Przemysław Majkut
Maciej Koniewski
Paulina Skórska
Instytut Badań Edukacyjnych

Wykorzystanie teorii odpowiedzi na wiązki zadań *(Testlet Response Theory)* w analizie wyników testów egzaminacyjnych

Wstęp

W testach egzaminacyjnych powszechnie wykorzystuje się wiązki zadań *(testlets)*. Głównym elementem wiązki zadań jest rdzeń (np. fragment tekstu, wykres, mapa, rysunek lub inny element graficzny). Do treści zawartych w rdzeniu zadawane są pytania (zadania testowe). Odpowiedzi na nie są ze sobą związane, tzn. wybór jednej z odpowiedzi może sugerować poprawną odpowiedź na pozostałe pytania w ramach wiązki. Z tego powodu zadania zamknięte w wiązkę mogą generować wariancję wyników testu inną niż związaną z różnicami w poziomie badanej umiejętności *(construct irrelevant variance,* CIV), ponieważ złamane może być założenie o lokalnej niezależności w ramach wiązki zadań.

Założenie o lokalnej niezależności zadań jest podstawowym założeniem teorii odpowiedzi na pozycje testowe *(item response theory,* IRT). Założenie to jest spełnione, gdy żadne z zadań w teście (treść zadania lub zestaw dostępnych odpowiedzi) nie stanowi wskazówki do prawidłowego rozwiązania innego zadania. Innymi słowy, gdy przy danym poziomie umiejętności ucznia odpowiedź na jedno zadanie jest statystycznie niezależna od odpowiedzi na inne zadanie. Złamanie założenia o lokalnej niezależności może prowadzić do błędnych oszacowań parametrów zadań i umiejętności uczniów (Li, Bolt i Fu, 2006).

Wiązkowanie zadań w testach stanowi problem w przypadku szacowania parametrów zadań i poziomu umiejętności uczniów za pomocą modeli IRT. Wykorzystanie modeli IRT nie jest standardem postępowania w przypadku opracowywania wyników testów w ramach polskiego systemu egzaminów zewnętrznych. Są jednak na dużą skalę wykorzystywane w ramach projektów badawczych, realizowanych w Instytucie Badań Edukacyjnych (IBE). Jako najważniejsze polskie przykłady analiz wyników egzaminacyjnych z wykorzystaniem modeli IRT należy wskazać opracowywanie wskaźników edukacyjnej wartości dodanej (EWD, ewd.edu.pl) oraz porównywalnych wyników egzaminacyjnych (PWE, pwe.ibe.edu.pl).

Jednym ze sposobów radzenia sobie ze złamaniem założenia o lokalnej niezależności w ramach wiązki zadań jest stosowanie modeli statystycznych uwzględniających powiązanie zadań testowych w wiązki. Modele tego typu zostały opracowane w ramach teorii odpowiedzi na wiązki zadań *(testlet response theory,* TRT, Wainer, Bradlow i Wang, 2007).

W polskiej literaturze naukowej brak jest opracowań dotyczących modeli TRT oraz sposobów ich wykorzystania w szacowaniu parametrów zadań i umiejętności uczniów. Prezentowany materiał jest pierwszą próbą starającą się wypełnić tę lukę. Zaprezentowane zostaną podstawowe informacje dotyczące modeli TRT. Dodatkowo porównane zostaną oszacowania parametrów zadań i umiejętności uczniów, wyliczone na podstawie modeli TRT i „tradycyjnych" modeli IRT.

Zakres i korzyści stosowania wiązek zadań

Twórcy standaryzowanych testów chętnie stosują wiązki zadań, ponieważ stanowią one efektywny sposób na przygotowywanie nowych zadań oraz są łatwe w procesie przeprowadzania testu. Tabela 1 prezentuje liczbę wiązek zadań w arkuszach z egzaminu gimnazjalnego z lat 2004-2015.

Wainer, Bradlow i Du (2000) podkreślają także, że za tworzeniem wiązek zadań przemawia redukcja atomistycznej natury pojedynczego zadania testowego. Pojedynczym zadaniom wielokrotnego wyboru *(multiple choice questions,* MCQ) zarzuca się często, że są oderwane od jakiegokolwiek kontekstu. Wiązki dostarczają kontekstowych informacji, dających głębszy i bardziej kompleksowy wgląd w naturę problemu rozwiązywanego przez ucznia. Zakłada się, że wiązki dobrze sprawdzają się w pomiarze procesów umysłowych, wiedzy i umiejętności wyższego rzędu i tych bardziej złożonych. Dzięki temu wiązki zadań w większym stopniu mogą reprezentować badaną umiejętność ucznia, co wpływa na podnoszenie trafności zarówno wiązki, jak i całego testu.

Tabela 1. Wiązki zadań w arkuszach egzaminu gimnazjalnego w latach 2004-2015

Część egzaminu	Rok	L. wiązek	L. pytań	Część egzaminu	Rok	L. wiązek	L. pytań
humanistyczna	2004	4	32	matematyka	2012	2	23
matematyczno-przyrodnicza	2004	2	35	j. polski	2012	4	23
humanistyczna	2005	4	34	przedmioty przyrodnicze	2012	7	26
matematyczno-przyrodnicza	2005	7	36	historia i WOS	2012	8	34
humanistyczna	2006	5	27	matematyka	2013	2	23
matematyczno-przyrodnicza	2006	6	35	j. polski	2013	4	23
humanistyczna	2007	4	31	przedmioty przyrodnicze	2013	4	28
matematyczno-przyrodnicza	2007	6	35	historia i WOS	2013	7	33
humanistyczna	2008	5	33	matematyka	2014	2	23
matematyczno-przyrodnicza	2008	7	34	j. polski	2014	4	23
humanistyczna	2009	5	30	przedmioty przyrodnicze	2014	4	28
matematyczno-przyrodnicza	2009	9	37	historia i WOS	2014	5	34
humanistyczna	2010	6	30	matematyka	2015	1	23
matematyczno-przyrodnicza	2010	9	37	j. polski	2015	3	22
humanistyczna	2011	4	30	przedmioty przyrodnicze	2015	4	28
matematyczno-przyrodnicza	2011	8	37	historia i WOS	2015	2	32

