Opis skalowania testów osiągnięć TOS3

2015-07-28

Streszczenie

Na potrzeby pomiaru osiągnięć na początku klasy czwartej szkoły podstawowej, w badaniu SUEK wykorzystano zestaw trzech testów osiągnięć, TOS3, mierzących poziom osiągnięć z zakresu świadomości językowej, czytania oraz matematyki. Niniejszy dokument opisuje, w jaki sposób zostały wyliczone wyniki uczniów w trzech wyróżnionych obszarach osiągnięć.

Wprowadzenie

Testy TOS3 pozwalają na pomiar osiągnięć szkolnych uczniów po zakończeniu nauki na I etapie edukacyjnym (tj. klasach I-III) z zakresu trzech umiejętności: czytania, świadomości językowej oraz matematyki. Więcej o samej konstrukcji testu można znaleźć w artykule Aleksandry Jasińskiej-Maciążek i Michała Modzelewskiego (2014). W tym miejscu opiszemy tylko najważniejsze cechy testów, które wpływają na sposób ich skalowania.

Każdy test występował w dwóch wersjach (A i B). Część zadań jest wspólna dla obu wersji – są to tak zwane zadania kotwiczące, pozostałe są unikalne dla wersji. Zgodnie z założeniami badania uczniowie rozwiązywali wszystkie zeszyty testowe z określonej wersji. Zadania z testów świadomości językowej oraz czytania zostały zestawione w ramach jednej grupy zeszytów testowych (po cztery zeszyty na wersję). Test matematyczny znajdował się w oddzielnych dwóch zeszytach (w każdej wersji). Nazwa zadania ma nastepującą budowę: [grupa zeszytów]_[wersja i numer zeszytu]_Zad_[numer zadania]. Np. zadanie P_A3_Zad_4 oznacza zadanie czwarte z trzeciego zeszytu z wersji A, z grupy zeszytów "polonistycznych" (test czytania lub test świadomości językowej).

Ogólna filozofia skalowania jest wspólna dla wszystkich trzech testów. Do skalowania zadań dwukategorialnych wykorzystano model Rascha (Rasch, 1980), a do zadań wielokategorialnych jego uogólnioną wersję w postaci modelu *partial credit* (Masters, 1982). Parametry zadań estymowano metodą brzegowej najwyższej wiarygodności (*marginal maximum likelihood*) (Bock i Aitkin, 1981). Wyniki uczniów wyliczane były jako średnia ich rozkładu *a posteriori* (*expected a posteriori*, EAP). Analizy były wykonane w pakiecie `TAM` (Kiefer, Robitzsch, i Wu, 2015) w środowisku **R** (R Core Team, 2015).

Analizy przebiegały zgodnie z ustaloną procedurą, na którą składały się następujące kroki:

1. Pozyskanie parametrów modelu.
    a. Dopasowanie modelu z wszystkimi zadaniami.
    b. Sprawdzenie dopasowania zadań.

c. Sprawdzenie DIF na wersję.
d. Sprawdzenie DIF na płeć.
e. Policzenie modelu ostatecznego. Zapisanie parametrów zadań.

2. Wyliczenie wyników uczniów z zafiksowanymi parametrami modelu.

   a. Policzenie średnich dla oddziałów klasowych w modelu z warunkowaniem po płci.
   b. Policzenie wyników uczniów w modelu z warunkowaniem po płci i średniej klasowej.
   c. Przeniesienie wyników na skalę 100/15.

W ramach kroków z punktu 1., brak odpowiedzi (kod 9) na zadania w danym zeszycie testowym TOS3 były traktowane jako braki danych, jeżeli spełnione były dwa warunki:

- uczeń nie udzielił odpowiedzi na żadne kolejne zadanie w zeszycie testowym oraz
- uczeń nie udzielił odpowiedzi na poprzedzające zadanie.

