Maciej Koniewski
Instytut Badań Edukacyjnych, Uniwersytet Jagielloński

Zastosowanie Analizy Czynnikowej i modelowania IRT w opracowaniu skal pomiarowych, na przykładzie skali „nauczanie pod egzamin"

Wstęp

Skale są podstawowym narzędziem pomiarowym zmiennych ukrytych. Najczęściej za pomocą skal mierzy się postawy. Postawa jest to złożona i względnie trwała struktura informacji, przekonań, emocji, skłonności w stosunku do jakiegoś obiektu, która uaktywnia się pod wpływem bodźca, wywołując określone zachowanie względem obiektu.

Można sobie wyobrazić postawę obywatela państwa wobec partii politycznej, która uaktywnia się pod wpływem karty do głosowania, wywołując oddanie lub nieoddanie głosu na konkretną partię w wyborach. Inny przykład to postawa konsumenta wobec coca-coli, która uaktywnia się w momencie, gdy przechodzi obok półki sklepowej, gdzie wystawione są puszki z coca-colą. Zachowaniem wywołanym określoną postawą może być zakup napoju.

Postawa może generować różne zachowania w zależności od jej nasilenia i kontekstu sytuacyjnego. Przychylność partii politycznej może wyrażać się oddaniem głosu w wyborach, poparciem programu partii w rozmowie przy stole rodzinnym, uczestnictwem w wiecach wyborczych lub marszach. Postawa może generować inne zachowania w czasie kampanii wyborczej, a inne tuż po wyborach. Inne zachowania, gdy partia dobrze rządzi krajem, inne, gdy źle. Podobnie konsument może ograniczyć się do kupna coca-coli, może też wychwalać jej walory smakowe wśród rodziny i przyjaciół. Może też kolekcjonować gadżety, co będzie wskaźnikiem silnego przywiązania do marki. Kupno orzeźwiającego napoju będzie bardziej prawdopodobne podczas upalnego lata niż jesienią.

Przywołane przykłady pokazują, że badanie postaw wyborców i konsumentów jest ważne, aby kreować pożądane postawy, a przez to wywoływać pożądane zachowania. Podobnie jak w badaniach społecznych i marketingowych, postawy badane są z wykorzystaniem skal także w badaniach psychologicznych i edukacyjnych. Psycholog za pomocą skali depresji chce postawić trafną diagnozę, aby zaproponować właściwe leczenie. Nauczyciel za pomocą testu kompetencji językowych chce zorientować się, jakie postępy w nauce poczynił uczeń.

W przytoczonych przykładach z badań społecznych, marketingowych, psychologicznych i edukacyjnych chodzi o pomiar ukrytej zmiennej, czyli takiej, której badacz nie może bezpośrednio zaobserwować. Może ją obserwować przez wskaźniki, czyli np. zachowania, lub gromadząc odpowiedzi na pytania w kwestionariuszu czy teście.

Innym niż opisane powyżej wykorzystaniem skali jest pomiar jakiegoś zjawiska, które nie jest dostępne bezpośredniej obserwacji, np. ze względu na wysokie koszty badania. W badaniu realizowanym przez Instytut Badań Edukacyjnych (IBE) w ramach projektu „Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej (EWD)" powzięto próbę zmierzenia nasilenia zjawiska „nauczania pod egzamin gimnazjalny". Najbardziej dokładnym sposobem byłyby obserwacje lekcji. Ponieważ obserwacje są bardzo kosztowne, zestaw stwierdzeń, które miały mierzyć konstrukt „nauczanie pod egzamin", umieszczono w kwestionariuszu kierowanym do uczniów.

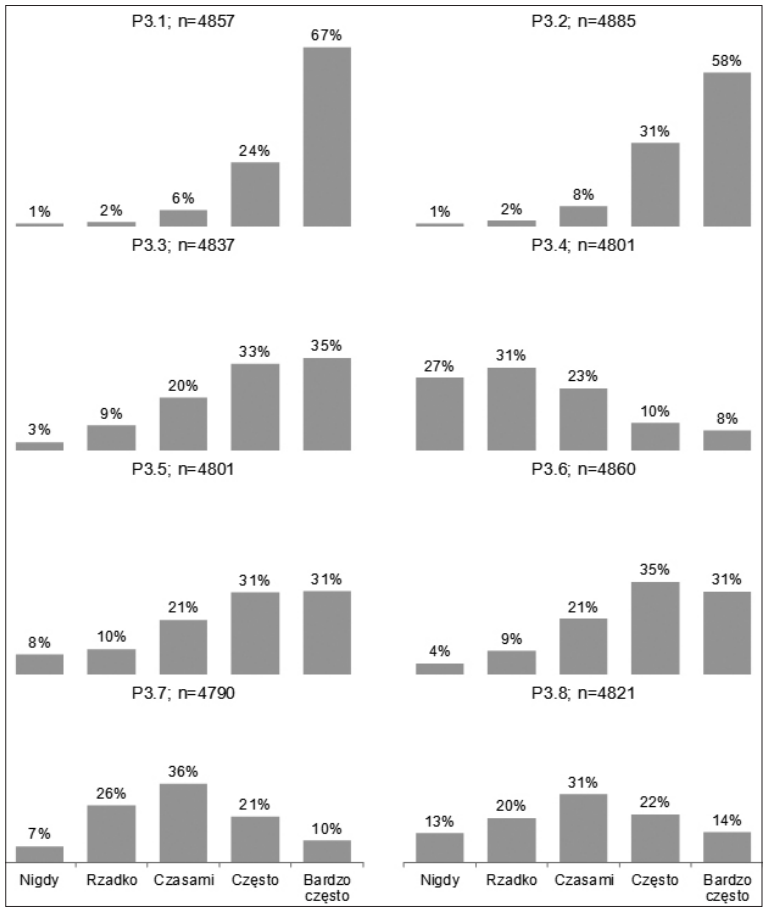
Skala „nauczanie pod egzamin" jest więc specyficzna o tyle, że bada dyspozycję nauczycieli do nauczania pod egzamin. Dane pochodzą jednak od uczniów. Skala bezpośrednio mierzy ekspozycję uczniów na praktyki nauczania pod egzamin. Jeśli zapewnimy, że uczniowie w oddziale szkolnym nauczani byli w danym okresie tylko przez jednego nauczyciela, to uśrednione wyniki uczniów na skali „nauczania pod egzamin" charakteryzują nauczyciela.

