POLSKIE FORUM PSYCHOLOGICZNE, 2001, TOM 1(6), NUMER 1, s. 53-60

CZTERY ŁATWOŚCI ZADAŃ WYBORU WIELOKROTNEGO

Kazimierz Ciżkowicz*
Wydział Pedagogiki i Psychologii
Akademia Bydgoska

FOUR EASINESSES OF MULTIPLE CHOICE TASKS

Summary. In this article the estimation of usefulness of multiple choice tasks was based on logical statistical procedures concerning the structure of data. Susceptibility of tasks on qualitative and quantitative standarisation procedures was connected with their suitability and reliability. Four realtions were distinguished: (1) realtion of comparing, (2) realtion of similarity, (3) realtion of domination, (4) realtion of preference. Each realtion was connected with different aspect of tasks casiness.

Wprowadzenie

W prezentowanym opracowaniu podejmuję próbę odpowiedzi na pytanie o przydatność zadań wyboru wielokrotnego (WW) w pomiarze testowym. Zdaniem znawcy klasycznej teorii testu D. Magnusona (1981, s. 292) „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". Dodam, iż niemal wszystkie własności zadania są funkcją ich trudności. Jej wskaźnikiem jest łatwość zadania. Przedstawię cztery odmiany wskaźnika łatwości, z którego można wnioskować o użyteczności zadań WW.

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 obserwa-

* Korespondencję kierować pod adresem: Kazimierz Ciżkowicz, email: cizbar@ab-byd.edu.pl

cji. 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 relacje 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. W tabeli 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 modelu skalowania dla każdej ćwiartki danych reprezentatywnych dla analizy zadań WW. Zakładamy, że każda z czterech macierzy obserwacji dotyczy badań pilotażowych poprzedzających zastosowanie zadań w egzaminach doniosłych.

Tabela 1. Cztery „ćwiartki" teorii danych

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

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 rodzaje 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 osoby badanej 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 ($k^*(k - 1)/2$). 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 rycinie 1A, zakodowano jedynką obserwację, że zadanie umieszczone w wierszu jest w opinii badanego łatwiejsze niż odpowia-

dające 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

Rycina 1. 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 badany uznał zadanie 1 za łatwiejsze niż 3, ponadto 3 za łatwiejsze niż 4, to również spodziewamy się, że zadanie 1 będzie łatwiejsze niż 4. Takiego właśnie dokonano wyboru (por. rycina 1A), czyli 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 istotności możemy wykluczyć z dalszej analizy oszacowania te osoby, których niespójność jest większa od krytycznej.

Decyzje pozostałych badanych z próby pilotażowej można zestawić w macierzy zbiorczej, sumując ich wybory cząstkowe dotyczące par poszczególnych zadań. Tym razem w danej klatce macierzy znajdujemy liczbę wyborów będącą sumą wyborów dokonanych przez wszystkich badanych. Subiektywny wskaźnik łatwości zadania jest frakcją jego wyborów w zbiorze *k* zadań i próbie badanych o liczebności *N*.

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

Na podstawie tego wskaźnika można uporządkować zadania na subiektywnym kontinuum trudności 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. Skale są niewspółmierne, co jednak nie wyklucza porównywania pozycji tego samego zadania na każdej z nich.

Sytuacją ostrzegawczą jest znaczne zróżnicowanie wartości subiektywnego i obiektywnego wskaźnika łatwości danego zadania. Odpowiada ona niezgodności pozycji zadania na porównywanych skalach, np. subiektywnie zadanie jest bardzo łatwe, a równocześnie prawie nie rozwiązywane lub w opinii większości badanych stawia wymogi znacznie przekraczające ich możliwości, a jednocześnie jego opuszczenia lub wybór któregoś z dystraktorów są bardzo rzadkie.

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 $\varphi = 4^{*}P_{ij} - 1$. 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% liczby popełnionych błędów w stosunku do sytuacji, gdy nie dysponujemy wiedzą o drugim zadaniu.

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

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 dendrogramu, co oznacza wzrost zróżnicowania wewnątrz grup na skutek łączenia zadań.

Empirycznie odkryta struktura kompetencji i/lub umiejętności zgodna z założoną w planie testu, potwierdza jego trafność teoretyczną. Zadania niezgodne z założoną strukturą bądź nie powiązane z wydzieloną wiązką zadań w dendrogramie, wykazują niższą przydatność. Ich włączenie do ostatecznej wersji testu powinno być rozważone.

Dominacje – wskaźnik obiektywnej łatwości zadania

Do macierzy prostokątnej danych prowadzi uwzględnienie w analizie dwóch różnych zbiorów, np. uczniów (wiersze macierzy) i zadań (jej kolumny), tak jak to przedstawiono na rycinie 1C. 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. Przyjęty model pomiaru umożliwia rangowe uporządkowanie zadań ze względu na ich trudność.

Problem liniowego powiązania kompetencji (umiejętności) ze wskaźnikiem łatwości jest wtedy mniej dotkliwy, gdy jest interpretowany jako prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie w populacji badanych. W takim modelu 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. tabela 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 zadania. Skala ta ma wyjaśnić zaobserwowane dominacje. W tym celu przypisujemy rangi poszczególnym obiektom, zarówno osobom, jak i odpowiedziom na zadania testowe.

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 ryciny 1C po dokonaniu permutacji).

Jeżeli skupimy uwagę na dwóch zadaniach i wybierzemy do analizy łatwiejsze i trudniejsze, to rozwiązanie trudniejszego powinno przesądzać o rozwiązaniu łatwiejszego, zaś nie rozwiązanie 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 nie rozwiązania 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.

