Edukacja 2013, 4(124), 42–66
ISSN 0239-6858

IRT i pomiar edukacyjny

Bartosz Kondratek, Artur Pokropek
Zespół Analiz Osiągnięć Uczniów, Instytut Badań Edukacyjnych*

Pod nazwą *item response theory* kryje się rodzina narzędzi statystycznych wykorzystywanych do modelowania odpowiedzi na rozwiązywane zadania oraz umiejętności uczniów. Modele IRT czynią to poprzez wprowadzenie parametryzacji, która określa: właściwości zadań oraz rozkład poziomu umiejętności uczniów. W artykule przedstawiony zostanie ogólny opis jednowymiarowego modelu IRT, przybliżone zostaną najczęściej stosowane modele dla zadań ocenianych dwupunktowo (2PLM, 3PLM, 1PLM) oraz wielopunktowo (GPCM), a także zarysowana zostanie problematyka estymacji poziomu umiejętności. Artykuł ma za zadanie wprowadzić czytelnika w techniczne szczegóły związane z modelowaniem IRT oraz przedstawić wybrane zastosowania praktyczne w pomiarze edukacyjnym. Wśród zastosowań praktycznych omówiono wykorzystanie IRT w analizie skomplikowanych schematów badawczych, zrównywaniu/łączeniu wyników testowych, adaptatywnym testowaniu oraz przy tworzeniu map zadań.

Słowa kluczowe: IRT, skalowanie, złożone schematy badawcze, zrównywanie, testowanie adaptatywne, mapowanie zadań.

Modele *item response theory* (IRT) w ciągu ostatniej dekady stały się podstawowym narzędziem statystycznym w rękach badacza zainteresowanego pomiarem edukacyjnym. W polskiej literaturze pomiarowej trudno jednak znaleźć publikacje, które przedstawiałyby statystyczne aspekty modelowania i wychodziłyby poza prosty aplikacyjny charakter opisu raportowanego badania. Czytelnik, chcący dowiedzieć się czegoś więcej o statystycznych aspektach modelowania, kierowany jest często do anglojęzycznych źródeł, które nierzadko okazują się trudne do zdobycia. Pomiar edukacyjny, opierający się na modelowaniu cech ukrytych, w Polsce jest wciąż dziedziną niszową i brak jest polskiego podręcznika kompilującego informacje o modelowaniu IRT. Artykuł, który przygotowaliśmy, do pewnego stopnia ma zapełnić tą przestrzeń. Stanowi wprowadzenie do modelowania IRT. Staramy się w nim opisać podstawy modelowania IRT, nie uciekając od statystycznych zagadnień oraz praktycznych wskazówek, kierujących czytelnika do trafnego wyboru modelu oraz do właściwych interpretacji. Artykuł jest skierowany zarówno do osób, które dopiero zaczynają poruszać się w tej dziedzinie badań, jak i do czytelników, którzy chcieliby uporządkować i poszerzyć wiedzę zdobytą wcześniej, a polskie źródła są dla nich niewystarczające.

W pierwszej części wprowadzamy uogólniony model IRT, dalej przedstawiamy

Artykuł powstał w ramach projektu systemowego „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" realizowanego przez Instytut Badań Edukacyjnych i współfinansowanego ze środków Europejskiego Funduszu Społecznego (Program Operacyjny Kapitał Ludzki 2007–2013, priorytet III: Wysoka jakość systemu oświaty).

* Adres do korespondencji: Bartosz Kondratek, Zespół Analiz Osiągnięć Uczniów, Instytut Badań Edukacyjnych, ul. Górczewska 8, 01-180 Warszawa. E-mail: b.kondratek@ibe.edu.pl

najczęściej przyjęte jego formy w praktyce badawczej. Wychodzimy od modelu najczęściej obecnie używanego (dwuparametrycznego), przedstawiając dalej modele rzadziej stosowanie (jedno- i trójparametryczny). Nie poprzestajemy na modelach dla pytań dychotomiczych – prezentujemy modele dla pytań o stopniowalnym charakterze poprawnej odpowiedzi. W dalszej części artykułu została omówiona kwestia dopasowania modelu do danych, bardzo istotna przy podejmowaniu decyzji o wyborze konkretnego modelu IRT.