W kwestionariuszu kierowanym do uczniów, umieszczono zestaw stwierdzeń, które miały mierzyć konstrukt „nauczanie pod egzamin". Uczniów proszono, aby wskazali „Jak często podczas nauki języka polskiego w gimnazjum:

P3.1. Nauczyciel przypominał o czekającym Was egzaminie gimnazjalnym.
P3.2. Nauczyciel zwracał uwagę na umiejętności, które będą sprawdzane podczas egzaminu gimnazjalnego.
P3.3. Nauczyciel podkreślał, jak bardzo wybór szkoły ponadgimnazjalnej zależy od wyniku na egzaminie gimnazjalnym.
P3.4. Nauczyciel omijał pewne zagadnienia, wyjaśniając, że nie będzie ich omawiać, ponieważ nie są one poruszane na egzaminie gimnazjalnym.
P3.5. Nauczyciel przygotowywał sprawdziany podobne do testów egzaminacyjnych, które rozwiązywaliśmy podczas lekcji w szkole.
P3.6. Nauczyciel, omawiając sprawdziany, nawiązywał do wymagań egzaminu gimnazjalnego.
P3.7. Na sprawdzianach pojawiały się głównie zadania zamknięte.
P3.8. Podczas całej lekcji rozwiązywaliśmy arkusze egzaminacyjne".

Ten sam zestaw pytań dotyczył nauczania matematyki. W artykule wykorzystano dla przykładu tylko skalę dotyczącą nauczania języka polskiego. Wykresy kolumnowe przedstawiają rozkłady odpowiedzi.

Wzory odpowiedzi na pytania P3.1, P3.2, P3.3, P3.6, P3.5 wskazują na przewagę odpowiedzi Często i Bardzo często. Z kolei we wzorach odpowiedzi na pytania P3.4, P3.7, P3.8 duży udział mają kategorie Nigdy i Rzadko. Stosunkowo popularne wśród nauczycieli praktyki nauczania pod egzamin to przypominanie o egzaminie gimnazjalnym (P3.1); zwracanie uwagi na umiejętności, które będą sprawdzane podczas egzaminu (P3.2); podkreślanie, jak bardzo wybór szkoły ponadgimnazjalnej zależy od wyniku na egzaminie (P3.3); nawiązywanie do wymagań egzaminu podczas omawiania wewnętrznych sprawdzianów (P3.6); przygotowywanie podobnych do egzaminu sprawdzianów wewnętrznych (P3.5).



Grafika 1. Rozkłady częstości odpowiedzi uczniów na pozycje skali „nauczanie pod egzamin"

Pozytywnie należy ocenić fakt, że omijanie przez nauczycieli treści programowych, które nie są objęte pytaniami egzaminacyjnymi (P3.4), jest rzadką praktyką. Potencjalnym wyjaśnieniem tego wyniku jest to, że nauczyciele nigdy nie mogą być pewni, że dane treści nie pojawią się na egzaminie. Są też zobowiązani do zrealizowania wszystkich treści podstawy programowej.

Nauczyciele języka polskiego stosują urozmaicone formy pytań na sprawdzianach wewnętrznych, nie ograniczają się do pytań zamkniętych (P3.7). Negatywnie należy ocenić fakt, że rozwiązywanie całych arkuszy egzaminacyjnych przez uczniów na lekcjach (P3.8) jest stosunkowo popularną praktyką. Stanowi urozmaicenie lekcji, na co wskazała jedna trzecia ankietowanych uczniów. Jednak aż 36% uczniów zadeklarowało, że ich nauczyciele języka polskiego często i bardzo często poświęcają całe lekcje na ćwiczenie rozwiązywania arkuszy egzaminacyjnych.

Zastosowanie Analizy Czynnikowej w opracowywaniu skal opartych na czynniku i skal czynnikowych

Są różne sposoby obliczania pozycji[1] respondentów na skali, np. wyborcy na skali poparcia politycznego partii, konsumenta na skali przywiązania do marki, ucznia na skali kompetencji językowych. Wyróżnić można (1) indeksy, (2) skale oparte na czynniku, (3) skale czynnikowe.