Przydatność zadań wykazujących niższe wartości współczynnika odtwarzalności powinna być rozważona. Zadania, które wykazują największą niezgodność ze skalogramem powinny być usunięte z ostatecznej wersji testu.

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 ją podejmiemy, rozważmy jaki rodzaj danych to umożliwi.

Przykładem relacji zgodności między elementami różnych zbiorów, czyli takiej macierzy obserwacji jaką przedstawiono na rycinie 1D, jest jej klasyfikacja. Ustalmy odpowiedniość między zbiorem punktów odpowiadających wyodrębnionej kategorii odpowiedzi na zadanie a zbiorem punktów odpowiadających klasom kompetencji badanych. 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. Jest to różnica łatwości zadania w górnej (P_g) i dolnej (P_d) grupie badanych, 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 operacyjną interpretację i taką 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 w analizowanym przypadku 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 – σ (Ciżkowicz, 1998). 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 wskaźnika podejmowanie decyzji ostrzegawczych dotyczących zadania wiązałbym raczej z analizą porównawczą wartości współczynnika σ, a nie jedną wybraną wartością krytyczną.

Zamiast konkluzji

Celem omówionych analiz porównawczych jest udzielenie pomocy konstruktorowi testu w podejmowaniu decyzji dotyczących przydatności zadań w egzaminach doniosłych. Logiczno-statystyczne procedury analizy zadań dotyczą struktury danych empirycznych z testowania próbnego. Zadania są oceniane pod względem ich stosowności i rzetelności, a rodzaj relacji między danymi odpowiada ich porównywaniu i podobieństwie oraz dominacji i preferencji. Z każdą wymieniona procedurą analizy związano określony wskaźnik łatwości zadania/zadań. W tabeli 2 wskaźnikom przyporządkowano podstawę decyzji ostrzegawczych dotyczącą włączenia zadania do ostatecznej wersji testu.

Tabela 2. Ocena przydatności zadań WW po testowaniu próbnym

Zakres analizy zadań	Relacje między danymi	Rodzaj wskaźnika	Podstawa decyzji ostrzegawczych
Stosowność	Porównywanie (A)	Subiektywna łatwość zadania (PS_i)	Znaczące zróżnicowanie subiektywnego i obiektywnego wskaźnika łatwości zadania
	Podobieństwo (B)	Zgodność łatwości zadań (P_{ij})	Pozycja zadania w strukturze skupieniowej testu niezgodna z jego planem
Rzetelność	Dominacje (C)	Obiektywna łatwość zadania (P_i)	Niezgodność zadania ze skalogramem; niska wartość współczynnika odtwarzalności
	Preferencje (D)	Niezgodność łatwości zadania ($P_g - P_d$)	Ujemna wartość wskaźnika; względnie mała wartość współczynnika σ

W ostatniej kolumnie tabeli wystąpiły określenia rozmyte, takie jak: znaczące zróżnicowanie, niezgodność, względnie mała wartość. Ich dokładne określenie jest możliwe wtedy, gdy konstruktor testu dokona wyboru między różnicowaniem bądź selekcjonowaniem badanych.

LITERATURA CYTOWANA

Brzeziński, J. (1997). *Metodologia badań psychologicznych.* Warszawa: PWN.

Ciżkowicz, K. (1998). Zastosowanie współczynnika Hellwiga w diagnostyce edukacyjnej. W: *Zeszyty Naukowe WSHE, Seria: Nauki Pedagogiczne,* t. 3. Włocławek: WSHE.

Coombs, C. H., Dawes, R. M., Tversky, A. (1997). *Wprowadzenie do psychologii matematycznej.* Warszawa: PWN.

Guilford, J. P. (1988). Tworzenie testu. W: J. Brzeziński (red.) *Problemy teorii, rzetelności, konstrukcji i analizy wyników testów psychologicznych.* Warszawa: PTP.

Ferguson, A. G., Takane, Y. (1997). *Analiza statystyczna w psychologii i pedagogice.* Warszawa: PWN.

Konarzewski, K. (2000). *Jak uprawiać badania oświatowe. Metodologia praktyczna.* Warszawa: WSiP.

Magnuson, D. (1981). *Wprowadzenie do teorii testów.* Warszawa: PWN.

Machowski, A. (1993). *Rzetelność testów psychologicznych. Dwa ujęcia modelowe.* Warszawa-Poznań: PWN.

Marek, T. (1989). *Analiza skupień w badaniach empirycznych. Metoda SAHN.* Warszawa: PWN.

Miles, M. B., Huberman, A. M. (2000). *Analiza danych jakościowych.* Białystok: Trans Humana.

Miluska, J. (1994). Psychologia płci jako wyzwanie dla edukacji. W: J. Brzeziński, L. Witkowski (red.) *Edukacja wobec zmiany społecznej.* Poznań-Toruń: Wydawnictwo Edytor.

Niemierko, B. (1975). *Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe.* Warszawa: WSiP.

Niemierko, B. (1990). *Pomiar sprawdzający w dydaktyce. Teoria i zastosowania.* Warszawa: PWN.

Niemierko, B. (1999). *Pomiar wyników kształcenia.* Warszawa: WSiP.

Niewiadomska-Bugaj, M. (1988). Analiza dyskryminacyjna. W: T. Bronka, E. Pleszczyńska (red.) *Teoria i praktyka wnioskowania statystycznego.* Warszawa: PWN.