Opis skalowania testów osiągnięć w VII etapie badania SUEK

2015-06-22

Streszczenie

Na potrzeby pomiaru osiągnięć na koniec szóstej klasy szkoły podstawowej, w badaniu SUEK wykorzystano zestaw trzech testów osiągnięć, TOS6, mierzących poziom osiągnięć z zakresu świadomości językowej, czytania oraz matematyki. Jako dodatkową informację o osiągnięciach uczniów wykorzystano rekordy wykonań ze sprawdzianu z 2014 r. Niniejszy dokument opisuje, w jaki sposób zostały wyliczone wyniki uczniów w trzech wyróżnionych obszarach osiągnięć.

Wprowadzenie

Ogólna filozofia skalowania jest wspólna dla wszystkich trzech testów. Każdy test był skalowany razem z odpowiadającymi mu zadaniami ze sprawdzianu w szóstej klasie z roku 2014. 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 TOS6 (ale nie na sprawdzianie) były traktowane jako braki danych, jeżeli spełnione były dwa warunki:

- uczeń nie udzielił odpowiedzi na żadne kolejne zadanie 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.

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 przeskalowane 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). Testy TOS6 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 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, następnie, które zadania ze sprawdzianu były uwzględnione w skalowaniu, a na końcu 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. KCA_1, dla zadania kotwiczącego z wersji A i jego odpowiednika z wersji B (w tym przykładzie zadania CB_7).

Czasami analizy wykazywały konieczność wprowadzenia oddzielnego parametru trudności dla chłopców i dziewczynek. W takiej sytuacji tworzono w zbiorze danych sztuczne zadania dla każdej z płci. Do wersji zadania dla danej płci skopiowano odpowiedzi uczniów o tej płci, a uczniom płci przeciwnej przypisywano braki danych. Analogicznie postępowano w przypadku wersji zadania dla drugiej płci.

Czytanie

Zadania kotwiczące

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

Zadanie kotwiczące	Wersja A	Wersja B
KCA_1	CA_1	CB_7
KCA_2	CA_2	CB_8
KCA_3	CA_3	CB_9
KCA_4	CA_4	CB_10
KCA_5	CA_5	CB_11
KCA_6	CA_6	CB_12
KCA_7	CA_7	CB_1
KCA_8	CA_8	CB_2
KCA_9	CA_9	CB_3
KCA_10	CA_10	CB_4
KCA_11	CA_11	CB_5
KCA_12	CA_12	CB_6

Wykorzystane zadania ze sprawdzianu

Oryginalnie wykorzystano 10 pierwszych zadań ze sprawdzianu, od s_1 do s_10.

Modyfikacje zadań

- **KCA_1**: zadanie usunięte.
- **KCA_12**: zrekodowanie kategorii 2 na 1.
- **CB_13**: oddzielny parametr trudności dla chłopców i dziewcząt.
- **s_2**: oddzielny parametr trudności dla chłopców i dziewcząt.
- **s_3**: oddzielny parametr trudności dla chłopców i dziewcząt.

Wykorzystane zadania

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

CA_13	CA_19	CB_19	KCA_7	CB_13k	s_7	s_3k
CA_14	CB_14	KCA_2	KCA_8	CB_13m	s_8	s_3m
CA_15	CB_15	KCA_3	KCA_9	s_1	s_9	
CA_16	CB_16	KCA_4	KCA_10	s_4	s_10	
CA_17	CB_17	KCA_5	KCA_11	s_5	s_2k	
CA_18	CB_18	KCA_6	KCA_12	s_6	s_2m	

Świadomość językowa

Zadania kotwiczące

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

Zadanie kotwiczące	Wersja A	Wersja B
KSA_1	SA_1	SB_1
KSA_4	SA_4	SB_4
KSA_5	SA_5	SB_5
KSA_7	SA_7	SB_9
KSA_9	SA_9	SB_11
KSA_12	SA_12	SB_13
KSA_13	SA_13	SB_14
KSA_14	SA_14	SB_16
KSA_15	SA_15	SB_17
KSA_17	SA_17	SB_18
KSA_18	SA_18	SB_19
KSA_20	SA_20	SB_21
KSA_22	SA_22	SB_22
KSA_23	SA_23	SB_24

Wykorzystane zadania ze sprawdzianu

Oryginalnie wykorzystano dwa ostatnie zadania ze sprawdzianu, tj. zadania s_25 i s_26.

Modyfikacje zadań

- **SA_2**: zadanie usunięte.
- **SA_24**: zadanie usunięte.
- **KSA_18**: skrócona skala oceny (kat. 2 została połączona z 1).
- **s_26**: dla kryteriów ocenianych na skalach dłuższych niż zero-jedynkowe, tj. s_26_1 ("pisze opowiadanie na zadany temat"; maks. 3 pkt.) oraz s_26_3 ("pisze poprawnie pod względem językowym"; maks. 2 pkt.), przyznano po jednym punkcie tylko tym uczniom, którzy zdobyli maksymalną liczbą punktów w danym kryterium. W efekcie zadanie oceniane jest na skali od 0 do 5 punktów.

Wykorzystane zadania

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

SA_3	SA_19	SB_8	KSA_1	KSA_13	KSA_22
SA_6	SA_21	SB_10	KSA_4	KSA_14	KSA_23
SA_8	SB_2	SB_12	KSA_5	KSA_15	s_25r
SA_10	SB_3	SB_15	KSA_7	KSA_17	s_26r
SA_11	SB_6	SB_20	KSA_9	KSA_18	
SA_16	SB_7	SB_23	KSA_12	KSA_20	

Matematyka

Zadania kotwiczące

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

Zadanie kotwiczące	Wersja A	Wersja B
KMA_4	MA_4	MB_4
KMA_6	MA_6	MB_6
KMA_8	MA_8	MB_7
KMA_9	MA_9	MB_8
KMA_10	MA_10	MB_9
KMA_12	MA_12	MB_12
KMA_13	MA_13	MB_13
KMA_14	MA_14	MB_14
KMA_15	MA_15	MB_16
KMA_16	MA_16	MB_17
KMA_18	MA_18	MB_18
KMA_22	MA_22	MB_22
KMA_23	MA_23	MB_23
KMA_24	MA_24	MB_24

Wykorzystane zadania ze sprawdzianu

Oryginalnie wykorzystano 14 środkowych zadań ze sprawdzianu, tj. zadania od s_11 i s_24.

Modyfikacje zadań

- **MA_25**: odpowiedzi B i D uznane za poprawne.

- **MB_25**: Zadanie usunięte.
- **MA_7**: oddzielny parametr trudności dla chłopców i dziewcząt.
- **s_17-s_20**: W skalowaniu uwzględniona suma punktów za te zadania.
- **s_22**: kategorie 1 i 2 zrekodowane na 1, kategorie 3 i 4 zrekodowane na 2.

Wykorzystane zadania

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

MA_1	MA_19	MB_3	MB_20	KMA_10	KMA_18	s_11	s_21
MA_2	MA_20	MB_5	MB_21	KMA_12	KMA_22	s_12	s_23
MA_3	MA_21	MB_10	KMA_4	KMA_13	KMA_23	s_13	s_24
MA_5	MA_25	MB_11	KMA_6	KMA_14	KMA_24	s_14	s_17_20r
MA_11	MB_1	MB_15	KMA_8	KMA_15	MA_7k	s_15	s_22r
MA_17	MB_2	MB_19	KMA_9	KMA_16	MA_7m	s_16	

Bibliografia

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

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.