Indeksy (1) to najczęściej suma odpowiedzi na pozycje (pytania) tworzące indeks. Przypisując wartości liczbowe kolejnym możliwym odpowiedziom i sumując je, otrzymamy indeks. Gdy skala składa się z 8 pozycji, przypisując 1 odpowiedzi Nigdy, 2 – Rzadko, 3 – Czasami, 4 – Często, 5 – Bardzo często, otrzymamy indeks, który przybiera wartości od 8 do 40. Kodując odpowiedzi od 0 do 4 lub przy zachowaniu kodowania od 1 do 5 i odjęciu od sumy ocen wartości 8, otrzymujemy łatwiejszy w interpretacji indeks o wartości minimalnej 0 i maksymalnej 32. Indeks możemy dodatkowo znormalizować do przedziału od 0 do 100, przez odjęcie od każdej odpowiedzi respondenta wartości minimalnej (w podanym przykładzie 1) i podzielenie otrzymanego wyniku przez wartość maksymalną (w podanym przykładzie 5). Otrzymaną zmienną możemy też wystandaryzować tak, aby średnia wynosiła 0, a odchylnie standardowe 1.

Bardziej dokładną niż indeks miarą zmiennej ukrytej jest skala oparta na czynniku (2). Aby skonstruować skalę opartą na czynniku, należy przeprowadzić Eksploracyjną (EFA) lub Konfirmacyjną Analizę Czynnikową (CFA). EFA to technika, którą stosuje się w sytuacji, gdy badacz nie ma wiedzy na temat struktury czynnikowej skali, ugruntowanej teorią i wynikami wcześniejszych badań. CFA można stosować, gdy badacz dysponuje modelem teoretycznym struktury czynnikowej. Skala oparta na czynniku to zazwyczaj (tak jak indeks) suma punktów (odpowiedzi na pytania), jednak dobór pozycji do skali nie jest arbitralny (tak jak to ma miejsce w przypadku indeksu), ale jest wynikiem analizy statystycznej. Kolejne kroki analizy to najczęściej: (1) ocena liczby wymiarów skali, (2) ocena rzetelności skali i na tej podstawie decyzja, które pozycje należy usunąć, a które pozostawić w skali, (3) ewentualnie zweryfikowanie przyjętej struktury czynnikowej w CFA.

EFA przeprowadzono w programie SPSS, wersja 20. Do wyodrębnienia czynników zastosowano metodę największej wiarygodności *(ML – Maximum Likelihood)*. Na podstawie uzyskanych odpowiedzi uczniów w badaniu realizowanym przez IBE nie było jednoznaczne, czy skalę można uznać za jednoczynnikową. Najczęściej stosowanymi kryteriami wyboru optymalnej liczby czynników są kryterium Kaisera (liczba zmiennych o wartościach własnych większych niż 1 służy za liczbę czynników), kryterium Cattella (liczbę czynników wyznacza przełamanie na wykresie osypiska), *parallel analysis* (porównanie wartości własnych na podstawie zebranych danych z wartościami własnymi oczekiwanymi w sytuacji, gdyby dane były generowane w sposób losowy) (Fabrigar i in., 1999).

[1] W artykule zamiennie stosowany jest termin „pytanie" (kwestionariuszowe) i „pozycja" (skali).

W przypadku skali „nauczanie pod egzamin", złożonej ze wszystkich ośmiu pozycji, stosunek początkowych wartości własnych pierwszego do drugiego czynnika wynosi 3,092 / 1,277 = 2,421. Optymalnie wartość własna pierwszego czynnika powinna być pięciokrotnie większa od wartości własnej drugiego czynnika. Pierwszy czynnik wyjaśnia 31% całkowitej wariancji danych, co nie jest satysfakcjonujące.

W takiej sytuacji optymalnie należałby wybrać dwa lub trzy czynniki. Jednak rozwiązania dwu- i trzyczynnikowe w tym przypadku nie są wygodne w interpretacji. Alternatywą jest usunięcie ze skali pozycji, które nisko korelują ze zmienną ukrytą i z innymi pozycjami skali. Na podstawie analizy macierzy korelacji, można wytypować pozycje P3.4 oraz P3.7 jako te, które najniżej korelują z pozostałymi pozycjami skali, jak i zmienną ukrytą.

Tabela 1. Macierz korelacji pozycji skali „nauczanie pod egzamin"

P3.1							
P3.2	,628						
P3.3	,423	,432					
P3.4	**,023**	**-,011**	**,140**				
P3.5	,306	,365	,326	,075			
P3.6	,357	,454	,366	,058	,531		
P3.7	**,088**	**,083**	**,166**	**,181**	**,284**	**,224**	
P3.8	,217	,256	,268	,083	,577	,387	,320

Ten sam wniosek można sformułować w oparciu o analizę rzetelności. Miara rzetelności Alfa Cronbacha dla skali z ośmioma pozycjami wynosi 0,734. Tylko po usunięciu P3.4 i P3.7 możliwe jest uzyskanie wyższej rzetelności. Miara rzetelności Alfa Cronbacha dla skali z sześcioma pozycjami, po usunięciu P3.4 i P3.7 wynosi 0,785.