Wiązki zadań mają także zalety dla uczniów. Najważniejsza to krótszy czas potrzebny na zapoznanie się z treścią zadań. Gdy istnieje wspólna treść do kilku zadań (wspólny rdzeń), uczeń zyskuje więcej czasu na zapoznanie się z nimi niż w sytuacji, gdy do każdego zadania podana jest osobna treść (Wainer, Bradlow i Wang, 2007).

Negatywne skutki ignorowania wiązek zadań

Ignorowanie wiązek zadań podczas szacowania parametrów zadań i umiejętności uczniów z użyciem modeli IRT może prowadzić do błędnych oszacowań rzetelności i błędów standardowych poziomu umiejętności uczniów (Bradlow, Wainer i Wang, 1999; Marais i Andrich, 2008; Sireci, Thissen i Wainer, 1991; Wainer i Wang, 2000; Yen, 1993). Ignorowanie wiązek zadań prowadzi także do niepoprawnego zrównywania wyników testów (Lee, Kolen, Frisbie i Ankenmann, 2002; Li, Bolt i Fu, 2005), błędnej interpretacji wielkości parametrów dyskryminacji zadań (Bradlow, Wainer i Wang, 1999; Wainer i Wang, 2000) oraz błędnej diagnozy zadań jako niedopasowanych do struktury danych (Marais i Andrich, 2008). Nawet jeśli oszacowania parametrów zadań, jak i poziomu umiejętności uczniów z modelu ignorującego oraz modelu uwzględniającego wiązkowanie zadań są zbliżone lub nawet identyczne, to funkcja informacyjna z modelu ignorującego wiązki zadań będzie niewłaściwa (Ip, 2010; Wainer i Wang, 2000).

Negatywy efekt wiązkowania zadań dla szacowania parametrów zadań i poziomu umiejętności uczniów może zostać dodatkowo wzmocniony w sytuacji, gdy prawidłowa odpowiedź do zadań w wiązce jest oznaczona zawsze tym samym symbolem, np. A, A, A (Koniewski, Majkut i Skórska, 2014). Na przykładzie analizy wyników części egzaminu z historii i wiedzy o społeczeństwie z 2013 r. wykazano, że w wiązce trzech zadań, w której poprawne odpowiedzi oznaczone są takim samym symbolem, poprawna odpowiedź na dwa z nich obniża szansę na poprawną odpowiedź na trzecie zadanie w wiązce o 27-52%, przy kontroli poziomu umiejętności ucznia. Szansa udzielenia poprawnej odpowiedzi na trzecie zadanie w wiązce jest niższa w przypadku zadań o większej trudności. Dodatkowo, gdy cała wiązka zadań ma cechy, które wpływają na zróżnicowane funkcjonowanie zadań *(differential item functioning,* DIF), między grupami zdających o określonych cechach wykorzystywanie wiązek może prowadzić do uzyskiwania stronniczych wyników, silniej niż w przypadku pojedynczych pytań obciążonych efektem DIF. Dlatego nowo tworzone wiązki zadań powinny podlegać wnikliwej ocenie, zarówno treściowej, jak i statystycznej, na podstawie wyników pilotażu.

Metody radzenia sobie z problemem łamania założenia o lokalnej niezależności w ramach wiązek zadań

Negatywne konsekwencje złamania założenia o lokalnej niezależności w przypadku analiz IRT stały się impulsem do poszukiwania metod, które pozwalałaby na szacowanie parametrów zadań i uczniów z uwzględnieniem wiązkowania zadań. Można wyróżnić dwa typy metod analizy wiązek zadań: zorientowane na wynik i zorientowane na zadania (Wilson i Adams 1995).

W przypadku pierwszego typu metodą postępowania jest sumowanie punktów za zadania w ramach wiązki i tworzenie z nich jednego zadania, tzw. „superzadania" (Bishop i Omar, 2002; Marais i Andrich, 2008; Sireci, Thissen, Wainer, 1991; Zenisky i in., 2002). Do analizy takich zadań kodowanych politomicznie stosuje się jednowymiarowe modele IRT do zadań wielopunktowych (najczęściej *generalized partial credit model,* GPCM; Muraki, 1992). W przypadku krótkich testów rozwiązanie to ma tę wadę, że pomniejsza pulę zadań, na podstawie których szacowany jest poziom konstruktu głównego, przez co obniża precyzję jego pomiaru. Nie jest bowiem brany pod uwagę wzór odpowiedzi ucznia na zadania składające się na wiązkę (Yen, 1993). Stosowanie metody sumowania zadań jest krytykowane ze względu na negatywne konsekwencje stosowania tego podejścia (m.in. Eckes, 2014). Należy jednak wskazać, że niewątpliwą korzyścią jest tu prostota modelu i ekonomia obliczeń. Metoda ta stosowana jest w analizach prowadzonych w ramach przywołanych we wstępie projektów badawczych EWD i PWE.