Programy pozwalające dopasować modele IRT zazwyczaj oferują całą kafeterię dostępnych estymatorów poziomu umiejętności ucznia, pozostawiając wybór użytkownikowi. Aby ten wybór ułatwić, metody szacowania parametru umiejętności opisano z wyszczególnieniem zasadniczych różnic między nimi. W drugiej części artykułu przedstawiamy zastosowania IRT w pomiarze edukacyjnym, które są unikalne dla tego typu metodologii (lub przynajmniej bardzo trudne do aplikacji w ramach klasycznej teorii testów): złożone schematy doboru zadań, łączenie i zrównywanie wyników, testowanie adaptatywne i mapowanie zadań.

Statystyczna charakterystyka modeli IRT

Jednowymiarowy model IRT w ujęciu ogólnym

Celem modelu IRT jest opisanie rozkładu prawdopodobieństwa wektora odpowiedzi $\boldsymbol{U} = (U_1, U_2, ..., U_n)$ udzielanych przez ucznia, którego wylosowano z pewnej populacji $\mathcal{P}$. W najogólniejszej postaci, jednowymiarowy model IRT można przedstawić w następującej postaci:

$$P(\boldsymbol{U} = \boldsymbol{u}|\mathcal{P}) = \int f(\boldsymbol{u}, \theta, \beta)\psi_{\mathcal{P}}(\theta)\, d\theta, \quad (1)$$

gdzie θ jest losową zmienną ukrytą opisującą poziom umiejętności uczniów; $\psi_{\mathcal{P}}(\theta)$ jest funkcją gęstości prawdopodobieństwa określającą rozkład zmiennej θ w populacji $\mathcal{P}$; $f(\boldsymbol{u}, \theta, \beta)$ jest funkcją, która określa prawdopodobieństwo zaobserwowania konkretnej wartości $\boldsymbol{u}$ wektora odpowiedzi $\boldsymbol{U}$, w zależności od poziomu umiejętności θ oraz wektora parametrów $\beta_i = (\beta_1, \beta_2, ...\beta_n)$, gdzie parametry zadania β_i również mogą być wektorami (np. dla dwuparametrycznego modelu logistycznego $\beta_i = (a_i, b_i)$.

Podstawowym założeniem jednowymiarowych modeli IRT jest faktoryzowanie się funkcji określającej prawdopodobieństwo całego wektora odpowiedzi $f(\boldsymbol{u}, \theta, \beta)$ do iloczynu tzw. funkcji charakterystycznych poszczególnych zadań:

$$f(\boldsymbol{u}, \theta, \beta) = \prod_{i=1}^{n} f_i(u_i, \theta, \beta_i). \quad (2)$$

Założenie (2) nosi nazwę lokalnej niezależności i ma bardzo istotne techniczne znaczenie przy szacowaniu parametrów modelu, ale samo w sobie stanowi również bardzo ważną teoretyczną przesłankę dotyczącą testu złożonego z zadań i. Mianowicie (2) stanowi, że w momencie, gdy poziom umiejętności θ jest znany, odpowiedzi na zadania testu są względem siebie statystycznie niezależne – poziom umiejętności θ wystarcza do wyjaśnienia wszystkich obserwowanych współzależności między zadaniami. Tym samym z założenia o lokalnej niezależności (2) wynika założenie o jednowymiarowym charakterze θ. Zarówno model (1), jak i założenie (2) można uogólnić do postaci wielowymiarowego poziomu umiejętności. Kompleksowe omówienie wielowymiarowych modeli IRT można znaleźć w publikacji Marka Reckase'a (2009).

Z wzoru (1) wynika, że parametry modelu IRT to zestawy: parametrów zadań β oraz parametrów określających rozkład umiejętności $\psi_{\mathcal{P}}$. Zazwyczaj przyjmuje się $\psi_{\mathcal{P}} = N(\mu_{\mathcal{P}}, \sigma^2_{\mathcal{P}})$, czyli że rozkład umiejętności jest określony przez rozkład normalny

o średniej $\mu_{\mathcal{P}}$ oraz wariancji $\sigma^2_{\mathcal{P}}$. Oszacowanie wartości parametrów β oraz parametrów rozkładu umiejętności na podstawie zebranych danych nosi nazwę kalibracji testu.

