dr Kazimierz Ciżkowicz
Akademia Bydgoska

Cztery łatwości zadań wyboru wielokrotnego

„W poszukiwaniu wiedzy nie ma punktu
zerowego, nie istnieje nieskończone źródło,
z którego tryskałaby pewność."

Leszek Kołakowski

Szkic do studium przypadku

Przed kilku laty Test Kompetencji Szkolnych Absolwenta Ósmej Klasy, opracowany w ośrodku wałbrzyskim, zastąpił w pięciu województwach tradycyjny egzamin do szkół średnich. Przykładowe pytania z tego testu, potem z wielu innych nieźle naśladujących wzorce unijne, po wydrukowaniu w gazetach wzbudziły emocje nauczycieli, rodziców, mniejsze uczniów i specjalistów. W kolejnych latach „twórcze" pomysły metodyków - pomiarowców wypełniały cyklicznie wkładki lokalnych gazet, na przemian zadaniami i ich rozwiązaniami. Wzrost popytu na prasę wtorkowo-sobotnią, zwiększona podaż na rynku korepetycji, rytuał przekazywania z pokolenia na pokolenie przyszłych licealistów skrzętnie gromadzonych baz zadań egzaminacyjnych, lawinowy przyrost zbiorów zadań testowych o zróżnicowanej jakości na półkach księgarskich, to wybrane wskaźniki nowego zjawiska oświatowego, które uruchomiła reforma edukacji w Polsce.

Media operowały częściej czernią niż bielą w tworzeniu portretu „Testomanii", jeśli użyć nagłówka jednego z doniesień. W grupie wyważonych publikacji za reprezentatywną można uznać rozmowę z maja 1998 roku, którą przeprowadzono w Gazecie Wyborczej z kierownikiem zespołu naukowego wspierającego merytorycznie wałbrzyski eksperyment. W centralnej części tekstu opatrzonego znamiennym tytułem „Testy Prawdy?" wystąpiły kolejne pytania: „Czy test jest idealny?", „(...)co taki test mierzy?". Pytania te, na pierwszy rzut oka, są naiwne, ale jak poucza W. Szymborska w wierszu „Schyłek wieku"

(...) nie ma pytań pilniejszych
od pytań naiwnych.

W odpowiedzi na wspomniane pytania wystąpiły w rozmowie takie terminy, jak: obiektywne narzędzia, pilotaż, rzetelność testu, trafność kryterialna, wystandaryzowany test. Próba ich objaśnienia zasiała zapewne ziarno wiedzy psychometrycznej u przeciętnego czytelnika i, co ważniejsze, przybliżając instrumentarium naukowe towarzyszące testowaniu prawdopodobnie wzbudziła w nim poczucie zaufania do tej formy oceny kompetencji uczniów.

Bez względu na wynik toczącego się równolegle dyskursu wokół pomiaru i treści egzaminów doniosłych, pytania te pozostają istotne zarówno w odniesieniu do diagnozy osiągnięć uczniów, jak i rozpoznawania kompetencji percepcyjnych i umysłowych w zewnątrzszkolnym badaniu (por. Niemierko 1999, Konarzewski 2000). Będą one powracały w kolejnych latach i to nie tylko pod adresem ministerialnej komisji egzaminacyjnej. Próbę odpowiedzi na nie podjęto również na tym sympozjum, zawężając ich zakres do przydatności zadań wyboru wielokrotnego (WW) w pomiarze testowym. Znawca klasycznej teorii testu D. Magnuson uzasadniłby to następująco: **„Charakter pojedynczego zadania określa wiarygodność całego testu. Nie znajdujemy w teście takiej właściwości, której nie można by wyprowadzić z właściwości pojedynczego zadania lub stosunków między nimi"** [1981, s. 292]. Dodam, iż niemal wszystkie własności zadania są funkcją ich trudności. Jej wskaźnikiem jest łatwość zadania. Tak jak niegdyś z zachowania białego konia

składanego w ofierze Świętowitowi — bogu o czterech obliczach — przewidywano przyszłość, tak z czterech odmian wskaźnika łatwości można wnioskować o użyteczności zadań WW. Ich wprowadzenie poprzedzę uwagami dotyczącymi podatności zadań WW na procedurę standaryzacji.