Metody analizy wyników testów z wiązkowaniem zorientowane na zadania zachowują informację o wzorze odpowiedzi ucznia na zadania składające się na wiązkę. Ich główną zasadą jest wprowadzanie dodatkowych wymiarów (czynników specyficznych wiązki, *testlet factors)*, które pozwalają modelować wariancję specyficzną dla wiązki zadań równocześnie z wariancją właściwą dla czynnika głównego (Bradlow, Wainer i Wang, 1999; Wainer, Bradlow i Du, 2000). Nawet jeśli oszacowania parametrów zadań i poziomu umiejętności uczniów z modelu pomijającego wiązkowanie zadań byłyby takie same jak te z modelu uwzględniającego czynniki wiązek, to wyniki z jednowymiarowego modelu IRT obciążone są błędem. Funkcja informacyjna zadań powinna być bowiem uśredniona przez rozkłady czynników wiązek (Ip, 2010; Wainer i Wang, 2000). Dlatego wskazane jest korzystanie z modeli TRT w porównaniu z jednowymiarowymi modelami IRT, w przypadku zdiagnozowania silnych efektów wiązek.

Modele *Testlet Response Theory*

Popularnymi modelami przeznaczonymi do analizy wyników testów zawierających wiązki zadań są modele opracowane w ramach konfirmacyjnej analizy czynnikowej *(confirmatory factor analysis,* CFA) lub wielowymiarowej teorii odpowiedzi na zadania testowe *(multidimensional item response theory,* MIRT):

1. Model podwójnego czynnika *(bifactor model;* Gibbons i Hedeker, 1992), w którym każde zadanie testowe jest wskaźnikiem konstruktu głównego oraz dodatkowo ładuje jeden ze specyficznych wymiarów reprezentujących wiązkę zadań, do której dane zadanie należy. Konstrukt główny to zmienna latentna, będąca przedmiotem głównego zainteresowania (np. kompetencje matematyczne, językowe). Z kolei wymiary specyficzne uwzględniają dodatkowe zależności między zadaniami testowymi, wynikające z faktu ich powiązania w wiązki. Dla zadań kodowanych dychotomicznie3-parametryczny model podwójnego czynnika to

$$P_i(y = 1|\theta_g,\theta_1,\ldots,\theta_k,\ldots,\theta_K) = c_i + (1-c_i)\ \Phi\,(a_{ig}\,\theta_g + a_{ik}\,\theta_k - b_i), \qquad (1)$$

gdzie P_i to prawdopodobieństwo prawidłowej odpowiedzi na zadanie testowe warunkowo ze względu na wartości parametrów zadania, wartości konstruktu

głównego θ_g, oraz wektora wartości czynników specyficznych θ_K, których jest tyle, ile wiązek zadań w teście. Parametr b_i jest stałą regresji (parametrem trudności) dla zadania *i*; parametr a_{ig} to ładunek (krzywa nachylenia regresji) dla zadania testowego na konstrukcie głównym oraz a_{ik} to ładunki dla zadania na wymiarach specyficznych; c_i to dolna asymptota. Aby model był identyfikowalny, średnia oraz wariancja każdego wymiaru (czynnika, θ) ustalane są odpowiednio na 0 oraz 1. Wszystkie wymiary w modelu traktowane są jako ortogonalne (niezależne). Możliwe jest zastosowanie tego modelu także do danych politomicznych przez zastosowanie adekwatnej funkcji łączącej zamiast funkcji logitowej lub probitowej (Fahrmeir i Tutz, 2001). Przez ustalenie parametru dolnej asymptoty na zero uzyskujemy model dwuparametryczny, natomiast przez ustalenie parametrów dyskryminacji zadań testowych na jeden lub inną stałą uzyskujemy model jednoparametryczny.

2. Model wiązki zadań *(testlet model;* Bradlow, Wainer i Wang, 1999) to model podwójnego czynnika mieszanych efektów, w którym ładunki (parametry dyskryminacji) zadań na wymiary specyficzne są zdefiniowane jako proporcjonalne do ładunków tych zadań na konstrukt główny w ramach każdej wiązki zadań (Li, Bolt i Fu, 2006)

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1+e^{-Da_i(\theta - b_i + \gamma_{k(i)})}}, \qquad (2)$$

gdzie $\gamma_{k(i)}$ to wektor wartości czynników specyficznych (zamiast θ_K, notacja za: DeMars, 2012, s. 106) i tak jak we wzorze na model podwójnego czynnika oznacza wariancję wspólną dla zadań w ramach wiązki, która nie jest wyjaśniona przez konstrukt główny, jest stałą 1,7, pozwalającą przejść z metryki normalnej na logistyczną (Birbaum, 1962). Podobnie jak w poprzednim wzorze, przez ustalenie parametru dolnej asymptoty na zero uzyskujemy model dwuparametryczny; przez ustalenie parametrów dyskryminacji zadań testowych na jeden lub inną stałą uzyskujemy model jednoparametryczny.

3. W modelu z czynnikiem wyższego rzędu *(higher order factor,* HOF) zadania testowe ładują tylko wymiary specyficzne. Korelacje między wymiarami specyficznymi są modelowane przez czynnik wyższego rzędu.

Model wiązki zadań jest formalnie tożsamy z modelem podwójnego czynnika z ograniczeniami, jak również z modelem z czynnikiem wyższego rzędu (Li, Bolt i Fu, 2006; Rijmen, 2010). W tradycji konfirmacyjnej analizy czynnikowej te relacje zostały także opisane jako modele dla zmiennych ciągłych (Yung, Thissen i McLeod, 1999). Autorzy ci wykazali, że model z czynnikiem wyższego rzędu jest formalnie ekwiwalentny z tzw. modelem hierarchicznym Schmida-Leimana (Schmid i Leiman, 1957). Jest to model podwójnego czynnika z ograniczeniami na ładunki czynnikowe zadań testowych w ramach wiązki, takimi, że są one proporcjonalne do ładunków tych zadań na konstrukt główny. Stąd też model wiązki zadań jest modelem z czynnikiem wyższego rzędu po transformacji Schmida-Leimana.