Podstawowe modele IRT

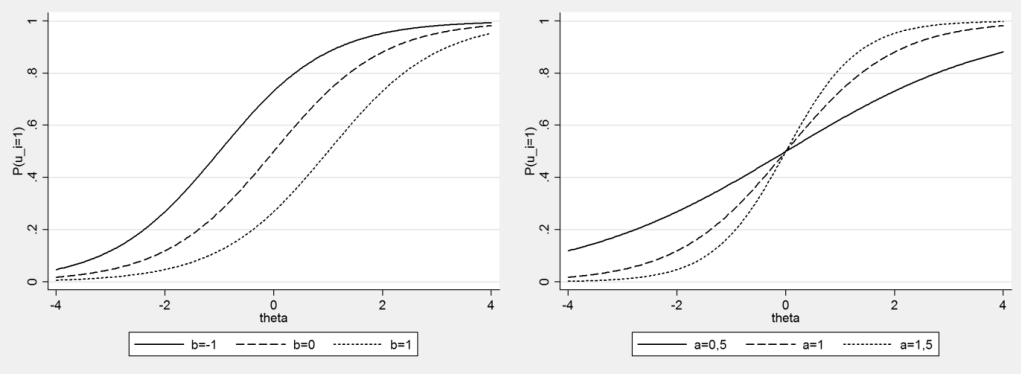
Różnica między jednowymiarowymi modelami IRT sprowadza się do postaci funkcji pojawiających się we wzorze (2), które określają prawdopodobieństwa uzyskania poszczególnych odpowiedzi, w zależności od poziomu umiejętności θ. Przedstawione w dalszej części modele IRT zostały sformułowane już w pionierskich pracach z zakresu IRT – można je znaleźć u Allana Birnbauma (1968), Georga Rascha (1960) oraz Fumiko Samejimy (1969). Wszystkie prezentowane modele będą się odwoływały do funkcji logistycznej. Dostępne są dla nich również wersje opierające się na krzywej skumulowanego rozkładu normalnego, które częstokroć są historycznie i teoretycznie (zob. Lord i Novick, 1968) pierwotne względem rozwiązań opartych na funkcji logistycznej. Ze względu na bardzo przyjazne matematyczne właściwości funkcji logistycznej modele *normal ogive* zostały w dużej mierze wyparte z praktycznych zastosowań modeli jednowymiarowych (stosowane są za to w modelach wielowymiarowych) przez dające bardzo zbliżone wyniki modele logistyczne, i nie zostaną w tym opracowaniu opisane. Relacja między modelami IRT opartymi na funkcji logistycznej a modelami opartymi na krzywej skumulowego rozkładu normalnego może być sprowadzona do relacji między logitową a probitową funkcją wiążącą w uogólnionych modelach liniowych/nieliniowych, gdyż modele IRT stanowią szczególny przypadek tychże (De Boeck i Wilson, 2004).

Modele dla zadań ocenianych dychotomicznie

W modelu 2PLM (*two-parameter logistic model*) prawdopodobieństwo udzielenia poprawnej odpowiedzi w zależności od poziomu umiejętności θ jest określone za pomocą funkcji, która zależy od parametrów a_i oraz b_i w następujący sposób:

$$P(u_i = 1|\theta, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}. \quad (3)$$

Parametr a_i nosi nazwę parametru dyskryminacji, natomiast parametr b_i nosi nazwę parametru trudności. Wykres funkcji określającej prawdopodobieństwo udzielenia odpowiedzi ocenianej na określoną liczbę punktów w zależności od poziomu



Rysunek 1. Przykładowe krzywe charakterystyczne w modelu 2PLM; z lewej zróżnicowany parametr trudności (a_i ustalony na 1), z prawej zróżnicowany parametr dyskryminacji (b_i ustalony na 0).

umiejętności ucznia nosi w IRT nazwę krzywej charakterystycznej zadania (*item characteristic curve*, ICC). Zależność pomiędzy wartościami parametrów modelu 2PLM a kształtem krzywej charakterystycznej zadania modelującej prawdopodobieństwo udzielenia poprawnej odpowiedzi zilustrowano na Rysunku 1.

