dr Barbara Ciżkowicz
Akademia Bydgoska

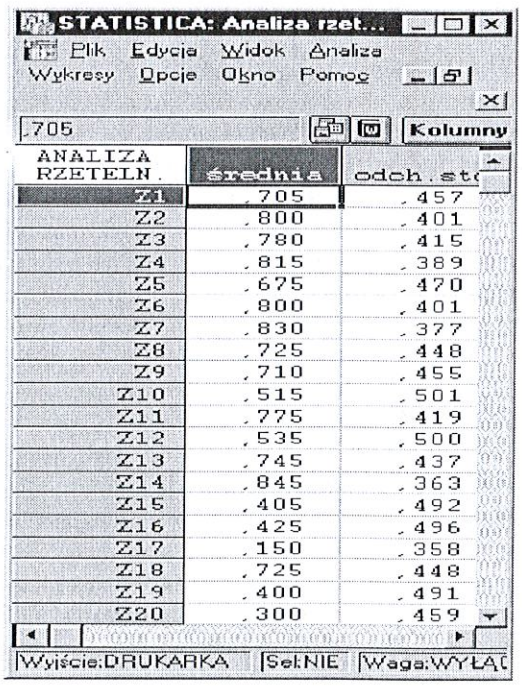
Zastosowanie pakietów statystycznych w analizie kompetencji uczniów

Jednym z utrudnień związanych ze stosowaniem testów w pomiarze dydaktycznym jest czasochłonność obliczeń towarzyszących analizom zarówno na etapie konstrukcji, jak i opracowywania wyników. Rosnący dostęp do komputerów osobistych oraz coraz szersza oferta oprogramowania użytkowego minimalizują ten problem. Diagnoza dydaktyczno-wychowawcza powinna być wspomagana przez zastosowanie komputerów w opracowywaniu trafnych i rzetelnych narzędzi badawczych, a także w analizie wyników. Dostępne profesjonalne pakiety statystyczne, takie jak: SPSS, STATGRAPHICS, SYSTAT, STATISTICA zawierają moduły umożliwiające analizę narzędzi pomiarowych i wyników pomiaru pedagogicznego. Ich niewątpliwą zaletą jest wyczerpująca lista analiz statystycznych możliwych do przeprowadzenia przy ich użyciu, począwszy od podstawowych statystyk, aż po zaawansowane metody wielozmiennowe. Umożliwiają one również wieloaspektową prezentację wyników w formie tabelarycznej i graficznej. Pakiety te mają jednak dwa istotne mankamenty: jednym są ich wysokie ceny, nie do zlekceważenia przy obecnej kondycji finansowej szkół, drugim — rozbudowane menu wymagające wprawy w skutecznym wyborze i realizacji pożądanych funkcji.

Obok wymienionych powyżej pakietów statystycznych dostępne są również mniejsze pakiety (GB-STAT) lub nawet pojedyncze programy (ITEMAN), których instalacja nie stawia wysokich wymagań dotyczących sprzętu komputerowego, a rezygnacja z rzadko wykorzystywanych modułów zmniejsza zajmowaną pamięć oraz pozwala tworzyć mniej rozbudowane menu, upraszczając użytkowanie.

Poniżej zaprezentowano możliwości zastosowania pakietu STATISTICA oraz programu ITEMAN™ w analizie testów.

Pakiet STATISTICA składa się z dwudziestu modułów, wśród których znajduje się *Analiza rzetelności/pozycji*. Moduł ten umożliwia analizę zadań testowych wielokrotnego wyboru oraz zadań wielopunktowych. Analiza przeprowadzana jest na poziomie pojedynczych zadań, wyników testów oraz baterii testów. Dla każdego zadania wyznaczany jest średni wynik oraz dyspersja odpowiedzi (ryc. 1). W przypadku zadań wielokrotnego wyboru punktowanych 0–1 średni wynik jest równy łatwości zadania. Obliczana jest również korelacja między zadaniami, której miarą można uczynić współczynnik korelacji liniowej Pearsona lub r tetrachoryczne. Wysokie wartości tych współczynników wskazują na homogeniczność zadań testu. Na ich podstawie wyliczana jest też średnia korelacja między zadaniami (ryc. 2). Dla każdego zadania program oblicza moc dyskryminacyjną, której wskaźnikiem jest współczynnik korelacji liniowej Pearsona, obliczany między wynikami zadania a wynikami w teście. Ponadto obliczane są zmiany w statystykach testu, które spowodowałyby



Ryc. 1. Łatwość i dyspersja zadań.

usunięcie zadania. Średni wynik testu, wariancja, odchylenie standardowe oraz współczynnik rzetelności α Cronbacha wyznaczane są z pominięciem rozpatrywanego zadania (ryc.2).