Wyniki i dyskusja

W celu zilustrowania zastosowania modelowania wyników testów za pomocą modelów uwzględniających wiązki zadań wybrano wyniki części egzaminu gimnazjalnego z historii i wiedzy o społeczeństwie z 2014 r. dla losowej próbki 1000 uczniów, pobranej z ogólnopolskiej populacji 362 752 zdających. Wybór testu z historii i wiedzy o społeczeństwie był podyktowany faktem, że od 2012 r. w polskich egzaminach gimnazjalnych w tej części egzaminu występowało najwięcej wiązek zadań. Test składał się z 33 zadań, z czego 14 tworzyło pięć wiązek: k1: 'z6.1', 'z6.2'; k2: 'z8.1', 'z8.2', 'z8.3'; k3: 'z10.1', 'z10.2', 'z10.3; k4: 'z15.1', 'z15.2', 'z15.3'; k5: 'z22.1', 'z22.2', 'z22.3'. Wszystkie zadania kodowane były dychotomicznie.

Modelowanie efektów wiązek zadań, które są nieistotne, sprawia, że model jest niepotrzebnie skomplikowany i zwiększa błędy oszacowań parametrów zadań. Efekt wiązki zadań jest istotny, gdy złamane jest założenie o lokalnej niezależności w parach lub grupach zadań ujętych w wiązki. Dlatego analizę należy rozpocząć od identyfikacji zadań, dla których złamane jest założenie o lokalnej niezależności. Chen i Thissen (1997) zaproponowali statystykę , która obliczana jest przez porównanie obserwowanych i oczekiwanych częstości w ramach każdej pary zadań testowych. Statystyki te w przybliżeniu odpowiadają standaryzowanym wartościom i są wysokie, jeśli w parze zadań złamane zostało założenie o lokalnej niezależności. Ponieważ rozkład statystyki jest zbliżony, ale nie identyczny z rozkładem standaryzowanym Z, wartości statystyki powyżej 2 lub 3 nie należy uważać za wysokie. Za wskaźnik złamania założenia o lokalnej niezależności uznaje się wartości większe niż 10, a wartości między 5 a 10 jako wskazujące na taką możliwość (Cai, Thissen i du Toit, 2011). Statystyki obliczono w programie IRTPRO 2.1 (Cai, Thissen i du Toit, 2011). Statystyki powyżej 10 odnotowano tylko w grupach zadań 'z6.1', 'z6.2'; 'z8.1', 'z8.2', 'z8.3'; oraz 'z22.1', 'z22.2', 'z22.3'. Spośród pięciu wiązek w teście, zidentyfikowano trzy, w których złamane zostało założenie o lokalnej niezależności.

Wykazano, że model podwójnego czynnika, model wiązki zadań i model z czynnikiem wyższego rzędu są formalnie ekwiwalentne (Li i in., 2006; Rijmen, 2010). Dlatego jako przykład do analiz oszacowano model Rascha (jednoparametryczny) wiązki zadań (mod1; Wang i Wilson, 2005) oraz model dwuparametryczny podwójnego czynnika (mod2; Reise, 2012). Modele obliczono za pomocą pakietu TAM dla R. W mod1 szacowanych było 37 parametrów, tj. 33 parametry trudności zadań oraz wariancje czterech czynników (konstrukt główny i trzy czynniki specyficzne odpowiadające trzem wiązkom zadań). Z kolei w mod2 szacowane były 74 parametry, tj. 33 parametry trudności zadań oraz 41 parametrów dyskryminacji (dla zadań ładujących konstrukt główny i czynnik specyficzny po dwa parametry).

DeMars (2012) porównała metody detekcji efektu wiązki zadań: (1) porównanie między modelem bez wiązek a modelem z wiązką; (2) porównanie między modelem z wszystkimi potencjalnymi wiązkami a modelem z wszystkimi potencjalnymi wiązkami za wyjątkiem jednej wiązki (porównanie „all-but-one"); (3) testowanie jednowymiarowości struktury danych (Stout, 1987; 2005). Metoda 2 okazała się najskuteczniejsza.

W tabeli 2 porównano modele z trzema wiązkami wobec modeli, które ignorowały kolejno wiązkę k1, k2 lub k5. W tabeli podano liczbę parametrów szacowanych w modelach, wartości logarytmu funkcji największej wiarygodności *(log-likelihood,* LL), wartości kryterium informacyjnego Akaike'a (AIC, Akaike, 1987), wartości skorygowanego Bayesowskiego kryterium informacyjnego Schwarza *(sample adjusted BIC,* SA-BIC, Schwarz, 1978; Sclove, 1987), rzetelność empiryczną (zwaną także brzegową) dla konstruktu głównego oraz wiązek zadań k1, k2 i k5, w których złamane było założenie o lokalnej niezależności.

Statystyki AIC i SA-BIC są przekształceniami -2LL, stąd im niższe wartości, tym lepsze dopasowanie modelu do danych. SA-BIC uwzględnia „karę" za dodawanie parametrów opartą o wielkość próby. Liczne badania symulacyjne (m. in. Enders i Tofighi, 2008) sugerują użyteczność SA-BIC do porównywania modeli. Na przewagę SA-BIC nad innymi miarami w kontekście porównywania modeli uwzględniających wiązki zadań wskazała DeMars (2012).