Oznacza to, że seria kodów 9 znajdująca się pod koniec danego zeszytu była traktowana jako braki danych, poza pierwszą "dziewiątką" z tej serii. Takie traktowanie braków odpowiedzi ma na celu zmniejszenie potencjalnego wpływu braku czasu na rozwiązania uczniów i tym samym przeciwdziała sztucznemu podnoszeniu trudności zadania. Należy przy tym pamiętać, że zadania z testu czytania i testu świadomości językowej sąsiadowały ze sobą w ramach zeszytu testowego. Oznacza to, że w przypadku tych testów należy określić zadania "nieosiągnięte" zanim przystąpi się do ich oddzielnego skalowania.

Na potrzeby wyliczania wyników uczniów wszystkie braki odpowiedzi były traktowane jako odpowiedzi błędne.

Przeniesienie wyników na skalę 100/15 było dokonane za pomocą dostosowanej wersji klasycznego wzoru na standaryzację wyników:

$$S_i = 15(\frac{X_i - E(X)}{\sigma_X}) + 100$$

Jako wynik punktowy ucznia (X_i) przyjęto oszacowanie EAP. $E(X)$ to średnia rozkładu wyników w badanej próbie, a σ_X to odchylenie standardowe tego rozkładu.

Analiza jakości

Jakość dopasowania danych do modelu, była sprawdzana z wykorzystaniem statystyk *infit* i *outfit* wyliczanych dla zadań. W przypadku, gdy wykryto zadanie, dla którego którakolwiek z tych statystyk wykraczała poza zakres <0,8; 1,2>, próbowano, tam, gdzie było to możliwe, podjąć próbę poprawy dopasowania poprzez zmianę klucza kodowego.

Gdy próby te zawodziły, zadanie było usuwane z dalszych analiz, a pozostałe zadania były skalowane po raz kolejny. Proces ten był powtarzany, aż do otrzymania zestawu zadań, które wykazywały akceptowalny poziom dopasowania.

Drugim krokiem w kontroli jakości było sprawdzenie zadań pod kątem zróżnicowanego ich funkcjonowania ze względu na wersję testu (tzw. *differential item functioning*, DIF). Jak już bowiem wspomniano, testy TOS3 posiadają dwie wersje zeszytów, w których część zadań jest unikalna dla wersji, a część jest wspólna (tzw. zadania kotwiczące). W ramach wieloaspektowego modelu Rascha (*multi-facet Rasch model*, zob. Linacre, 1994) dopasowano model, w którym wyliczano interakcję parametrów zadań z wersją testu. W modelu takim uwzględniano także wyraz pozwalający oszacować różnicę w poziomie umiejętności dla grup uczniów wyróżnionych ze względu na rozwiązywaną wersję testu. Zadania, dla których bezwzględna różnica trudności pomiędzy wersjami wynosiła więcej niż 0,4 logita, traktowane były jako funkcjonujące inaczej w dwóch wersjach testu. Zadania te były następnie analizowane treściowo i jeżeli podejrzenia co do ich funkcjonowania się potwierdzały, zadania te przestawały być traktowane jako kotwiczące w ramach dalszych analiz (były traktowane jako zadania unikalne dla wersji).

Trzecim i ostatnim elementem analizy jakości testów było sprawdzenie, czy zadania funkcjonują w różny sposób w zależności od płci ucznia. Analiza przebiegała analogicznie do tej opisanej wyżej. Dopasowano model, w którym obok przeciętnego poziomu umiejętności chłopców i dziewczynek modelowano interakcję trudności zadań z płcią. Zadania, dla których różnica w trudności ze względu na płeć przekraczała 0,4 logita traktowane były jako potencjalnie obciążone ze względu na płeć. Następnie sprawdzano treść tych zadań i w zależności od oceny ekspertów albo wyróżniano dwie wersje zadania (zadanie dla chłopców i dla dziewczynek), albo pozostawiano je bez zmian. Pierwsze rozwiązanie było stosowane w sytuacji, gdy istniały mocne przesłanki za hipotezą mówiącą, że zadanie odwoływało się do umiejętności w różnym stopniu opanowanym przez chłopców i dziewczynki, np. w wyniku treningu związanego z płcią lub cech i kompetencji niezwiązanych z mierzoną umiejętnością. Zadanie było pozostawiane bez zmian, gdy obserwowana różnica mogła zostać przypisana przedmiotowym różnicom pomiędzy chłopcami i dziewczynkami w zakresie badanej przez zadanie umiejętności.