Rzetelność testu może być szacowana na kilka sposobów: α Cronbacha pozwala ocenić rzetelność metodą zgodności wewnętrznej, wzór Spearmana-Browna — metodą połówkową.

Uwzględniony został też w programie szczególny przypadek szacowania rzetelności metodą zgodności wewnętrznej, zaproponowany przez Hoyta (analiza wariancji). Oprócz analizy pojedynczych zadań wyznaczane są charakterystyki rozkładu wyników testowania takie, jak: średni wynik w teście, wariancja, odchylenie standardowe, współczynnik skośności oraz kurtozy rozkładu (ryc. 3). Ujemna wartość współczynnika skośności wskazuje na lewoskośny rozkład wyników, a dodatnia — na prawoskośny. Kurtoza informuje o skupianiu poszczególnych wyników wokół średniej arytmetycznej. Jeśli wartość współczynnika jest dodatnia, wówczas rozkład wyników — w porównaniu z rozkładem normalnym — ma kształt spiczasty, a gdy współczynnik przyjmuje wartość ujemną, wskazuje to na spłaszczenie rozkładu.

STATISTICA: Analiza rzetelności i pozycji - [Zestawienie skal...

Plik Edycja Widok Analiza Wykresy Opcje Okno Pomoc

11,955 Kolumny Wiersze

Dalej... Alfa Cronbacha: ,779522 Alfa standaryzowan
Średnia kor. między poz.: ,147837

zmienna	Śred.gdy usunięte	War.gdy usunięte	OdSt.gdy usunięte	Poz-Cał. Korel.	Alfa gdy usunięta
Z1	11,955	13,753	3,7085	,2848	,7743
Z2	11,860	13,910	3,7297	,2862	,7739
Z3	11,880	13,706	3,7021	,3414	,7705
Z4	11,845	13,871	3,7244	,3121	,7724
Z5	11,985	13,335	3,6517	,4006	,7664
Z6	11,860	13,900	3,7283	,2897	,7737
Z7	11,830	14,241	3,7737	,1912	,7790
Z8	11,935	13,411	3,6621	,4019	,7665
Z9	11,950	13,388	3,6589	,4008	,7665
Z10	12,145	12,924	3,5950	,4872	,7597
Z11	11,885	13,812	3,7164	,3023	,7730
Z12	12,125	13,179	3,6303	,4132	,7654
Z13	11,915	13,578	3,6848	,3601	,7693
Z14	11,815	14,181	3,7657	,2248	,7771
Z15	12,255	12,950	3,5986	,4906	,7596

Gotowy | Wyjście:DRUKARKA | Sel:NIE | Waga:WYŁĄCZONA

Ryc. 2. Analiza zadań wielokrotnego wyboru punktowanych 0–1.