„Zjadacze liczb" vs „zapatrzeni w pępek"

Proces poznania naukowego można ujmować jako „grę z naturą", a jej reguły są przedmiotem zainteresowania metodologii. Znajomość osobliwości konstruowania zadań, w tym WW, przez badacza kompetencji i formalnie poprawnego interpretowania jego wyników przez analityka - praktyka jest więc istotnym elementem jego/ ich świadomości metodologicznej (Brzeziński, 1984). Można też dodać - jej, jeśli uwzględnimy paradygmat kobiecości, jeden z wielu względnie nowych na obrzeżach filozofii nauki. W programie nauki proponowanej przez feministki uznaje się, że w procesie badawczym męskie podejście znamienne jest uporządkowaniem, ilością, manipulacją i kontrolą. Należy je uzupełnić kobiecym, które odznacza się obserwacją w warunkach naturalnych, wrażliwością na istotę, wzorcami jakościowymi [por. Miluska, 1994]. Interpretuje to przesłanie jako konieczność komplementarnej ilościowo-jakościowej analizy danych, w szczególności wyników zadań WW oraz ich kontekstu.

Aby ocenić przydatność tych zadań, należy w analizowaniu ich właściwości zastosować procedury cykliczne, ukierunkowujące uczenie się eksperta (twórcy zadania, sędziego kompetentnego) w toku jego analizy. Te procedury są znaczące w podejściu jakościowym, czyli badaczy „zapatrzonych w pępek" [Miles i Huberman, 2000]. Należy je uzupełnić analizami sekwencyjnymi, prowadzonymi według jasno zdefiniowanych metod, co charakteryzuje podejście ilościowe, czyli „zjadaczy liczb". Zadania zamknięte WW są szczególnie podatne na powyższe procedury, których celem jest uszlachetnianie zadań. Poddajemy ocenie nie tylko poprawność merytoryczną, językową i redakcyjną zadania, lecz również ich stosowność będącą aspektem jakościowym trafności oraz rzetelność zadania, czyli aspekt ilościowy trafności [Niemierko, 1990].

Na uwagę zasługuje tu również propozycja zgłoszona przez J. Brzezińskiego [1998], który z punktu widzenia zasady holizmu, przedstawił „itemową" analizę trafności testu opartą na analizie wariancji. Łączy ona procedurę ustalania mocy dyskryminacyjnej itemów i ich trafności (a nie całego testu!), co autor określa mianem ustalania ich dobroci. Inne podejście, stosowane w standaryzacji testów psychologicznych, jest związane z teorią wielopoziomowego różnicowania, w której o przydatności zadania rozstrzyga współczynnik skalowalności [Machowski,1993]. Poprzestając na tych przykładach, z szerokiego nurtu nieklasycznych teorii testu, w dalszej części poddamy analizie rodzaje danych i relacji między nimi, które mogą stanowić podstawę szacowania różnorodnych aspektów łatwości zadań WW, a w konsekwencji ocenę ich przydatności.

Cztery ćwiartki teorii danych

Jeśli przyjmiemy, że teoria pomiaru określa warunki konstruowania testu, a skalowanie jest procesem przypisywania liczb ujawnionym w testowaniu właściwościom badanych, to przenikanie się znaczeń tych pojęć ujmuje w spójny system klasyfikacyjny teoria danych.

W sposób uproszczony, ale czytelny, przedstawia ją C.H.Coombs (1977) jako wynik skrzyżowania dwóch dychotomicznych podziałów dotyczących macierzy obserwacji. Po sprecyzowaniu stanowiska teoretycznego badacz może wyodrębnić między jednostkami analizy rodzaj relacji: porządku, bliskości. Ze względu na zasięg diagnozowanych relacji między elementami można je analizować z osobna bądź w powiązaniu. W pierwszym przypadku ustalamy relację między elementami jednego zbioru, a w drugim dwu zbiorów — dokładniej podzbiorów hipotetycznego zbioru, które mają zróżnicowaną interpretację empiryczną (np. osób i zadań). Odpowiednio do tych sytuacji dane stanowią kwadratową i symetryczną macierz lub prostokątną (zwykle jest zróżnicowana liczebność osób i zadań) i nie wykazującą symetrii względem głównej przekątnej. Na rycinie 1, wzorując się na pomyśle Coombsa i wprowadzonej terminologii, przedstawiono cztery „ćwiartki" teorii danych. Oznaczono je literami A, B , C , D, co podkreśla, moim zdaniem, elementarność tego ujęcia. Poniżej przedstawimy po jednym istotnym modelu skalowania dla zilustrowania macierzy danych i zarazem reprezentatywnym dla analizy zadań WW. Zakładamy, że macierz obserwacji dotyczy badań pilotażowych poprzedzających zastosowanie zadań w egzaminach doniosłych.

