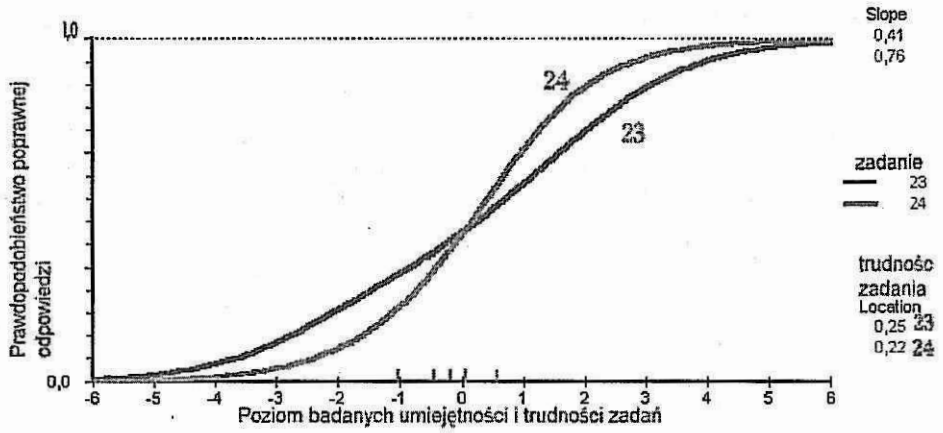
gdzie c_{in} jest parametrem określającym prawdopodobieństwo udzielenia poprawnej odpowiedzi w danym zadaniu przy braku koniecznej wiedzy, czyli zgadując odpowiedź.

Dla zadań otwartych punktowanych według ustopniowanego schematu oceniania posiadających zróżnicowaną moc różnicującą dochodzi czwarty parametr F_{ik} związany z progami i oznacza trudność progu k w zadaniu i punktowanym 1, 2, 3...

JAK CZYTAĆ INFORMACJE O ZADANIACH UZYSKANE Z WYKORZYSTANIEM TEORII ANALIZY ODPOWIEDZI NA ZADANIE TESTOWE

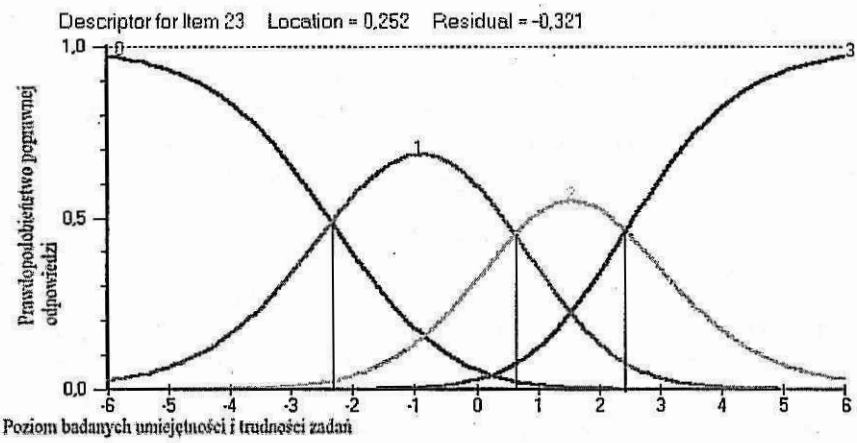
W tym rozdziale przeanalizujemy charakterystyki kilku zadań, próbując zinterpretować znaczenie uzyskanych danych dla doskonalenia procesu tworzenia i arkuszy egzaminacyjnych oraz analizy wyników. Do analizy wykorzystano oprogramowanie komputerowe RUMM2010 Rasch Unidimensional Measurement Models.

Zacznijmy od sprawdzenia, w jakim stopniu wybrany model analizy przystaje do wyników empirycznych na przykładzie trzech zadań z testu zastosowanego w badaniach osiągnięć z geografii w 1998 r. w klasach III szkół średnich województwa krakowskiego. Przyjrzyjmy się dokładniej charakterystykom zadań.