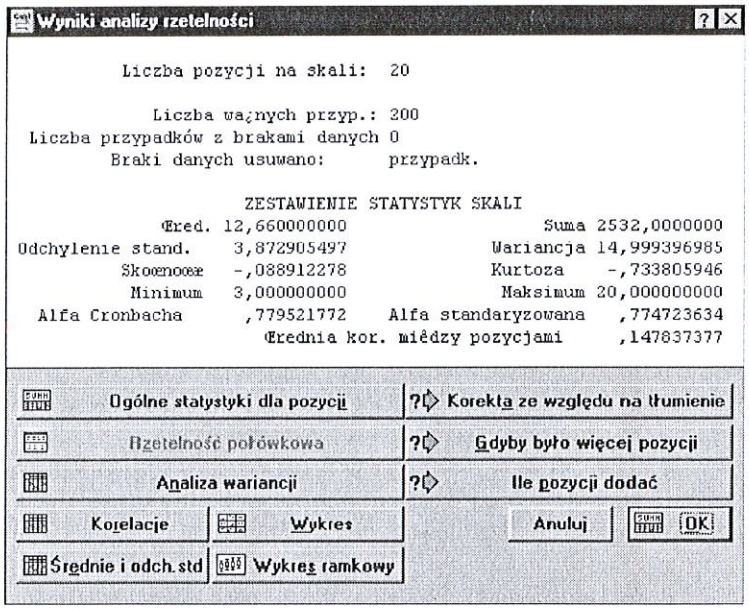
Dodatkowe funkcje wbudowane w moduł pozwalają obliczyć, o ile zadań należałoby wydłużyć test, żeby uzyskać określoną rzetelność *(Ile pozycji dodać, por. ryc. 3)*. W tym celu zastosowano wzór proroczy Spearmana-Browna. Program pozwala też odpowiedzieć na pytanie: jak zmieni się rzetelność w przypadku wydłużenia testu o określoną ilość zadań ? *(Gdyby było więcej pozycji)*. Oczywiście są to oszacowania czynione przy założeniu, że dodawane zadania cechują charakterystyki podobne do przeciętnego zadania w teście.

Oprócz szacowania rzetelności i obliczania bezpośrednio związanych z nią statystyk dla zadań pakiet Statistica może być wykorzystany w ocenie trafności. Jedna z metod szacowania trafności testu polega na porównaniu jego wyników z wynikami narzędzia badawczego przyjętego jako kryterium. Miarą trafności testu jest korelacja między tymi wynikami. Moduły zawarte w pakiecie STATISTICA pozwalają obliczyć stosowny współczynnik korelacji bez względu na rodzaj skali, na której odbył się pomiar kryterium. Tak wyznaczona wartość współczynnika korelacji ulega obniżeniu wskutek niepełnej rzetelności narzędzi badawczych. W opcji *Korekta na tłumienie* (ryc. 3) można obliczyć poprawkę Spearmana na obniżenie współczynnika korelacji, tzn. wartość, do której wzrosłaby korelacja, gdyby narzędzia badawcze nie były obciążone błędem [Niemierko, 1975]. W specyficznych przypadkach do oceny trafności można też wykorzystać analizę czynnikową lub analizę skupień. W sytuacji kiedy założenia przyjęte na etapie konstrukcji testu pozwalają oczekiwać większej homogeniczności czy to określonych grup zadań, czy określonych grup badanych, można szukać potwierdzenia tych oczekiwań w skupieniach lub czynnikach łączących obiekty lub zmienne.

Zakładanie zbiorów zawierających wyniki badań przeznaczonych do analizy oraz możliwości wykonywania operacji na tych zbiorach zapewnia moduł *Zarządzanie danymi/MFM*. Umiejętne posługiwanie się funkcjami zawartymi w tym module jest niezbędnym warunkiem pełnego wykorzystania możliwości obliczeniowych całego pakietu.

ITEMAN™ jest programem analizy testów w zintegrowanym pakiecie MicroCAT™ przeznaczonym do konstrukcji, zarządzania oraz analizowania konwencjonalnych i komputerowych testów. Niska cena i prostota obsługi stwarza realną szansę korzystania z niego w praktyce zarówno przez nauczycieli, jak i zawodowych ewaluatorów.

ITEMAN™ umożliwia analizę testów składających się z zadań zamkniętych wielokrotnego wyboru i zadań typu prawda-fałsz, które są punktowane 0–1 oraz skal postaw lub testów wielopunktowych (punktowanych od 1–9 punktów). Pozwala on również analizować baterie testów składające się z testów o różnej konstrukcji. Dopuszczalnych jest maksymalnie 10 testów w baterii.



Ryc. 3. Statystyki rozkładu wyników testu składającego się z 20 zadań wielokrotnego wyboru punktowanych 0–1.

Zbiór danych, na którym program przeprowadza analizę, zawiera informacje dotyczące sposobu odpowiadania badanych na każdą pozycję testu. Program umożliwia analizę wyników obejmujących 30 000 badanych, z których każdy rozwiązuje maksymalnie 250 zadań.