Tabela 2. Wyniki analizy rzetelności skali „nauczanie pod egzamin"

Pozycja	Korelacja pozycji	Alfa Cronbacha po usunięciu pozycji
P3.1	,465	,706
P3.2	,493	,700
P3.3	,471	,698
P3.4	**,126**	**,770**
P3.5	,596	,668
P3.6	,559	,679
P3.7	**,316**	**,728**
P3.8	,502	,691

n = 4351

Należy pamiętać, że miara rzetelności Alfa Cronbacha może być stosowana tylko dla skal jednoczynnikowych. Po usunięciu P3.4 i P3.7 stosunek początkowych wartości własnych pierwszego do drugiego czynnika wynosi 2,988/1,056 = 2,829. Wynik wciąż daleki od optymalnej pięciokrotnej różnicy, jednak o 0,408 lepszy od wersji skali z ośmioma pozycjami. Pierwszy czynnik

wyjaśnia 40% całkowitej wariancji danych. Można to obliczyć przez dodanie kwadratów ładunków czynnikowych i podzielenie otrzymanej sumy przez liczbę czynników. W przypadku skali „nauczanie pod egzamin" to $(0{,}667^2 + 0{,}720^2 + 0{,}580^2 + 0{,}634^2 + 0{,}665^2 + 0{,}504^2) / 6 = 0{,}400$.

Wypracowaną w toku EFA skalę można poddać weryfikacji w CFA. Do przeprowadzenia CFA wykorzystano program Mplus, wersja 7. Wykorzystano estymację ważonych najmniejszych kwadratów *(WLS – Weighted Least Squares)*. Zdecydowano się na wykorzystanie estymacji WLS, ponieważ jest metodą rekomendowaną w literaturze dla dużych prób (>200) i wysokich ładunków czynnikowych (>0,7) (Nestler, 2013). W przypadku małej liczby wskaźników, jak ma to miejsce w przypadku skali „nauczanie pod egzamin", WLS jest rekomendowaną metodą, zamiast innej często wykorzystywanej metody dla zmiennych kategorialnych, tj. *DWLS (Diagonally Weighted Least Squares)* (Woods, 2002).

Jako miary dopasowania raportowane są statystyka χ^2 (Chi^2), miara relatywnego dopasowania CFI oraz miara przybliżonego dopasowania RMSEA. Stosuje się te miary jako standardowe i najczęściej stosowane w publikacjach raportujących wyniki CFA (McDonald i Ho 2002).

Statystyka χ^2 powinna być możliwie niska przy możliwie wielu stopniach swobody dla dobrego dopasowania. Przy istotności powyżej 0,05 uznaje się model za dobrze dopasowany. Jednak w przypadku dużych prób (>200) wynik testu jest zawsze istotny (Jöreskog, 1969). Dlatego w ocenie dopasowania kierować należy się raczej miarami RMSEA i CFI.

Zgodnie uznaje się RMSEA niższe niż 0,05 jako wskaźnik dobrego dopasowania, a wartość w przedziale 0,05-0,08 jako akceptowalne dopasowanie. 90% przedział ufności dla RMSEA nie powinien zwierać wartości > 0,8. Miara CFI powyżej 0,9 wskazuje na dobre dopasowanie (McDonald i Ho, 2002).

Dla skali „nauczanie pod egzamin", dla próby n = 4914 wartość statystyki χ^2 wynosi 854,837 przy 9 stopniach swobody przy istotności $p<0{,}001$. RMSEA wynosi 0,138 znacznie powyżej wartości 0,05, którą przyjmuje się za górną granicę dobrego dopasowania. 90% przedział ufności dla RMSEA przyjmuje zakres od 0,131 do 0,146. CFI wynosi 0,927, czyli powyżej dolnej granicy akceptowalnego dopasowania. Należy więc wnioskować, że ocena dopasowania do danych skali „nauczanie pod egzamin" nie jest jednoznacznie dobra. Interpretacja wyników analiz z wykorzystaniem skali „nauczanie pod egzamin" musi być ostrożna.

Tabela 3. Zmienne tworzące skalę „nauczanie pod egzamin" wraz z ładunkami i błędami standardowymi (w nawiasach)

P3.1	0,844 (0,009)
P3.2	0,862 (0,007)
P3.3	0,654 (0,010)
P3.3	0,841 (0,007)
P3.6	0,756 (0,008)
P3.8	0,698 (0,009)

Sumując odpowiedzi uczniów na wybrane do skali pozycje, można skonstruować skalę opartą na czynniku. Analogicznie jak w przypadku indeksu, nowo utworzoną zmienną, w celu ułatwienia interpretacji, można poddać normalizacji, np. do zakresu 0-100 lub standaryzacji.