Ostateczna postać testów

W dalszej części dokumentu opisano jakie zadania i z jakimi modyfikacjami posłużyły do wyliczenia wyników uczniów. Najpierw opisano jakie zadania są kotwiczące w obu wersjach testu, a nastepnie umieszczono informację o wykorzystanych zadaniach i uwzględnionych modyfikacjach.

Zadania kotwiczące powstawały poprzez złączenie odpowiedzi uczniów na zadania z poszczególnych wersji pod jedną etykietą. Etykieta ta przyjmowała zawsze postać nazwy zadania z wersji A z przedrostkiem "K", np. KM_A1_Zad_2, dla zadania kotwiczącego z wersji A i jego odpowiednika z wersji B (w tym przykładzie zadania M_B1_Zad_3).

Czytanie

Zadania kotwiczące

Tablica 1: Zadania kotwiczące w podziale na wersje. Test czytania.

Zadanie kotwiczące	Wersja A	Wersja B
KP_A2_Zad_1	P_A2_Zad_1	P_B2_Zad_4
KP_A2_Zad_2	P_A2_Zad_2	P_B2_Zad_6
KP_A2_Zad_4	P_A2_Zad_4	P_B2_Zad_2
KP_A2_Zad_5	P_A2_Zad_5	P_B2_Zad_1
KP_A2_Zad_6	P_A2_Zad_6	P_B2_Zad_3
KP_A2_Zad_7	P_A2_Zad_7	P_B2_Zad_5
KP_A3_Zad_5	P_A3_Zad_5	P_B3_Zad_9
KP_A3_Zad_6	P_A3_Zad_6	P_B3_Zad_7
KP_A3_Zad_8	P_A3_Zad_8	P_B3_Zad_10
KP_A3_Zad_9	P_A3_Zad_9	P_B3_Zad_6
KP_A4_Zad_10	P_A4_Zad_10	P_B4_Zad_11
KP_A4_Zad_11	P_A4_Zad_11	P_B4_Zad_13
KP_A4_Zad_12	P_A4_Zad_12	P_B4_Zad_10
KP_A4_Zad_14	P_A4_Zad_14	P_B4_Zad_15
KP_A4_Zad_15	P_A4_Zad_15	P_B4_Zad_14

Modyfikacje zadań

- **P_B1_Zad_15**: dopuszczono odpowiedź "B" (kod 2) jako poprawną (obok oryginalnie uznanej za poprawną odpowiedzi "C", czyli kodu 3).

Wykorzystane zadania

Tablica 2: Zadania wykorzystane w wyliczaniu wyniku w zakresie umiejętności czytania.

P_A1_Zad_11	P_A4_Zad_4	P_B1_Zad_9	P_B4_Zad_8	KP_A3_Zad_8
P_A1_Zad_12	P_A4_Zad_5	P_B1_Zad_10	P_B4_Zad_9	KP_A3_Zad_9
P_A1_Zad_13	P_A4_Zad_6	P_B1_Zad_11	P_B4_Zad_12	KP_A4_Zad_10
P_A1_Zad_14	P_A4_Zad_7	P_B1_Zad_12	KP_A2_Zad_1	KP_A4_Zad_11
P_A1_Zad_15	P_A4_Zad_8	P_B1_Zad_13	KP_A2_Zad_2	KP_A4_Zad_12
P_A2_Zad_3	P_A4_Zad_9	P_B1_Zad_14	KP_A2_Zad_4	KP_A4_Zad_14
P_A3_Zad_4	P_A4_Zad_13	P_B1_Zad_15	KP_A2_Zad_5	KP_A4_Zad_15