Analiza testów składających się z zadań wielokrotnego wyboru dostarcza informacji o poszczególnych zadaniach w teście, o wynikach uzyskanych przez każdego uczestnika badań oraz o rozkładzie wyników całego testu w badanej zbiorowości. Jako pierwszy pojawia się wynik analizy zadań wchodzących w skład testu. Dla każdego z nich drukowany jest kolejny numer tego zadania w teście i numer zadania w podteście (por. zał. 1). Dalej program oblicza łatwość zadania rozumianą jako stosunek liczby osób, które rozwiązały zadanie poprawnie do wszystkich, którzy rozwiązywali test. W następnej kolejności drukowany jest wskaźnik mocy różnicującej zadania D_{27}. Jest on obliczany na podstawie łatwości zadania w grupie 27% osób posiadających najwyższy wynik w teście i 27% — o najniższym wyniku. Dokładniejszą analizę mocy różnicującej zadań umożliwia obliczany również w programie współczynnik dyskryminacji, którym w zależności od decyzji prowadzącego badania może być współczynnik biseryjny lub punktowo-biseryjny. Informacje te są szczególnie przydatne na etapie konstrukcji testu. Należy przypomnieć, że współczynniki mocy dyskryminacyjnej zadania określają wysokość korelacji wyników zadania z wynikami testu. Każda z tych miar może przyjmować wartości od –1 do 1. Niska (w zależności od liczby zdań od 0,10 dla testu 100 zadaniowego do 0,45 dla 5 zadaniowego) lub ujemna wartość mocy dyskryminacyjnej wskazuje na brak jednorodności danego zadania z pozostałymi zadaniami testu [Niemierko,1999].

W zależności od rodzaju testu inne są oczekiwania co do łatwości oraz minimalnej wartości mocy różnicującej zadań. Ujemna moc różnicująca zadania w teście osiągnięć szkolnych jest efektem rozwiązywania tego zadania przez uczniów, którzy uzyskali niski wynik w teście przy jednoczesnym nierozwiązywaniu go przez uczniów, którzy z pozostałymi zadaniami radzili sobie dobrze, uzyskując wysoki wynik testowania. Dlatego też w tej sytuacji konieczna jest powtórna analiza treści zadania i jego korekta lub usunięcie go z testu. Jeżeli konstruowany test ma być testem różnicującym, łatwość zadania powinna przyjmować wartość bliską 0,5, a D_{27} i współczynnik mocy różnicującej powinny osiągać możliwie wysokie wartości. Nieco wyższe są optymalne łatwości w przypadku zadań zamkniętych. W zależności od liczby odpowiedzi, z których badany wy-

biera odpowiedź prawidłową łatwość powinna przyjmować wartość od 0,75 (dla czterech) do 0,85 (dla dwóch). Nieodpowiednie wartości tych parametrów dyskwalifikują zadanie w teście różnicującym. Nieco inne są implikacje określonych wartości tych miar w teście sprawdzającym. W tym przypadku niewłaściwa łatwość czy moc różnicująca nie decydują o usunięciu zadania z testu, a jedynie stanowią wskazanie do jego powtórnej analizy.

W zadaniach wielokrotnego wyboru program dostarcza również informacji o łatwości i mocy różnicującej poszczególnych dystraktorów. Pozwala to wyeliminować nieatrakcyjne dystraktory oraz wychwycić i przeprowadzić powtórną analizę tych, których moc różnicująca jest wyższa niż moc różnicująca werstraktora.

Dla wnikliwego ewaluatora informacje o sposobie rozwiązywania poszczególnych zadań przez uczniów są równie ważne na etapie konstrukcji testu, co na etapie badania osiągnięć uczniów. Pozwalają one ustalić, jakie treści nauczania zostały opanowane i w jakim stopniu. Analiza atrakcyjności dystraktorów umożliwia zidentyfikowanie błędów najczęściej popełnianych przez uczniów. Wiedza o efektach kształcenia uzyskana w badaniu powinna znaleźć odbicie nie tylko w ocenianiu uczniów, ale i w ewentualnych korektach nauczycielskiego systemu kształcenia.