Na Rysunku 1 widać, że zmiana parametru b_i przesuwa wykres równolegle do osi θ. Im b_i będzie większe, tym mniejsze będzie prawdopodobieństwo udzielenia poprawnej odpowiedzi na to zadanie dla uczniów o ustalonym poziomie umiejętności – stąd nazwa parametru. W modelu 2PLM parametr trudności wyznacza punkt umiejętności $\theta = b_i$, w którym prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie wynosi 0,5 – parametr trudności można zatem w modelu 2PLM bezpośrednio odnieść do skali wyznaczonej przez zmienną umiejętności θ. Dodatkowo, $\theta = b_i$ jest punktem przegięcia krzywej charakterystycznej w modelu 2PLM i wskazuje otoczenie θ, w którym krzywa jest najbardziej stroma.

Parametr dyskryminacji natomiast określa w 2PLM wspomnianą stromość krzywej charakterystycznej. Obserwujemy na Rysunku 1, że im większa jego wartość, tym silniejsze jest w punkcie przegięcia nachylenie krzywej (parametr a_i jest w 2PLM równy pierwszej pochodnej liczonej w punkcie $\theta = b_i$). Im bardziej wykres ICC jest w pewnym punkcie stromy, tym większą dane zadanie ma zdolność do różnicowania uczniów o poziomie umiejętności w lewym sąsiedztwie tego punktu, od uczniów o poziomie umiejętności znajdującym się w prawym sąsiedztwie tego punktu. Parametr a_i informuje zatem, jak dobrze dane zadanie różnicuje uczniów w otoczeniu $\theta = b_i$, stąd też jego nazwa.

Gdyby zredukować model 2PLM dany wzorem (3) do postaci, w której wartość dyskryminacji wszystkich zadań jest równa jedności, to powstanie model 1PLM (*one-parametr logistic model*), nazwany na cześć duńskiego matematyka Georga Rascha także jego nazwiskiem. Krzywe w modelu Rascha będą zatem względem siebie równoległe, tak jak na wykresie z lewej strony Rysunku 1. Będąca konsekwencją ustalenia parametru dyskryminacji równoległość krzywych charakterystycznych w modelu Rascha z jednej strony usztywnia model, powodując, że zazwyczaj będzie on gorzej dopasowany do danych, ale z drugiej strony, niesie ze sobą kilka wartych odnotowania zalet.

Model ten odznacza się wieloma korzystnymi właściwościami matematycznymi, w szczególności jest jedynym z prezentowanych modeli, w którym wynik sumaryczny w teście jest statystyką dostateczną dla oszacowania poziomu umiejętności ucznia (Wright i Stone, 1979). Może mieć to pozytywne implikacje praktyczne, gdyż pozwala na przykład na łatwą konwersję między sumą punktów a skonstruowaną skalą. W przypadku modeli o większej liczbie parametrów, aby określić wynik ucznia na skali θ, potrzebna jest znajomość całego wektora odpowiedzi.

Wracając do prezentowanego na Rysunku 1 przykładu z trzema zadaniami o różnej dyskryminacji, zauważamy, że dla uczniów, których odpowiedzi osiągają wartość $\theta = 0$, zadania mają taką samą trudność. Jednak dla uczniów najsłabszych zadanie o najniższej dyskryminacji jest zadaniem najłatwiejszym, zadanie o dyskryminacji 1 jest od niego trudniejsze, a zadanie o najwyższej dyskryminacji jest najtrudniejsze. Gdy popatrzymy na uczniów o poziomie umiejętności powyżej 0, porządek trudności zadań odwraca się – najłatwiejszym jest zadanie najbardziej dyskryminujące, a najtrudniejszym zadanie najmniej dyskryminujące. Opisana interakcja relatywnej (względem innych zadań) trudności zadania z poziomem umiejętności budzi pewne zastrzeżenia

zwolenników modelu Rascha. W sposób intuicyjny ilustruje, dlaczego sumaryczny wynik w teście nie jest statystyką dostateczną dla modeli dopuszczających nierównoległość ICC – odpowiedź poprawna na zadanie w takich modelach ma lokalnie różną wagę i różne znaczenie dla oceny poziomu umiejętności ucznia.