Zasięg relacji	Rodzaj relacji	
	Porządku (dominacji)	Bliskości (zgodności)
Jeden zbiór (z osobna)	Porównanie (A)	Podobieństwa (B)
Dwa zbiory (w powiązaniu)	Dominacje (C)	Preferencje (D)

Ryc.1. Cztery „ćwiartki" teorii danych

Porównywanie — wskaźnik subiektywnej łatwości zadania

Oszacowanie stosowności zadania umożliwia analiza zgodności czynności wykonywanej przez badanego dla rozwiązania zadania z kategorią sprawdzanych czynności. Kategoria taka obejmuje rodzaj kompetencji (np. czyta, pisze) i umiejętności (np. przeprowadza operacje logiczne, rozwiązuje problemy) związane z materiałem humanistycznym, matematycznym lub przyrodniczym. Jedną z empirycznych metod wglądu w proces rozwiązywania zadań WW jest informacja zwrotna od ucznia z grupy pilotażowej uzyskana bezpośrednio po rozwiązaniu testu.

Proponuję badanie względnej trudności zadań testu techniką porównywania parami, co odpowiada skalowaniu jednego zbioru danych według relacji porządku — ćwiartka (A) teorii danych. Kierując się planem testu i opinią ekspertów, należy wybrać próbę kwotową zadań i ich numery połączyć losowo w pary. Próba zadań (k) powinna być mała, gdyż liczba koniecznych porównań szybko wzrasta z jej liczebnością ($k \cdot (k-1) / 2$). I tak dla 6 zadań — reprezentantów testu — liczba porównywanych par wynosi aż 15.

Zadaniem badanego jest podkreślenie w każdej parze zadań z przedstawionej listy numeru tego zadania, które jest według niego łatwiejsze, wybór jest więc wymuszony. W macierzy danych, takich jak przedstawiono na ryc.2 A, zakodowano jedynką obserwację, że zadanie umieszczone w wierszu jest w opinii badanego łatwiejsze niż odpowiadające mu w kolumnie. Zero oznacza sytuację przeciwną. Macierz danych jest więc kwadratowa i symetryczna.

	1	2	3	4
1	x	0	1	1
2	1	x	1	0
3	0	0	x	1
4	0	1	0	x

A, B

	1	2	3	4
A	1	1	1	0
B	0	1	1	0
C	1	1	0	1
D	1	0	0	0
E	0	1	0	0

C, D

Ryc.2. Hipotetyczne macierze danych odpowiadające „ćwiartkom" A, B, C, D teorii danych

Stosując metodę porównywania parami, możemy dokonać oceny niespójności wyborów, czyli pogwałcenia przez badanego zasady przechodniości. Jeżeli skupimy uwagę na zadaniu 1, 3, 4, to ponieważ uczeń uznał zadanie 1 za łatwiejsze niż 3, ponadto 3 niż 4, to również spodziewamy się, że zadanie 1 będzie łatwiejsze niż 4. Ponieważ takiego dokonano wyboru, to w tej triadzie zadań oceny są spójne. Łatwo stwierdzić, iż niespójna jest triada zadań 2, 3, 4. Można ocenić, czy stopień spójności danych jest większy niż ten, jaki

mielibyśmy prawo oczekiwać przy losowym dokonywaniu wyborów (por. Ferguson, Takane,1997, s.449). I na tej podstawie przy założonym poziomie ufności możemy wykluczyć z dalszej analizy oszacowania tych uczniów, których niespójność jest większa od krytycznej.