W analizie zadań wielokrotnego wyboru program dopuszcza wprowadzenie w parametrach jednej prawidłowej odpowiedzi. Jeśli jednak istnieje uzasadnienie, by uznać jako poprawne dodatkowo inne odpowiedzi, można dołączyć je, tworząc osobny zbiór zwany zbiorem wyjątków.

Analiza skal postaw lub testów zbudowanych z zadań wielopunktowych dopuszcza zadania (pozycje) kodowane od 1–9 punktów (por. zał. 2). Podany na początku opisu numer zadania w teście i numer zadania w podteście ułatwia przyporządkowanie wynikom analizy odpowiadającej im treści. Kolumna „Item Mean" zawiera wartość średnią, która w przypadku skali postaw jest średnią aprobatą pozycji, a w przypadku zadań — średnim wynikiem osiąganym w zadaniu przez badanych. Podana dalej wartość wariancji pozwala ocenić wielkość dyspersji wyników. Miarą mocy różnicującej zadań jest w tym przypadku współczynnik korelacji liniowej Pearsona. Jego wartość jest umieszczana w kolumnie obok wariancji (Item-Scale Correlation). Ze względu na dopuszczane przez program różne warianty przetwarzania niekompletnych danych przy każdym zadaniu wyprowadzana jest liczba badanych uwzględnionych w analizie zadania. Dalej zamieszczony jest rozkład częstości wyborów odpowiedzi. W polu „klucz" znajduje się znak „+" lub „–", w zależności od deklaracji w zbiorze danych. Umożliwia to analizę w testach psychologicznych i skalach postaw pozycji odwróconych bez dokonywania zamiany punktacji przy wprowadzaniu danych. Rozkład punktów i wartość wariancji pozwala ocenić zróżnicowanie wyników lub zróżnicowanie odpowiedzi, dostarczając dokładnych informacji o funkcjonowaniu zadania w badanej zbiorowości. Interpretacja i wykorzystanie tych informacji w znacznym stopniu zależy od narzędzia badawczego, którego elementem jest rozpatrywane zadanie. Ten wariant programu umożliwia też analizę pytań zamkniętych kwestionariusza ankiety.

Na zakończenie analizy poszczególnych zadań drukowane są wyniki zbiorcze dotyczące każdego podtestu (por. zał. 3). W wydruku tym zamieszczane są kolejno informacje o: liczbie badanych uwzględnionych w analizie, liczbie zadań w podteście, średnim wyniku w podteście. Wariancja i odchylenie standardowe dają wgląd w wielkość zróżnicowania wyników. Wyznaczane są również momenty trzeciego i czwartego rzędu, na podstawie których obliczany jest współczynnik skośności i kurtozy rozkładu. W podsumowaniu zamieszczony jest też minimalny i maksymalny wynik w teście oraz wartość mediany. W zależności od rodzaju testu podawana jest średnia łatwość zadań oraz średnia moc dyskryminacyjna. Szczególnie ważny na etapie konstrukcji testu jest, używany do szacowania rzetelności, współczynnik zgodności wewnętrznej α Cronbacha. Wysokie wartości współczynnika α Cronbacha wskazują na wysoką rzetelność testu użytego w badaniach, co stanowi warunek konieczny trafności stosowanego narzędzia badawczego. W literaturze podawane są najniższe dopuszczalne wartości współczynnika, które zależą od rodzaju i doniosłości wniosków formułowanych na podstawie danych empirycznych. Niższe wymagania stawiane są przy porównaniach międzygrupowych, wyższe przy porównaniach interpersonalnych. W przypadku niezadowalającej wartości współczynnika α Cronbacha należy poddać analizie pojedyncze zadania, usuwając lub zmieniając te, których moc różnicująca jest niska.

W przypadku, kiedy analizowany test stanowi baterię testów, na zakończenie drukowana jest tablica interkorelacji między poszczególnymi podtestami. Pozwala to ocenić, czy testy są jednorodne pod względem badanych umiejętności lub postaw.

Oprócz analizy testu program redaguje zbiór z indywidualnymi wynikami badanych. W zbiorze tym zamieszczone są dla każdego badanego wyniki uzyskane w poszczególnych testach wyszczególnionych w parametrach. W zależności od wybranej opcji może to być suma punktów lub średnia liczba punktów przypadających na zadanie.