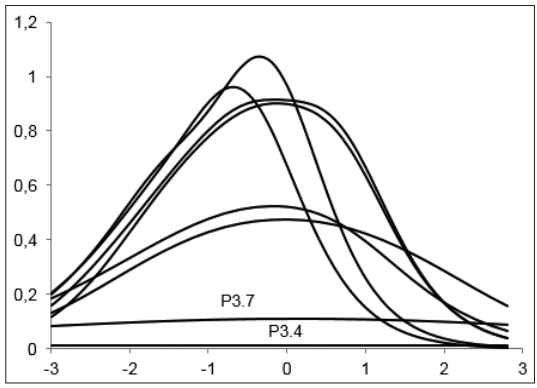
Skala oparta na czynniku może być także sumą zważonych przez ładunki czynnikowe odpowiedzi na pytania skali. Ładunek czynnikowy to korelacja danej pozycji skali z mierzoną za pomocą skali zmienną ukrytą.

Najbardziej dokładną miarą zmiennej ukrytej jest skala czynnikowa (3). Jest najbardziej zbliżona do ważonej skali opartej na czynniku. Jednak w przypadku skali czynnikowej, utworzenie nowej zmiennej z wartościami respondentów na skali nie jest ani prostą, ani ważoną sumą punktów odpowiedzi. Jest natomiast generowana automatycznie za pomocą algorytmów wykorzystywanych w EFA lub CFA.

Zastosowanie modelowania IRT w opracowywaniu skal opartych na czynniku i skal czynnikowych

Alternatywną techniką obliczania wartości respondentów na skali jest stosowanie modeli Teorii Odpowiedzi na Pozycje *(IRT – Item Response Theory)*. Jak pokazało wielu autorów CFA dla zmiennych kategorialnych *(CCFA – Categorical Confirmatory Factor Analysis)* i IRT są formalnie tożsame (np. Takane i de Leeuw, 1987; Kamata i Bauer, 2008; McDonald, 1999). Zdecydowana popularność CFA nad IRT w opracowywaniu skal jest wynikiem starszej i szerszej tradycji w stosowaniu i nauczaniu CFA, a także łatwości obliczeń, co było ważne, kiedy komputery nie miały takiej mocy obliczeniowej jak współcześnie.

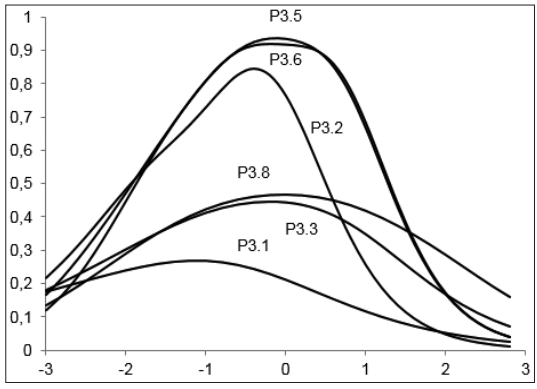
Istnieje wiele modeli IRT, które można podzielić na modele jedno- i wielowymiarowe oraz na modele dedykowane dla danych kodowanych zero-jedynkowo lub danych kodowanych większą liczbą kategorii. Dane kodowane zero-jedynkowo to np. odpowiedzi uczniów na pytania testowe, gdzie uczeń za prawidłową odpowiedź otrzymuje jeden punkt, a za błędną zero punktów. Dane kodowane większą liczbą kategorii to np. *Nigdy – Rzadko – Czasami – Często – Bardzo często*, czyli dokładnie tak, jak w przypadku omawianej tu skali „nauczanie pod egzamin". Dedykowanym modelem IRT do tego typu danych jest model GRM *(Graded Response Model)*, zaproponowany przez Samejima (1969). Dwuparametryczny (2PL) model GRM dopasowano do odpowiedzi uczniów na pozycje P3.1-P3.8. Skalowanie przeprowadzono w programie Mplus, wersja 7. W efekcie skalowania możliwe jest wygenerowanie wykresu funkcji *informacyjnych (IIF – Item Information Functions)* poszczególnych pytań, będących pozycjami (elementami) skali „nauczanie pod egzamin". Oś *y* na wykresie opisuje informację, jaką poszczególne pytanie wnosi do całej skali. Oś *x* to skala standardowa (Z), która opisuje nasilenie praktyk nauczania pod egzamin. Im wyższa wartość, tym dany uczeń (w swojej subiektywnej ocenie) jest wystawiony na bardziej intensywne praktyki nauczania pod egzamin ze strony nauczyciela. Pozycja P3.4 (krzywa niemal równoległa i położona najbliżej osi odciętych) oraz pozycja P3.7 (krzywa nieco wyżej, także niemal równoległa do osi odciętych) wnoszą najmniej informacji do skali „nauczanie pod egzamin".



Wykres 1. Krzywe informacyjne pozycji skali „nauczanie pod egzamin"

Podobnie jak w przypadku wniosków z EFA, poglądowej analizy macierzy korelacji, analizy rzetelności, także w oparciu o wykres krzywych informacyjnych, można zdecydować o usunięciu ze skali pozycji P3.4 („Nauczyciel omijał pewne zagadnienia, wyjaśniając, że nie będzie ich omawiać, ponieważ nie są one poruszane na egzaminie gimnazjalnym") oraz P3.7 („Na sprawdzianach pojawiały się głównie zadania zamknięte").