Tabela 2. Statystyki podsumowania porównania modeli „all-but-one"

Krótka nazwa modelu	L. parametrów	LL	AIC	SA-BIC	Rzetelność EAP dim1	k1	k2	k5
mod1	37	-18777,6	37629,1	37693,0	0,758	0,283	0,500	0,395
mod1-k1	36	-18791,9	37655,9	37718,1	0,763	nd	0,503	0,406
mod1-k2	36	-18864,2	37800,4	37862,6	0,779	0,292	nd	0,420
mod1-k5	36	-18838,6	37749,2	37811,4	0,755	0,287	0,507	nd
mod2	74	-18477,8	37103,6	37231,5	0,806	0,381	0,473	0,633
mod2-k1	72	-18497,1	37138,3	37491,6	0,807	nd	0,473	0,631
mod2-k2	71	-18573,8	37289,6	37412,3	0,819	0,318	nd	0,628
mod2-k5	71	-18542,5	37227,0	37349,7	0,806	0,361	0,471	nd

W modelu Rascha wiązki zadań największy efekt należy odnotować dla wiązki k2 (zadania 'z8.1', 'z8.2', 'z8.3'). Natomiast po uwzględnieniu parametru dyskryminacji w modelu podwójnego czynnika to wiązka k5 (zadania 'z22.1', 'z22.2', 'z22.3') ma najsilniejszy efekt. Miary AIC i SA-BIC sugerują, że najlepsze modele to te, które uwzględniają wszystkie trzy wiązki zadań, zidentyfikowane jako istotne za pomocą statystyki. Stąd żadna z tych wiązek nie powinna być ignorowana.

Oszacowania parametrów zadań, jak i pozycji uczniów na konstrukcie głównym (kompetencje historyczne i w zakresie wiedzy o społeczeństwie) zestawiono z oszacowaniami uzyskanymi z modeli 1PL (mod1PL) i 2PL (mod2PL), ignorujących fakt występowania wiązek zadań, oraz z modeli, w których zadania występujące w wiązkach potraktowano jako zadania politomiczne (odpowiednio mod1PCM oraz mod2GPCM). Modele obliczono w pakiecie mirt dla R.

Jak wskazuje tabela 3, klasyczne modele IRT (zarówno jednoparametryczny model Rascha, jak i model dwuparametryczny) charakteryzują się najgorszym dopasowaniem do danych, w których zaobserwowano lokalną zależność zadań w wiązkach. Pozwalają to stwierdzić wartości AIC i SA-BIC, które dla tych modeli przyjmują najwyższe wartości. Wyniki te potwierdzają obecne w literaturze wnioski, zawierające podkreślenie, że klasyczne modele IRT nie odzwierciedlają adekwatnie struktury danych, w których występują wiązki zadań (Marais i Andrich, 2008).

Tabela 3. Statystyki podsumowania analizowanych modeli

Krótka nazwa modelu	Opis modelu	L. par.	LL	AIC	SA-BIC	Rzetelność EAP dim1	k1	k2	k5
mod1	model Rascha wiązki zadań	37	-18777,6	37629,1	37693,0	0,758	0,283	0,500	0,395
mod1PL	klasyczny model Rascha	33	-18948,2	37964,4	38023,3	0,778	nd		
mod1PCM	klasyczny *Partial Credit Model*	30	-17451,9	34971,7	35030,6	0,759			
mod2	model dwupar. podwójnego czynnika	74	-18477,8	37103,6	37231,5	0,806	0,381	0,473	0,633
mod2PL	klasyczny model dwuparametryczny	66	-18655,3	37442,7	37557,0	0,820	nd		
mod2GPCM	klasyczny *Graded Partial Credit Model*	60	-17183,8	34489,5	34595,1	0,759			

Do analizowanych danych lepiej dopasowane są modele, które z problemem wiązek radzą sobie poprzez sumowanie punktów w wiązkach i tworzenie „superzadań". Wskazują na to niższe wartości kryteriów AIC i SA-BIC dla klasycznego modelu PCM w porównaniu z modelem Rascha wiązki zadań oraz dla klasycznego modelu GPCM w porównaniu z modelem dwuparametrycznego podwójnego czynnika.

W przypadku analizowanych danych modele traktujące wiązkę jako jedno zadanie wielopunktowe są najlepszym wyborem. W literaturze naukowej i badawczej podkreśla się, że wykorzystanie tych modeli może być problematyczne i prowadzić do obniżenia precyzji szacowania głównego konstruktu wskutek zmniejszenia puli zadań składających się na test (Yen, 1993). Wskazany problem dotyczy głównie krótkich testów, z dużą ilością wiązek i/lub wiązek, na które składa się duża liczba zadań. Analizowany test jest stosunkowo długi, a wiązki są niewielkie (maksymalnie trzy zadania w wiązce), więc stosowanie modeli przeznaczonych do zadań wielopunktowych jest wystarczające. Każdorazowo decyzja o wyborze między modelem z „superzadaniami" a modelami wiązki zadań powinna bazować na analizie dopasowania do konkretnych danych.

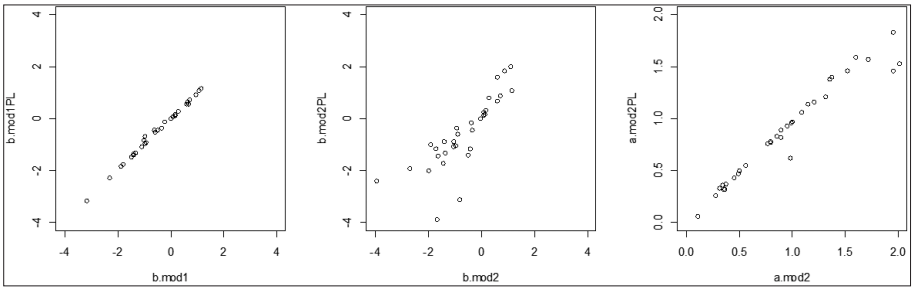
Tabela 3 prezentuje także wartości pozwalające ocenić rzetelność głównego konstruktu. W wypadku klasycznych modeli IRT wskaźniki rzetelności brzegowej wynoszą odpowiednio 0,778 dla klasycznego modelu Rascha oraz 0,820 dla klasycznego modelu dwuparametrycznego. Są one wyższe niż wskaźniki rzetelności brzegowej oszacowanej dla modeli z „superzadaniami" i modeli wiązki zadań. Wyniki te potwierdzają wnioski z innych badań, tj. że problemem klasycznych modeli IRT jest przeszacowywanie rzetelności konstruktu głównego, sprawdzanego w teście (Sireci, Thissen i Wainer, 1991; Wainer i Wang, 2000).