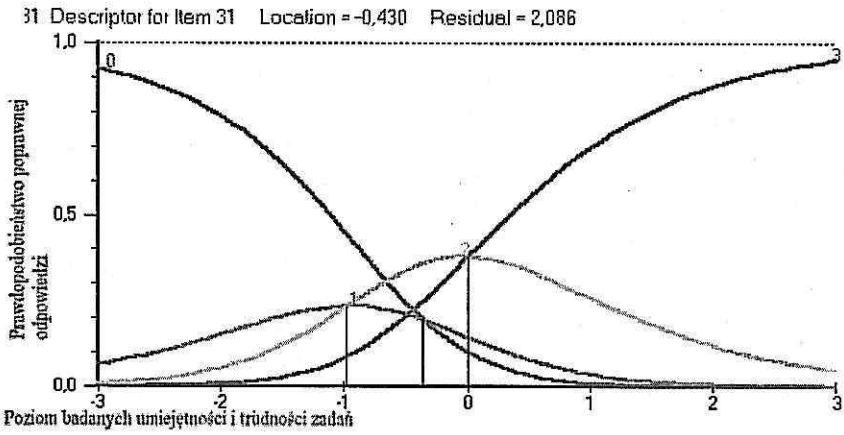


Ryc. 7. Zadania o różnej mocy różnicującej i podobnej trudności

można zauważyć na ryc. 7, dla uczniów o poziomie badanych umiejętności poniżej średniej (poniżej 0) istnieje wyższe prawdopodobieństwo poprawnego rozwiązania zadania 23 niż zadania 24. Natomiast dla uczniów lokujących się na znormalizowanej skali powyżej średniej prawdopodobieństwo poprawnego rozwiązania zadania 23 jest niższe niż zadania 24. Różnice mocy różnicującej zadań, których charakterystyki przedstawione są na ryc. 7, mają istotne znaczenie pomiarowe. Zadanie 24 o bardziej stromej charakterystyce i wyższej mocy różnicującej daje znaczny wzrost prawdopodobieństwa poprawnej odpowiedzi już przy niewielkim wzroście poziomu badanej umiejętności. I tak np. jeżeli poziom umiejętności ucznia zmieni się w tym przypadku o jedno odchylenie standardowe, to przyrost prawdopodobieństwa poprawnej odpowiedzi na zadanie 24 ma dwa razy większą wartość niż dla zadania 23. Uzyskane na podstawie analizy



Ryc. 8. Uszeregowane progi uzyskania 1, 2 i 3 punktów w zadaniu 23



Ryc. 9. Zaburzenie uszeregowania progów w zadaniu punktowanym od 1 do 3

informacje mogą mieć istotne znaczenie przy tworzeniu arkusza egzaminacyjnego o pożądanych właściwościach pomiarowych.

Obydwa zadania, których wyniki zostały zacytowane, były zadaniami otwartymi, za które uczeń mógł uzyskać od 1 do 3 punktów. Zobaczmy teraz, jak funkcjonowały progi ustalone w schemacie oceniania zadania 23.

Pierwsza krzywa, oznaczona cyfrą 0, obrazuje prawdopodobieństwo uzyskania 0 punktów przez uczniów o różnym poziomie badanych umiejętności. Kolejne krzywe, oznaczone cyframi 1, 2 i 3, ilustrują prawdopodobieństwa uzyskania odpowiednio 1, 2 i 3 punktów. Punkt przecięcia krzywej 0 i krzywej 1 odpowiada pierwszemu progowi. Możemy odczytać z wykresu, że uczeń o poziomie umiejętności wyrażonej w standaryzowanej skali wskaźnikiem –2,3 ma takie samo prawdopodobieństwo uzyskania zarówno 0 punktów, jak i 1. Dla tego zadania to prawdopodobieństwo wynosi 0,5. Uczniowie ulokowani na skali umiejętności na pozycji –1 mają największe prawdopodobieństwo uzyskania jednego punktu. Kolejny próg ma wartość +0,7 i odpowiada przecięciu się krzywej prawdopodobieństwa uzyskania 1 i 2 punktów. Oznacza to, że uczniowie o poziomie umiejętności 0,7 odchylenia standardowego powyżej średniej mają takie samo prawdopodobieństwo uzyskania zarówno 1, jak i 2 punktów. Uzyskanie 2 punktów w tym zadaniu jest najbardziej prawdopodobne dla uczniów o poziomie umiejętności 1,5 na skali, czyli 1,5 odchylenia standardowego powyżej średniej. Jednakowe prawdopodobieństwo uzyskania 2 i 3 punktów mają uczniowie, dla których poziom badanych umiejętności daleko odbiega od przeciętnej i wynosi +2,4. Dla uczniów powyżej tego punktu na skali najbardziej prawdopodobny wynik to 3 punkty. Przedstawiona charakterystyka dostarcza podstaw do wnioskowania, że omawiane zadanie zostało poprawnie skonstruowane. Można także przypuszczać, że oceniający go egzaminatorzy nie mieli większych wątpliwości, jak oceniać uczniów, którzy wykazali się różnym poziomem opanowania umiejętności badanych danym zadaniem.