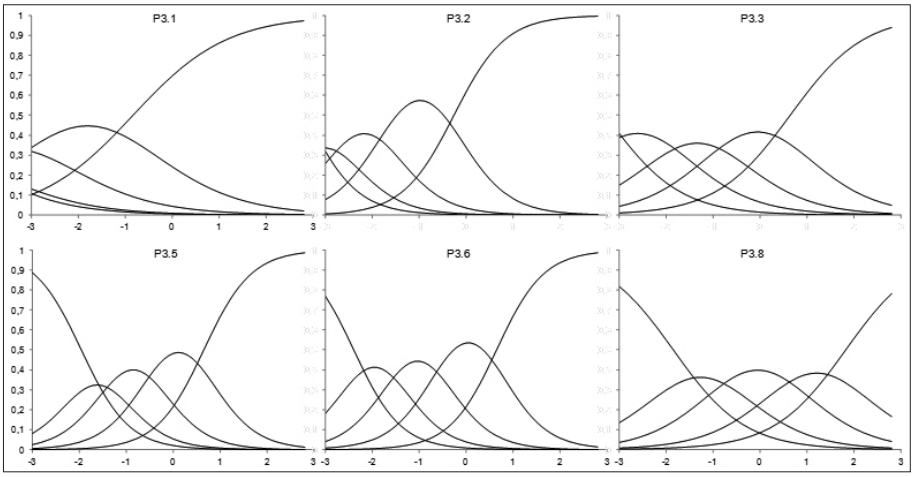
Poniższy wykres to prezentowany już wcześniej wykres krzywych informacyjnych. Tym razem z pominięciem pozycji P3.4 i P3.7. Najwięcej informacji do skali „nauczanie pod egzamin" wnoszą pozycje P3.5 i P3.6. Krzywe informacyjne tych pozycji niemal się nakładają. Można więc rozważyć usunięcie jednej z nich ze skali.



Wykres 2. Krzywe informacyjne pozycji skali „nauczanie pod egzamin" po usunięciu redundantnych pozycji

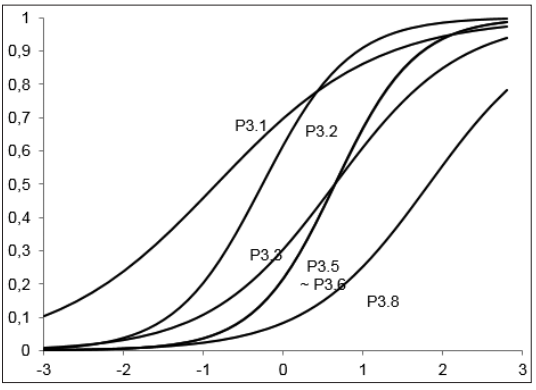
Kolejne wykresy prezentują krzywe charakterystyczne pozycji skali *(ICC – Item Characteristic Curves)*. Na każdym wykresie widocznych jest pięć linii, ponieważ na każde pytanie uczeń mógł udzielić odpowiedzi: *Nigdy – Rzadko – Czasami – Często – Bardzo często.* Im rozkłady są bardziej strome i smukłe, tym większą moc dyskryminującą ma pozycja, tzn. tym bardziej różnicuje uczniów. Tak więc pozycje P3.2 i P3.6 bardziej różnicują uczniów niż pozycje

P3.3 i P3.8. Obserwacja ta jest zgodna z wnioskami płynącymi z obserwacji prostych rozkładów odpowiedzi, prezentowanych wcześniej. Odpowiedzi na pozycje o mniejszej mocy dyskryminującej, rozkładają się bardziej równomiernie między kategoriami odpowiedzi (Nigdy – Rzadko itd.). Z kolei w pozycjach o wysokim parametrze dyskryminacji da się zaobserwować wyraźnie dominujące kategorie odpowiedzi. Odpowiedzi mają często rozkład skośny.



Grafika 2. Wykresy krzywych charakterystycznych odpowiedzi dla pozycji skali „nauczanie pod egzamin"

Ponieważ porównanie krzywych charakterystycznych pytań, które budują skalę „nauczanie pod egzamin" na jednym wykresie, jest trudne, jeśli uwzględnia się wszystkie kategorie odpowiedzi, można zsumować tylko niektóre. Poniższy wykres prezentuje krzywe charakterystyczne pytań, które sumują odpowiedzi uczniów z kategorii Bardzo często i Często.



Wykres 3. Krzywe charakterystyczne dla sumy kategorii 4+5 (b. często + często) dla pozycji skali „nauczanie pod egzamin"