Właściwym indeksem rzetelności dla wyników testów, w których zastosowano wiązki zadań jest hierarchiczny współczynnik omega *(coefficient omega hierarchical,* Zinbarg i in., 1997; 2005), który wyniósł 0,841 dla modelu Rascha wiązki zadań oraz 0,870 dla modelu podwójnego czynnika. Hierarchiczny współczynnik omega definiuje się jako procent nieważonej wariancji surowych ocen czynnikowych, którą można przypisać czynnikowi głównemu.

Dostępne dowody empiryczne wskazują, że stosowanie klasycznych modeli IRT ignorujących obecność wiązek zadań w strukturze danych może powodować problemy z poprawnością oszacowań parametrów zadań (Bradlow,

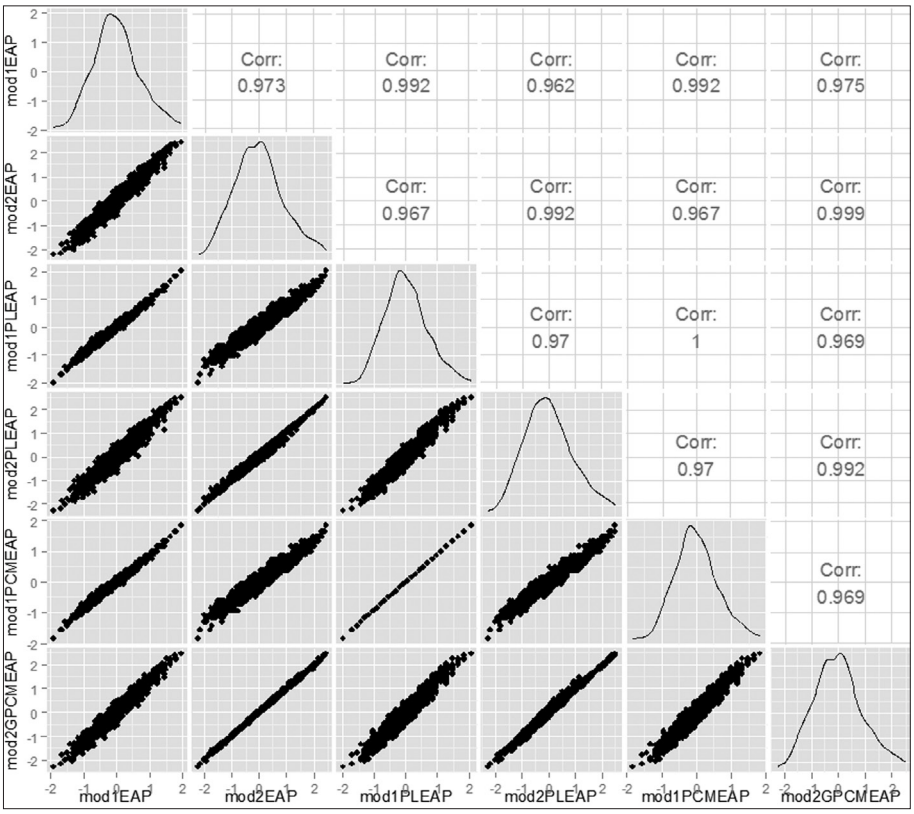
Wainer i Wang, 1999; Wainer i Wang, 2000). Jak wskazuje wykres 1, problem w niewielkim stopniu dotyczy modelu jednoparametrycznego Rascha, gdyż parametry trudności oszacowane w klasycznym modelu IRT i modelu wiązki zadań są wysoko ze sobą skorelowane. Wybór modelu w tym przypadku nie będzie powodował znacznych konsekwencji dla oszacowania parametrów zadań w teście. Może mieć to jednak poważne konsekwencje w przypadku modelu dwuparametrycznego, zwłaszcza dla oszacowania parametru trudności. Jak wskazuje wykres 2, korelacja parametrów trudności z klasycznego modelu IRT oraz modelu podwójnego czynnika jest słabsza. Na wykresie nie zaprezentowano jednego zadania, które stanowi przypadek odstający. Jest to trzecie i ostatnie zadanie z22.3, składające się na wiązkę. Z zadaniem można się zapoznać, pobierając arkusz egzaminacyjny ze strony cke.edu.pl.

Analiza tego zadania ujawniła, że klasyczny model IRT szacuje trudność tego zadania jako wynoszącą 8,912, podczas gdy model wiązki wskazuje na trudność wynoszącą 4,467. Użycie klasycznego modelu IRT prowadziłoby do wniosku, że zadanie jest dwukrotnie trudniejsze niż pozwala stwierdzić model wiązki zadania. Klasyczny model IRT, ignorując strukturę danych, pomija fakt, że uczniowie, rozwiązując zadanie, mogli korzystać ze strategii eliminacji – ich zadaniem było dopasowanie organów władzy (cztery możliwości) do trzech rodzajów uprawnień ustawodawczych. Odpowiadając na trzecie i ostatnie w wiązce zadanie, prawdopodobnie uczniowie wybierali odpowiedź nie z czterech możliwości dostępnych w zadaniu, ale z dwóch, które pozostały po rozwiązaniu dwóch poprzednich zadań w wiązce. W związku z tym zadanie było prostsze, gdyż nawet uczniowie posiadający niski poziom wiedzy z zakresu historii i WOS mieli wysokie prawdopodobieństwo zgadnięcia poprawnej odpowiedzi.