Całkiem inaczej wygląda charakterystyka zadania 32 w tym samym teście. Jest to również zadanie punktowane od 1 do 3.

Jak można zauważyć na podstawie rysunku, krzywa opisująca prawdopodobieństwo uzyskania jednego punktu ma swoje maksimum dla uczniów o poziomie umiejętności –1. Jest to też drugi próg dla zadania, co oznacza, że prawdopodobieństwo uzyskania 1 i 2 punktów jest jednakowe. Dla ucznia o takim poziomie umiejętności prawdopodobieństwo uzyskania 0 punktów wynosi $P(\theta = -1, 0) = 0{,}47$, prawdopodobieństwo uzyskania 1 lub 2 punktów jest jednakowe i wynosi 0,2, natomiast uzyskanie 3 punktów jest mało prawdopodobne, ponieważ $P(\theta = -1, 3) = 0{,}09$. Pierwszy próg odpowiadający równemu prawdopodobieństwu uzyskania 0 lub 1 punktów znajduje się pomiędzy progiem drugim oraz trzecim i wynosi –0,4, czyli $P(\theta = -0{,}4, 0) = P\theta = -0{,}4\ 1) = 0{,}2$. Istnieje jakaś przyczyna lub kilka przyczyn, które powodują zaburzenie kolejności progów w charakterystyce tego zadania. Jakaś przyczyna powoduje, że dla uczniów o poziomie umiejętności –0,7 prawdopodobieństwo uzyskania 0 lub 2 punktów jest jednakowe i wynosi 0,3, podczas gdy prawdopodobieństwo uzyskania przez tych uczniów 1 punktu jest znacznie niższe.

Wynik przeprowadzonej analizy nie daje odpowiedzi, dlaczego to zadanie funkcjonuje w taki sposób. Jest jednak wyraźnym sygnałem, że należy dokonać analizy treści zadania ze szczególnym uwzględnieniem schematu oceniania. Może przyczyna tkwi w samym ocenianiu przez różnych egzaminatorów, którzy mają problem decyzyjny, czy za dany typ rozwiązania przyznać 1 czy też 2 punkty. Może jest to zadanie, które można rozwiązać na o wiele więcej sposobów, niż to przewidzieli autorzy i egzaminatorzy, a może po prostu istnieje ukryty błąd w treści zadania lub w schemacie oceniania. Odpowiedzi na tego typu pytania mogą dostarczyć wyniki analizy jakościowej zadania, a także sposobu oceniania uczniowskich rozwiązań.

Teoria odpowiedzi na zadanie testowe dostarcza więc o wiele więcej informacji na temat mocy różnicującej i funkcjonowania zadań, niż to ma miejsce w przypadku klasycznej teorii testu i teorii Guttmana. Przejście od zaobserwowanej ilości punktów w teście dla danego ucznia do jego poziomu umiejętności i wyrażonej w takiej samej skali trudności zadań jest centralnym zagadnieniem, którym zajmuje się IRT. Jest także punktem wyjścia do pogłębionych analiz jakościowych wyników egzaminowania w kontekście jakości arkuszy egzaminacyjnych i procesu oceniania prac przez zewnętrznych egzaminatorów.

LITERATURA

Andrich D., 2000, *Wykłady kursu: applying Rasch measurement*, Paryż.

Bieniek A., 2000, *Teoria analizy odpowiedzi na pytanie testowe,* materiały do kursu prowadzonego przez Centrum Otwartej i Multimedialnej Edukacji UW, http//www.come.uw.pl.

Perline R., Wright B. D., Wainer H., 2000, *The Rasch model as additive conjoint measurement*, http//www.rasch.org/memo 24.htm.

Suanthong S., Randal E., 2000, *An investigation of factors affecting test equation in latent trait theory,* „Journal Applied Measurement. Constructing Variables" 1, 1.

Szaleniec H., Szmigel M. K., 2001, *Egzaminy zewnętrzne. Podnoszenie kompetencji nauczycieli w zakresie oceniania zewnętrznego,* Kraków.

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