P_A3_Zad_7	P_B1_Zad_5	P_B3_Zad_8	KP_A2_Zad_6
P_A4_Zad_1	P_B1_Zad_6	P_B4_Zad_5	KP_A2_Zad_7
P_A4_Zad_2	P_B1_Zad_7	P_B4_Zad_6	KP_A3_Zad_5
P_A4_Zad_3	P_B1_Zad_8	P_B4_Zad_7	KP_A3_Zad_6

Świadomość językowa

Zadania kotwiczące

Tablica 3: Zadania kotwiczące w podziale na wersje. Test świadomości językowej.

Zadanie kotwiczące	Wersja A	Wersja B
KP_A1_Zad_5	P_A1_Zad_5	P_B1_Zad_2
KP_A1_Zad_6	P_A1_Zad_6	P_B1_Zad_1
KP_A1_Zad_9	P_A1_Zad_9	P_B1_Zad_3
KP_A1_Zad_10	P_A1_Zad_10	P_B1_Zad_4
KP_A2_Zad_9	P_A2_Zad_9	P_B2_Zad_8
KP_A2_Zad_10	P_A2_Zad_10	P_B2_Zad_10
KP_A2_Zad_11	P_A2_Zad_11	P_B2_Zad_12
KP_A2_Zad_12	P_A2_Zad_12	P_B2_Zad_13
KP_A2_Zad_13	P_A2_Zad_13	P_B2_Zad_15
KP_A3_Zad_1	P_A3_Zad_1	P_B3_Zad_4
KP_A3_Zad_2_1	P_A3_Zad_2_1	P_B3_Zad_3_1
KP_A3_Zad_2_2	P_A3_Zad_2_2	P_B3_Zad_3_2
KP_A3_Zad_10	P_A3_Zad_10	P_B3_Zad_14
KP_A3_Zad_11	P_A3_Zad_11	P_B3_Zad_13
KP_A3_Zad_13	P_A3_Zad_13	P_B3_Zad_12
KP_A3_Zad_14	P_A3_Zad_14	P_B3_Zad_16

Modyfikacje zadań

W teście świadomości językowej nie dokonano żadnych modyfikacji w porównaniu do oryginalnego kodowania.

Wykorzystane zadania

Tablica 4: Zadania wykorzystane w wyliczaniu wyniku w zakresie świadomości językowej.

P_A1_Zad_1	P_A2_Zad_16	P_B3_Zad_1	KP_A1_Zad_5	KP_A3_Zad_1
P_A1_Zad_2	P_A3_Zad_3	P_B3_Zad_2	KP_A1_Zad_6	KP_A3_Zad_2_1
P_A1_Zad_3	P_A3_Zad_12	P_B3_Zad_5	KP_A1_Zad_9	KP_A3_Zad_2_2
P_A1_Zad_4	P_A3_Zad_15	P_B3_Zad_11	KP_A1_Zad_10	KP_A3_Zad_10
P_A1_Zad_7	P_B2_Zad_7	P_B3_Zad_15	KP_A2_Zad_9	KP_A3_Zad_11
P_A1_Zad_8	P_B2_Zad_9	P_B4_Zad_1	KP_A2_Zad_10	KP_A3_Zad_13
P_A2_Zad_8	P_B2_Zad_11	P_B4_Zad_2	KP_A2_Zad_11	KP_A3_Zad_14
P_A2_Zad_14	P_B2_Zad_14	P_B4_Zad_3	KP_A2_Zad_12	
P_A2_Zad_15	P_B2_Zad_16	P_B4_Zad_4	KP_A2_Zad_13	

Matematyka

Zadania kotwiczące

Tablica 5: Zadania kotwiczące w podziale na wersje. Test umiejętności matematycznych.