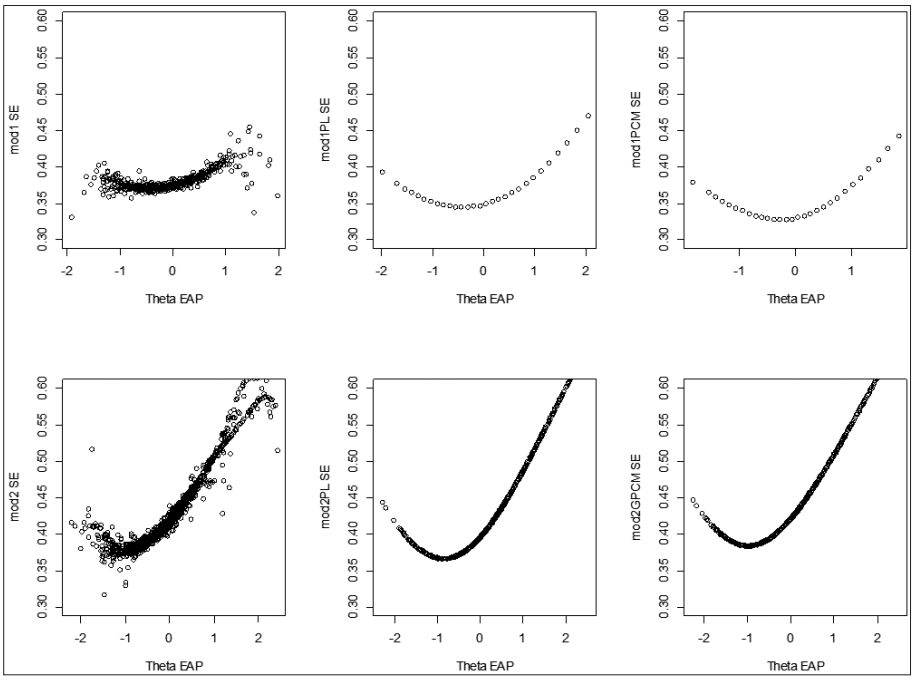
Wykresy 1-3. Porównanie parametrów trudności w modelu Rascha wiązki zadań vs. modelu 1PL (wykres 1); w modelu podwójnego czynnika vs. modelu 2PL (wykres 2) oraz porównanie parametrów dyskryminacji w modelu podwójnego czynnika vs. modelu 2PL (wykres 3). Na wykresie 2 nie umieszczono parametru b dla zadania z22.3, który w modelu 2PL wyniósł 8,912, a w modelu podwójnego czynnika 4,467

Jak pokazuje wykres 4, wybór pomiędzy modelami klasycznymi IRT, modelami wiązki zadań i modelami z „superzadaniami" ma relatywnie niewielkie konsekwencje dla oszacowania poziomu umiejętności uczniów. Rozkład umiejętności uczniów jest podobny, niezależnie od wybranego modelu. Korelacje pomiędzy oszacowaniami parametrów uczniów pochodzących z różnych modeli są bardzo wysokie i wynoszą od 0,967 do 1.



Wykres 4. Rozkłady umiejętności uczniów i ich korelacje w zastosowanych modelach

Dla samych oszacowań parametrów uczniów wybór modelu w sytuacji występowania wiązek zadań w strukturze testu nie ma relatywnie dużego znaczenia, wpływa natomiast w widoczny sposób na błędy standardowe poziomu umiejętności uczniów. Jak pokazuje wykres 5, wybór modelu wiązki zadań wpływa na obniżenie błędów standardowych oszacowań poziomu umiejętności uczniów znajdujących się na krańcach rozkładu umiejętności, tj. wśród uczniów najzdolniejszych i najmniej zdolnych. Wybór modelu wiązki zadań będzie miał więc znaczenie w kontekście zwiększenia precyzji oszacowań umiejętności uczniów o najwyższych i najniższych umiejętnościach.



Wykres 5. Oszacowania błędów standardowych i parametrów uczniów w analizowanych modelach

Podsumowanie

Otrzymane rezultaty wskazują, że modele odpowiedzi na wiązki zadań są użyteczną metodą szacowania parametrów zadań i uczniów w testach egzaminacyjnych, wykorzystywanych w polskim systemie oświaty. Jednak zasadność ich zastosowania jest uzależniona od specyfiki danego testu. Użyteczność modeli TRT rośnie wraz ze zwiększaniem się liczby zadań w wiązkach i liczby samych wiązek w teście. Modele wiązki zadań wydają się także dokładniej szacować umiejętności uczniów o najniższych i najwyższych umiejętnościach niż klasyczne modele IRT. Z tego względu należy rozważyć szersze korzystanie z tego rodzaju modeli w przypadku analiz wykorzystujących wyniki polskich egzaminów zewnętrznych.

Bibliografia

1. Akaike, H. (1987). *Factor analysis and AIC*. Psychometrika, 52, 317-332.
2. Birnbaum, A. (1962). *On the Foundations of Statistical Inference.* Journal of the American Statistical Association, 57(298), 269-306.
3. Bradlow, E. T., Wainer, H. i Wang, X. (1999). *A Bayesian random effects model for testlets.* Psychometrika, 64, 153-168.