Na wykresie widać tylko pięć krzywych, mimo że pozycji w skali jest sześć. Jest tak, ponieważ krzywe pozycji P3.5 i P3.6 są niemal identyczne i nakładają się na siebie. Z analizy poniższego wykresu można wnioskować, które praktyki nauczania pod egzamin są bardziej popularne, a które mniej. Jak widać, skala ma charakter gradacyjny, tzn. na przykład nauczyciele zwracający uwagę na umiejętności, które będą sprawdzane podczas egzaminu gimnazjalnego (P3.2), zwykle też przypominają swoim uczniom o egzaminie gimnazjalnym (P3.1) itd. Przypominanie o egzaminie (P3.1) jest czynnością mniej absorbującą, tym samym bardziej popularną wśród nauczycieli. Z kolei zwracanie uwagi na umiejętności, które będą sprawdzane podczas egzaminu (P3.2), jest już czynnością wymagającą większego zaangażowania, przez to nieco mniej popularną. Krzywa charakterystyczna dla P3.2 jest bardziej stroma niż krzywa dla P3.1. Dlatego też można wnioskować, że raczej większość nauczycieli przypomina uczniom o egzaminie (P3.1) i jest to praktyka powszechna. Z kolei praktyka zwracania uwagi na umiejętności, które będą sprawdzane podczas egzaminu (P3.2), dzieli nauczycieli na grupę tych, którzy taką praktykę stosują, i tych, którzy praktyki tej nie stosują. Im bardziej stroma krzywa, tym łatwiej można wyróżnić dwie grupy nauczycieli, tych, którzy stosują, i tych, którzy nie stosują danej praktyki nauczania pod egzamin. Po P3.1 i P3.2 kolejnymi pod względem częstości stosowania przez nauczycieli praktykami jest podkreślanie, jak bardzo wybór szkoły ponadgimnazjalnej zależy od wyniku na egzaminie gimnazjalnym (P3.3); przygotowywanie sprawdzianów podobnych do testów egzaminacyjnych (P3.5); nawiązywanie do wymagań egzaminu gimnazjalnego podczas omawiania sprawdzianów wewnętrznych (P3.6). Te trzy praktyki (P3.3, P3.5, P3.6) są stosowane przez nauczycieli niemal tak samo często. Praktyką nauczania pod egzamin, która lokuje się w górnych wartościach skali „nauczania pod egzamin", jest poświęcanie całych lekcji na rozwiązywanie arkuszy egzaminacyjnych (P3.8). Uczniowie, którzy wskazali, że są często lub bardzo często poddawani takiej praktyce przez swoich nauczycieli języka polskiego, otrzymywali wysokie wartości na skali „nauczanie pod egzamin". Wśród sześciu różnych praktyk nauczania pod egzamin, praktyka poświęcania całych lekcji na rozwiązywanie arkuszy egzaminacyjnych jest najrzadziej stosowana.

Wnikliwego czytelnika zachęcam to porównania wniosków z analizy wykresów krzywych charakterystycznych dwuparametrycznego modelu GRM przedstawionych powyżej z wnioskami z analizy rozkładów odpowiedzi uczniów, opisanych na początku artykułu. Z łatwością można zauważyć, że wnioski te są spójne. Także wykresy krzywych charakterystycznych i rozkładów częstości są komplementarne.

Porównanie skal

Opracowywanie skal pomiarowych w oparciu o wykresy wygenerowane w modelowaniu IRT jest atrakcyjne i szybsze dla wprawnego analityka niż tradycyjna Analiza Czynnikowa. Najlepsze efekty w aspekcie „rozumienia" danych, odkrywania i weryfikacji struktury czynnikowej daje analiza danych w ramach CFA i modelowania IRT, co jest możliwe w ramach jednego programu Mplus.

Wykorzystanie Analizy Czynnikowej i modelowania IRT do opracowywania skal pomiarowych daje bardzo zbliżone rezultaty. Poniżej prezentowana jest macierz korelacji skal opartych na czynniku, będących prostą sumą lub sumą ważoną odpowiedzi. Dla porównania prezentowane są też skale czynnikowe: EFA metodą ML, CFA metodą ML i WLS, a także skala skonstruowana z wykorzystaniem modelu GRM. Korelacje skal między sobą są bardzo wysokie i wynoszą od 0,956 do 0,999.

Tabela 4. Korelacje skal otrzymanych różnymi sposobami

Wynik egzaminu części polonistycznej (3PLM)							
Skala op. na czynn. (suma) 6 pozycji	**,035**						
Skala op. na czynn. z wag. EFA (ML) 6 poz.	**,045**	,998					
Skala op. na czynn. z wag. CFA (ML) 6 poz.	**,042**	,999	,999				
Skala czynnikowa EFA (ML) 6 poz.	**,068**	,978	,987	,983			
Skala czynnikowa CFA (ML) 6 poz.	**,059**	,971	,975	,973	,970		
Skala czynnikowa CFA (WLS) 6 poz.	**,056**	,971	,974	,974	,965	,996	
Skala czynnikowa GRM 2PL (ML) 6 poz.	**,049**	,976	,977	,977	,956	,989	,987

n = między 4635 a 4857

Tę samą macierz korelacji zaprezentowano dla danych uśrednionych na poziomie oddziałów. Do porównania dodano zmienną „efektywność nauczyciela", którą stanowią reszty z trzypoziomowego modelu regresji (HLM), uwzględniającego wynik sprawdzianu, płeć, inteligencję mierzoną testem matryc Ravena, wyższą z liczby lat nauki rodziców, zamożność gospodarstwa domowego, liczbę osób w gospodarstwie domowym, lokalizację gimnazjum.

Także na zagregowanym poziomie korelacje skal są bardzo wysokie i wynoszą od 0,924 do 0,999. Nauczanie pod egzamin koreluje na poziomie około 0,120 z wynikami egzaminu z części polonistycznej i około 0,190 z efektywnością nauczyciela.