W jednej tablicy zestawimy wybory pozostałych uczniów z próby pilotażowej, zakładając zróżnicowany poziom ich zdolności. Tym razem w danej klatce macierzy znajdujemy liczbę wyborów będącą sumą wyborów dokonanych przez wszystkich badanych. Wprowadźmy subiektywny wskaźnik łatwości zadania, jako frakcję jego wyborów w zbiorze k zadań i próbie badanych o liczebności N:

$$\overline{PS_i} = \frac{\text{liczba wyborów...i-tego...zadania}}{(k-1) \cdot N}$$

Na podstawie tego wskaźnika można uporządkować zadania na subiektywnym kontinuum łatwość - trudność i porównać z „obiektywnym" wskaźnikiem łatwości zadań, który wyznacza się jako częstość podawania poprawnych odpowiedzi przez badanych na zadanie. Oba wskaźniki mają ten sam zakres wartości, który zmienia się od zera do jeden. Obie skale są niewspółmierne, co jednak nie wyklucza porównywania pozycji tego samego zadania na każdej z nich. Sytuacją ostrzegawczą jest niezgodność pozycji zadania, np. subiektywnie zadanie jest bardzo łatwe, a jednocześnie prawie nie rozwiązywane w próbie lub gdy w opinii większości badanych zadanie stawia wymogi znacznie przekraczające ich możliwości, a jednocześnie jego opuszczenia lub wybór któregoś z dystraktorów jest bardzo rzadki.

Podobieństwo — wskaźnik zgodności łatwości zadań

Wyniki dwóch zadań punktowanych dychotomicznie można zebrać w tablicy czteropolowej będącej skrzyżowaniem poprawnych (1) i niepoprawnych (0) odpowiedzi na każde z nich. Najczęściej stosowaną miarą współzmienności jest współczynnik φ Yule'a, szczególny przypadek współczynnika korelacji liniowej Pearsona. Jego wartość jest zależna od łatwości poszczególnych zadań oraz proporcji osób udzielających poprawnej odpowiedzi w obu zadaniach. Inaczej jest to frakcja badanych, dla których wartości obydwu zmiennych wynoszą jeden, co nazwiemy wskaźnikiem zgodności łatwości zadań (P_{ij}). Gdy dwa zadania wykazują łatwość umiarkowaną, równą 0,5, to współczynnik korelacji czteropolowej upraszcza się do postaci $\varphi = 4 \cdot P_{ij} - 1$. Rozważmy próbę standaryzacyjną o liczebności 100 i załóżmy, że oba zadania rozwiązało 40 badanych. Współczynnik korelacji przyjmie wartość dodatnią i wysoką równą 0,6. Kładąc nacisk na operacyjną interpretację, można kwadrat współczynnika φ traktować jako miarę skuteczności przewidywania. W analizowanym przykładzie przy zaliczaniu poszczególnych osób do jednej z dwóch kategorii zadania wiedza o tym, do jakiej kategorii należą te osoby w drugim zadaniu powoduje zmniejszenie o 36% ilości popełnionych błędów w stosunku do sytuacji, gdy nie dysponujemy wiedzą o drugim zadaniu.

Macierz danych odpowiadających ćwiartce (B) z ryc. 1 będzie więc macierzą interkorelacji par wszystkich zadań testu. Miarą bliskości jest wartość bezwzględna z φ. Jeśli przyjmiemy wartość krytyczną, odpowiadającą przyjętemu poziomowi istotności, to możemy zakodować jedynką zadania powyżej wartości krytycznej i zerem poniżej. Macierz danych będzie analogiczna do tej, którą analizowano w przypadku porównań (ryc. 2B).

Na podstawie macierzy interkorelacji można ustalić odległość między parami zadań według semimetryki: $d_{ij} = 1 - |\varphi_{ij}|$. Na tej podstawie przeprowadzamy analizę struktury skupieniowej testu, której graficzną postacią jest dendrogram (por. Marek, 1989). Prowadzi to do wyodrębnienia grup zadań najbardziej podobnych do siebie. Pomocniczym kryterium podziału jest znacząca zmiana współczynnika skupienia między kolejnymi poziomami analizy, co oznacza wzrost zróżnicowania wewnątrz grup na skutek łączenia zadań.

Jeśli empirycznie odkryta struktura kompetencji i/lub umiejętności jest zgodna z założoną w planie testu, to mamy informacje o jego trafności teoretycznej. Zadania niezgodne z założoną strukturą, bądź nie powiązane z innymi, wykazują niższą przydatność niż pozostałe i ich włączenie do ostatecznej wersji testu powinno być rozważone.

Dominacje — wskaźnik obiektywnej łatwości zadań

Do macierzy prostokątnej danych prowadzi uwzględnienie w analizie dwóch różnych zbiorów, np. uczniów (wiersze macierzy) i zadań (jej kolumny), tak jak to przedstawiono na ryc. 2C. Wartość jeden w macierzy oznacza, że badany z danego wiersza rozwiązał zadanie z odpowiedniej kolumny, zaś zero — brak rozwiązania lub rozwiązanie nieprawidłowe. Wskaźnik obiektywnej łatwości zadania, to frakcja poprawnych rozwiązań, która zgodnie z przyjętym tu modelem pomiaru umożliwia rangowe uporządkowanie zadań ze względu na ich trudność.