4. Cai, L., Thissen, D. i du Toit, S. H. C. (2011). *IRTPRO for Windows.* [Computer software]. Lincolnwood, IL: Scientific Software International.
5. Chen, W.-H., Thissen, D. (1997). *Local dependence indexes for item pairs using item response theory.* Journal of Educational and Behavioral Statistics, 22, 265-289.
6. DeMars, Ch.E. (2012). *Confirming Testlet Effects. A*pplied Psychological Measurement, 36(2), 104-121.
7. Eckes, T. (2014). *Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach.* Language Testing, 31(1), 39-61.
8. Enders, C.K., i Tofighi, D. (2008). *The impact of misspecifying class-specific residual variances in growth mixture models.* Structural Equation Modeling: A Multidisciplinary Journal, 15(1), 75-95.
9. Fahrmeir, L. i Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer New York.
10. Gibbons, R., D., Hedeker, D., R., (1992). *Full-information item bi-factor analysis.* Psychometrika, 57, 423-436.
11. Ip, E. H. (2010). *Interpretation of the three-parameter testlet response model and information function.* Applied Psychological Measurement, 34, 467-482.
12. Koniewski, M., Majkut, P. i Skórska, P. (2014). *Zróżnicowane funkcjonowanie zadań testowych ze względu na wersję testu.* Edukacja, 1(126), 79-94.
13. Lee, G., Kolen, M. J., Frisbie, D. A. i Ankenmann, R. D. (2002). *Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets.* Applied Psychological Measurement, 25, 357-372.
14. Li, Y., Bolt, D. M. i Fu, J. (2006). *A comparison of alternative models for testlets.* Applied Psychological Measurement, 30, 3-21.
15. Li, Y., Bolt, D. M. i Fu, J. (2005). *A test characteristic curve linking method for the testlet model.* Applied Psychological Measurement, 29, 340-356.
16. Linn, R. L., Levine, M. V., Hastings, C. N. i Wardrop, J. L. (1981). I*tem bias in a test of Reading comprehension*. Applied Psychological Measurement, 5, 159-173.
17. Marais, I. i Andrich, D. (2008). *Formalizing dimension and response violations of local independence in the unidimensional Rasch model.* Journal of Applied Measurement, 9, 200-215.
18. Muraki, E. (1992). *A generalized partial credit model: Application of an EM algorithm.* Applied Psychological Measurement, 16, 159-176.
19. Reise, S. P. (2012). *The rediscovery of bifactor measurement models.* Multivariate Behavioral Research, 47, 667-696.
20. Rijmen, F. (2010). *Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model.* Journal of Educational Measurement, 47, 361-372.
21. Schmid, J. i Leiman., J. N. (1957). *The development of hierarchical factor solutions.* Psychometrika, 22, 53-61.
22. Schwarz, G. (1978). *Estimating the dimension of a model.* Annals of Statistics, 6, 461-464.
23. Sclove, S. L. (1987). *Application of model-selection criteria to some problems in multivariate analysis*. Psychometrika, 52, 333-343.
24. Sireci, S. G., Thissen, D., Wainer, H. (1991). *On the reliability of testlet-based tests.* Journal of Educational Measurement, 28, 237-247.
25. Stout, W. (1987). *A nonparametric approach for assessing latent trait unidimensionality.* Psychometrika, 52, 589-617.
26. Stout, W. (2005). Dimtest (Version 2.0) [Computer software]. Champaign, IL: William Stout Institute for Measurement.
27. Wainer, H. i Kiely, G. L. (1987). *Item clusters and computerized adaptive testing: A case for testlets*. Journal of Educational Measurement, 24, 185-201.

28. Wainer, H. i Lewis, C. (1990). *Toward a psychometrics for testlets*. Journal of Educational Measurement, 27, 1-14.
29. Wainer, H. i Wang, X. (2000). *Using a new statistical model for testlets to score TOEFL*. Journal of Educational Measurement, 37, 203-220.
30. Wainer, H., Bradlow, E. i Wang, X., (2007). *Testlet Response Theory and Its Applications*. Cambridge University Press.
31. Wainer, H., Bradlow, E. T. i Du, Z. (2000). *Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing.* W W. J. van der Linden i C. A. W. Glas (red.), *Computerized adaptive testing: Theory and practice* (245–270). Boston, MA: Kluwer-Nijhoff.
32. Wang,W.-C., Wilson, M. (2005). *The Rasch testlet model.* Applied Psychological Measurement, 29, 126-149.
33. Wilson, M., Adams, R. J. (1995). *Rasch models for item bundles*. Psychometrika, 60, 181-198.
34. Yen, W. M. (1993). *Scaling performance assessments: Strategies for managing local item dependence*. Journal of Educational Measurement, 30, 187-213.
35. Yung, Y., Thissen, D. i McLeod, L. D. (1999). *On the relationship between the higher-order factor model and the hierarchical factor model.* Psychometrika, 64, 113-128.
36. Zenisky, A., Hambleton, R. K., L. i Sireci, S. G. (2002). *Identification and evaluation of local item dependencies in the Medical College Admissions Test*. Journal of Educational Measurement, 39, 291-309.
37. Zinbarg, R.E., Barlow, D.H. i Brown, T.A. (1997). *Hierarchical structure and general factor saturation of the anxiety sensitivity index: Evidence and implications.* Psychological Assessment, 9, 277-284.
38. Zinbarg, R.E., Revelle, W., Yovel, I. i Li, W. (2005). *Cronbach's α, Revelle›s β, and McDonald›s ωh: Their relations with each other and two alternative conceptualizations of reliability*. Psychometrika, 70, 123-133.