Tabela 5. Korelacje skal otrzymanych różnymi sposobami z efektywnością nauczyciela i wynikami części polonistycznej egzaminu

Efektywność nauczyciela								
Wynik egzaminu części polonistycznej (3PLM)	**,524**							
Skala op. na czynn. (suma) 6 pozycji	**,140**	**,004**						
Skala op. na czynn. z wag. EFA (ML) 6 poz.	**,191**	**,105**	,947					
Skala op. na czynn. z wag. CFA (ML) 6 poz.	**,191**	**,106**	,948	,999				
Skala czynnikowa EFA (ML) 6 poz.	**,202**	**,122**	,924	,990	,985			
Skala czynnikowa CFA (ML) 6 poz.	**,207**	**,131**	,937	,978	,975	,975		
Skala czynnikowa CFA (WLS) 6 poz.	**,208**	**,132**	,941	,979	,979	,971	,997	
Skala czynnikowa GRM 2PL (ML) 6 poz.	**,205**	**,134**	,948	,983	,983	,969	,992	,993

n = 260

Konkludując, należy wskazać, która skala jest „najlepsza". Ponieważ błąd pomiaru powoduje obniżenie korelacji, można stwierdzić, że te skale, które wyżej korelują z wynikami egzaminu, są skalami bardziej dokładnymi. W tym przypadku będą to skale czynnikowe uzyskane w CFA i GRM. Można porównać jakość dopasowania skal. Dla CFA (ML) kryterium informacyjne Akaike (AIC) wynosi 67787,411, a Bayesowskie (BIC) 67982,406. Dla GRM 2PL (ML) AIC wynosi 67649,549, a BIC 67844,544. AIC i BIC są niższe dla modelu GRM, co przy tej samej liczbie stopni swobody (30) sugeruje lepsze dopasowanie do danych modelu GRM.

Bibliografia

1. Woods, C. M. (2002). Factor analysis of scales composed of binary Items: Illustration with the Maudsley Obsessional Compulsive Inventory. *Journal of Psychopathology and Behavioral Assessment*, *24*(4), 215-223.
2. Takane, Y., & Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408. Springer New York.
3. Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores.* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.
4. Nestler, S. (2012). A Monte Carlo study comparing PIV, ULS and DWLS in the estimation of dichotomous confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology, 66*(1), 127-143.
5. McDonald, R. P., Ho, M-H. R. (2002). Principles and Practice in Reporting Structural Equation Analyses. *Psychological Methods, 7*(1), 64-82.
6. McDonald, R. P. (1999). Test theory: A unified approach. *Mahwah, NJ: Lawrence Earlbaum*.
7. Kamata, A., i Bauer, D. J. (2008). A Note on the Relation Between Factor Analytic and Item Response Theory Models. *Structural Equation Modeling A Multidisciplinary Journal*, *15*(1), 136-153.
8. Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183-202.
9. Forero, C. G., Maydeu-Olivares, A., Gallardo-Pujol, D. (2009). Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation. *Structural Equation Modeling A Multidisciplinary Journal*, *16*(4), 625-641.
10. Fabrigar, L. R. i in. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods, 4*(3), 272-299.

Aneks (syntaksy)

```
TITLE: Uczenie pod egzamin (skala) CFA
DATA:
FILE IS „C:\(...)\(...).dat";
VARIABLE:
NAMES ARE P3.1 P3.2 P3.3 P3.4 P3.5 P3.6 P3.7 P3.8 kfullid;
USEVARIABLE ARE P3.1 P3.2 P3.3 P3.5 P3.6 P3.8;
MISSING ARE all (9);
IDVARIABLE IS kfullid;
CATEGORICAL ARE all;
ANALYSIS:
ESTIMATOR IS ML;
MODEL:
Factor BY P3.1 P3.2 P3.3 P3.5 P3.6 P3.8;
Factor@1;
SAVEDATA: SAVE = FSCORES;
FILE IS (...).dat;

TITLE: Uczenie pod egzamin (skala) GRM
DATA:
FILE IS „C:\(...)\(...).dat";
VARIABLE:
NAMES ARE P3_1 P3_2 P3_3 P3_4 P3_5 P3_6 P3_7 P3_8 kfullid;
USEVARIABLE ARE P3_1 P3_2 P3_3 P3_5 P3_6 P3_8;
```

```
MISSING ARE all (9);
IDVARIABLE IS kfullid;
CATEGORICAL ARE all;
ANALYSIS:
ESTIMATOR IS ML;
LINK IS LOGIT;
MODEL:
SKALA BY P3_1 P3_2 P3_3 P3_5 P3_6 P3_8;
 [P3_1$1-P3_8$1];
 [P3_1$2-P3_8$2];
 [P3_1$3-P3_8$3];
 [P3_1$4-P3_8$4];
 [SKALA@0]; SKALA@1;
OUTPUT: STDYX;
RESIDUAL TECH10;
SAVEDATA: SAVE = FSCORES;
FILE IS (...).dat;
PLOT: TYPE IS PLOT1;
TYPE IS PLOT2;
TYPE IS PLOT3;
```