Rozbudowany identyfikator w zbiorze wejściowym umożliwia umieszczenie w nim informacji o wartościach zmiennych objaśniających (niezależnych). Pozwala to analizować wyniki w określonych podgrupach np. wprowadzenie do identyfikatora rozróżnienia ze względu na płeć pozwala analizować łącznie wszystkie wyniki, a także rozłącznie wyniki uzyskane przez chłopców i dziewczęta.

Podsumowując można stwierdzić, że decyzje odnośnie wyboru odpowiedniego oprogramowania powinny być podyktowane możliwościami i potrzebami badacza. Unifikacja podstawowych okien programów i menu głównego, konsekwentnie realizowana przez firmy zajmujące się oprogramowaniem, ułatwia poruszanie się po pakiecie osobom posiadającym doświadczenie w obsłudze standardowego oprogramowania użytkowego.

Kolejnym kryterium jest przeznaczenie i zasięg prowadzonych analiz. STATISTICA posiada szeroką ofertę sposobów prezentacji wyników oraz moduł zarządzania danymi umożliwiający sprawne operacje na zbiorach począwszy od łączenia czy dzielenia zbiorów, poprzez ich sortowanie, po przekodowywanie zmiennych i ich standaryzację. Jest to szczególnie przydatne, gdy przeprowadzane są analizy zbiorcze wyników uzyskiwanych przez różne zespoły badawcze, a efekty tych analiz mają być publikowane.

ITEMAN jest jedynym programem umożliwiającym przeprowadzanie tak obszernej analizy testów w tak prosty sposób. Nauka obsługi tego programu nie wymaga wiele czasu, co pozwala skoncentrować się na analizie wyników i ich interpretacji. Prostota obsługi ITEMAN-a umożliwia jego realizację nawet przez osoby początkujące. Warte podkreślenia jest to, że analiza wyboru dystraktorów dostarczana przez program może stanowić dla nauczyciela ważne źródło informacji o błędach w posiadanej wiedzy uczniów. Wersja 3.5 ITEMAN-a nie jest wolna od usterek. Jedną z nich jest brak zabezpieczenia programu w przypadku wystąpienia zadań z zerową wariancją. Jednak zidentyfikowany błąd nie utrudnia dalszej analizy.

Zał.1.

		Item Statistics			Alternative Statistics				
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.
1	1-1	.58	.63	.55	1	.16	.33	.10	-.32
					2	.58	.22	.85	.55
					3	.08	.06	.00	-.04
					4	.18	.39	.05	-.37
					Other	.00	.00	.00	

Zał.2.

Seq. No.	Scale -Item	Item Mean	Item Var.	Item-Scale Correlation	N per Item	Alternative	Proportion Endorsing
25	3-5	3.833	2.847	.82	48	1	.08
						2	.17
						3	.17
						4	.27
						5	.13
						6	.10
						7	.08
						Other	.04

Zał.3.

Scale Statistics

Scale:	1	2	3
N of Items	10	10	10
N of Examinees	50	50	50
Mean	4.800	7.180	3.935
Variance	5.160	6.988	2.145
Std. Dev.	2.272	2.643	1.465
Skew	0.217	-0.909	0.057
Kurtosis	-0.981	0.018	-1.423
Minimum	1.000	0.000	1.667
Maximum	9.000	10.000	6.333
Median	5.000	8.000	4.100
Alpha	0.636	0.821	0.943

Literatura

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

Niemierko B. (1975), *Testy osiągnięć szkolnych*. Warszawa: WSiP.

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

Niemierko B. *Drogi i bezdroża ewaluacji*, *Kwartalnik pedagogiczny* 1996, nr 3-4.

Niemierko B. (1997), *Między oceną szkolną a dydaktyką. Bliżej dydaktyki*. Warszawa: WsiP.

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

Noworol Cz. (1998), *Analiza skupień w badaniach empirycznych. Rozmyte modele hierarchiczne*. Warszawa: PWN.

Zakrzewska M. (1994), *Analiza czynnikowa w budowaniu i sprawdzaniu modeli psychologicznych*. Poznań: Wydawnictwo Naukowe UAM.