Problem liniowego powiązania kompetencji (umiejętności) z wskaźnikiem łatwości jest tu mniej dotkliwy niż wtedy, gdy jest on interpretowany jako prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie w populacji badanych. Przyjmuje się, że prawdopodobieństwo to jest funkcją trudności zadań, a związek ten ma kształt esowatej krzywej kumulatywnej rozkładu normalnego (Guilford, 1988).

Powracając do interpretacji odpowiadającej skali porządkowej, przyjmiemy, że jedynka w klatce macierzy oznacza „zdominowanie" przez ucznia odpowiadającego mu zadania, zaś zero — przypadek przeciwny. Tak zinterpretowany zbiór informacji można analizować według metody skalogramowej, co odpowiada ćwiartce (C) teorii danych (por. ryc.1).

Metodę skalogramu Louis Guttman przedstawił ponad pół wieku temu w artykule dotyczącym skalowania danych jakościowych. Można ją interpretować dwojako: jako kryterium lub technikę. W pierwszym znaczeniu jest procedurą testowania opisywanej trafności danego modelu pomiaru, sprawdzaniem hipotezy o jednowymiarowej ukrytej zmiennej, będącej podstawą „dominacji" badanych względem zbioru przedstawianych im możliwości (zadań). Jeżeli traktujemy ją jako technikę skalowania, czyli przypisywania symboli (liczb, rang) obiektom lub właściwościom (Coombs, 1997 s.57), to konstruujemy jednowymiarową przestrzeń opierając się na geometrycznej reprezentacji obserwowanych odpowiedzi na bodźce. Skala ta ma wyjaśnić zaobserwowane dominacje. W tym celu przypisujemy rangi poszczególnym obiektom, zarówno osobom, jak i odpowiedziom na zadania testowe według przyjętego wcześniej modelu pomiaru.

Jednoczesne uporządkowanie zadań testu według malejącej łatwości oraz badanych według malejącego wyniku surowego testowania prowadzi do macierzy „trójkątnej", jako wynik permutacji kolumn i wierszy. Umożliwia to uporządkowanie na jednym kontinuum zadań i badanych. Ponieważ dla większości danych rzeczywistych struktura taka jest dobra tylko w przybliżeniu, to osoby i zadania wyłamują się z liniowego uporządkowania (np. dane z ryc. 2C po dokonaniu permutacji).

Jeżeli skupimy uwagę na zadaniach i wybierzemy do analizy łatwiejsze i trudniejsze, to rozwiązanie trudniejszego powinno przesądzać o rozwiązaniu łatwiejszego, zaś nierozwiązanie łatwiejszego determinuje taki sam wynik trudniejszego. Stąd w każdej tabeli czteropolowej, jakie analizowano powyżej, powinna wystąpić liczba zero w klatce, która odpowiada rozwiązaniu trudniejszego i nierozwiązaniu zadania łatwiejszego. Takie zdeterminowanie wystąpi rzadko i wówczas w tym polu, gdzie powinno być zero, wystąpią względnie małe liczebności badanych, świadcząc, że skala nie jest „doskonała". Frakcja tych liczebności odjęta od jedności to współczynnik odtwarzalności zadania. Jego wartość powyżej 0,90 wskazuje na zadowalającą jednorodność wyników — poprawność skalogramu. Zadania wykazujące obniżoną odtwarzalność powinny być wyłączone z ostatecznej wersji testu, jako nieprzydatne z punktu widzenia założonego modelu pomiaru.

Preferencje — wskaźnik niezgodności łatwości zadania

Czy można przewidywać, do jakiej klasy wyników testowania należy badany na podstawie rozwiązania wybranego zadania z tego testu? Jakie zebrać dane w badaniach pilotażowych o poszczególnych zadaniach, aby zwiększyć szansę trafnej predykcji o rozkładzie kompetencji w badanej próbie (populacji)? Odpowiedzi na takie pytania może dostarczyć analiza dyskryminacyjna. Zanim wykorzystamy ją w rozwiązaniu tych problemów, rozważmy jaki rodzaj danych to umożliwi.