Zadanie kotwiczące	Wersja A	Wersja B
KM_A1_Zad_2	M_A1_Zad_2	M_B1_Zad_3
KM_A1_Zad_4	M_A1_Zad_4	M_B1_Zad_2
KM_A1_Zad_5_1	M_A1_Zad_5_1	M_B1_Zad_6_1
KM_A1_Zad_5_2	M_A1_Zad_5_2	M_B1_Zad_6_2
KM_A1_Zad_5_3	M_A1_Zad_5_3	M_B1_Zad_6_3
KM_A1_Zad_7	M_A1_Zad_7	M_B1_Zad_9
KM_A1_Zad_8	M_A1_Zad_8	M_B1_Zad_10
KM_A1_Zad_9	M_A1_Zad_9	M_B1_Zad_5
KM_A1_Zad_10	M_A1_Zad_10	M_B1_Zad_7
KM_A1_Zad_12	M_A1_Zad_12	M_B1_Zad_11
KM_A2_Zad_2	M_A2_Zad_2	M_B2_Zad_9
KM_A2_Zad_4	M_A2_Zad_4	M_B2_Zad_2
KM_A2_Zad_6	M_A2_Zad_6	M_B2_Zad_10
KM_A2_Zad_7	M_A2_Zad_7	M_B2_Zad_5
KM_A2_Zad_9	M_A2_Zad_9	M_B2_Zad_7
KM_A2_Zad_10	M_A2_Zad_10	M_B2_Zad_3

Modyfikacje zadań

M_B2_Zad_1_1: Zadanie usunięte.

Wykorzystane zadania

Tablica 6: Zadania wykorzystane w wyliczaniu wyniku w zakresie umiejętności matematycznych.

M_A1_Zad_1	M_A2_Zad_8	M_B1_Zad_13	M_B2_Zad_13	KM_A1_Zad_9
M_A1_Zad_3	M_A2_Zad_11	M_B1_Zad_14	M_B2_Zad_14	KM_A1_Zad_10
M_A1_Zad_6	M_A2_Zad_12	M_B1_Zad_15	M_B2_Zad_15	KM_A1_Zad_12
M_A1_Zad_11	M_A2_Zad_13	M_B1_Zad_16	M_B2_Zad_16	KM_A2_Zad_2
M_A1_Zad_13	M_A2_Zad_14	M_B2_Zad_1_2	KM_A1_Zad_2	KM_A2_Zad_4
M_A1_Zad_14	M_A2_Zad_15	M_B2_Zad_1_3	KM_A1_Zad_4	KM_A2_Zad_6
M_A1_Zad_15	M_A2_Zad_16	M_B2_Zad_4	KM_A1_Zad_5_1	KM_A2_Zad_7
M_A1_Zad_16	M_B1_Zad_1	M_B2_Zad_6	KM_A1_Zad_5_2	KM_A2_Zad_9
M_A2_Zad_1	M_B1_Zad_4	M_B2_Zad_8	KM_A1_Zad_5_3	KM_A2_Zad_10
M_A2_Zad_3	M_B1_Zad_8	M_B2_Zad_11	KM_A1_Zad_7	
M_A2_Zad_5	M_B1_Zad_12	M_B2_Zad_12	KM_A1_Zad_8	

Literatura cytowana

Bock, R. D., i Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi:10.1007/BF02293801

Jasińska, A., i Modzelewski, M. (2014). Testy osiągnięć szkolnych TOS3: Przykład narzędzia skonstruowanego z wykorzystaniem modelu Rascha. *Edukacja*, *2*(127), 85–107.

Kiefer, T., Robitzsch, A., i Wu, M. (2015). *TAM: Test analysis modules*. Pobrano z http://CRAN.R-project.org/package=TAM

Linacre, J. M. (1994). *Many-facet rasch measurement* (2nd ed.). Chicago: MESA Press.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Pobrano z http://www.R-project.org/

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press.