Przykładem relacji zgodności między elementami różnych zbiorów, czyli takiej macierzy obserwacji jaką przedstawiono na ryc. 2D, jest jej klasyfikacja. Ustalmy odpowiedniość między zbiorem punktów odpowiadających wyodrębnionej kategorii odpowiedzi na zadanie a zbiorem punktów odpowiadających klasom

kompetencji uczniów. Jeśli porównamy punkty z jednego zbioru z punktami drugiego zbioru, to na tej podstawie możemy określić ich wzajemną bliskość.

W celu uproszczenia analizy dokonamy podziału dychotomicznego zarówno badanych, jak i odpowiedzi na zadania. W przypadku badanych po ich niemalejącym uporządkowaniu według surowych wyników testowania dokonamy podziału dychotomicznego medianą na klasę kompetencji wysokich i niskich. Jeśli próba pilotażowa jest duża (N = 300) możemy zastosować kwartyle i wyodrębnić dwie klasy o bardzo wysokich i bardzo niskich kompetencjach. W przypadku zadań zamkniętych WW będzie to podział dychotomiczny według kategorii odpowiedzi. Ustalmy teraz odpowiedniość między wyodrębnionymi klasami (punktami), wprowadzając wskaźnik niezgodności łatwości zadania jako różnicę łatwości zadania w górnej (P_g) i dolnej (Pd) grupie uczniów, odpowiednio o wysokich i niskich kompetencjach. Wskaźnik taki jest znany od lat czterdziestych w psychometrii i stosowany jako miara mocy dyskryminacyjnej zadań. W pomiarze dydaktycznym na użytek testów nauczycielskich upowszechnił go w Polsce B. Niemierko [1975], jako wygodny w obliczeniach, zastępczy wskaźnik mocy różnicującej zadań D_{50}.

Poniżej podano jego nową operacyjną interpretację i taką jego modyfikację, która pozwoli uznać go za „pełnoprawny" współczynnik mocy różnicującej zadań. Przywołując analizę dyskryminacyjną i przypadek binarnych zmiennych, można udowodnić, że wskaźnik mocy predykcyjnej przy minimalizacji prawdopodobieństwa błędnej decyzji [Niewiadomska-Bugaj, 1988] upraszcza się do wskaźnika oceny rozbieżności dwóch rozkładów, gdy jedna ze zmiennych ma rozkład symetryczny. Odpowiada to dokładnie bezwzględnej wartości różnic łatwości, czyli $|P_g - P_d|$. Jeśli rozważymy przykład zadania i wyników testowania, które pozwalają ustalić, że $P_g = 0{,}7$ zaś $P_d = 0{,}1$, to ich różnicy można nadać następującą interpretację.

Dysponując informacją o zróżnicowaniu badanych według tego, czy rozwiązali to zadanie, eliminujemy 60% błędów przewidując czy badany należy do klasy wysokich bądź niskich kompetencji, w odniesieniu do przypadku, gdybyśmy nie dysponowali taką informacją.

W innym miejscu udowodniono ponadto, że pierwiastek z bezwzględnej różnicy łatwości jest równy współczynnikowi Hellwiga — φ [Ciżkowicz, 1999]. Oznacza to powiązanie miary zależności stochastycznej ze wskaźnikiem predykcji. W rozważanym przykładzie wartość współczynnika zależności stochastycznej wyniosłaby 0,78, czyli moc różnicująca tego zadania byłaby bardzo wysoka.

Użyteczność rozważanej miary w analizie zadań WW wydaje się dość oczywista. Zadania o ujemnej wartości wskaźnika ($P_g - P_d$) powinny być usunięte z końcowej wersji testu. W przypadku wartości dodatniej podejmowanie decyzji ostrzegawczych dotyczących zadania wiązałbym raczej z analizą porównawczą wartości współczynnika φ, a nie jedną wybraną wartością krytyczną.

Zakończenie

Zastosowanie przedstawionych procedur analizy zadań testowych, wspomagające ocenę ich przydatności w egzaminach doniosłych, ma charakter hipotetyczny. Wnioski praktyczne, które można by sformułować są zależne od decyzji dotyczących modelu pomiaru. Są one odmienne w przypadku różnicowania czy selekcjonowania badanych. Będą też zróżnicowane, gdy oczekujemy prostokątnego, normalnego bądź dwumodalnego rozkładu wyników testowania.