Zwolennicy modelu Rascha argumentują natomiast, że miary umiejętności ucznia konstruowane za pomocą modelu Rascha lokują wyniki na skali przedziałowej, podczas gdy dla modeli o większej liczbie parametrów nie jest to zasadne (Wright, 1983; DeMars, 2010). Relacja „pomiaru" ukrytych zmiennych umiejętności w sensie psychometrycznym, jaki umożliwiają modele IRT, do pomiaru w rozumieniu typowym dla nauk ścisłych jest bardzo ciekawym i ważnym tematem, który jednak wykracza poza ramy tego artykułu. Warto w tym punkcie zaznaczyć, że teza mówiąca o tym, że model Rascha umożliwia pomiary na skali przedziałowej w rozumieniu Stanleya Stevensa, wzbudza wiele kontrowersji i nie jest ogólnie podzielana od momentu jej sformułowania, aż do dziś. Przegląd krytycznej dyskusji nad przedziałowością skal powstałych w wyniku zastosowania modeli IRT, rozumianą w klasycznym ujęciu Stevensa, przedstawili Michael Kolen i Robert Brennan (2004). Krytyczne ujęcie tematu na gruncie aksjomatycznej teorii pomiaru można znaleźć u Andrew Kyngdona (2011).

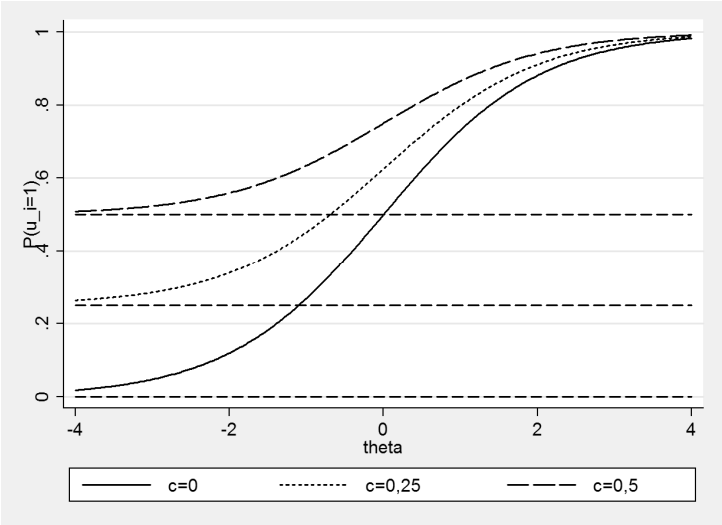
Trójparametryczny model logistyczny (*three-parameter logistic model*, 3PLM) powstaje natomiast poprzez uogólnienie 2PLM wyrażonego wzorem (3) w taki sposób, aby dolna asymptota przypadała powyżej zera. Uzyskuje się to poprzez c_i wprowadzenie dodatkowego parametru w następujący sposób:

$$P(u_i = 1|\theta, a_i, b_i, c_i) = \\ = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}}. \qquad (4)$$

Krzywą charakterystyczną w modelu 3PLM można zatem postrzegać jako średnią ważoną pomiędzy przeważonym przez c_i prawdopodobieństwem udzielenia odpowiedzi prawidłowej wynoszącym 1 na całym zakresie umiejętności θ oraz prawdopodobieństwem udzielenia odpowiedzi prawidłowej, zgodnie z modelem 2PLM przeważonym przez $(1 - c_i)$. W konsekwencji uzyskujemy krzywe, których dolna asymptota jest równa parametrowi c_i (Rysunek 2). Widać również, że 2PLM można postrzegać jako szczególny przypadek 3PLM, gdy parametr $c_i = 0$.

Krzywe z niezerowym parametrem c_i sugerują, że uczniowie o bardzo niskim poziomie mierzonej umiejętności mają dodatnie prawdopodobieństwo udzielenia odpowiedzi prawidłowej na dane zadanie. 3PLM często okazuje się przydatny do modelowania odpowiedzi na zadania wyboru, gdzie istnieje możliwość odgadnięcia odpowiedzi prawidłowej, w związku z czym parametr c_i bywa nazywany parametrem zgadywania (*guessing*). Jednak interpretacja odgadywania odpowiedzi prawidłowej nie zawsze jest w pełni uzasadniona do wyjaśnienia konkretnego poziomu c_i sugerowanego przez model. Zatem c_i ogólniej określa się jako parametr „pseudozgadywania" (*pseudo-guessing*).

Analizując krzywe na Rysunku 3 widzimy, że przy ustaleniu wartości a_i oraz b_i, wzrost wartości parametru c_i powoduje zmniejszenie zdolności zadania do różnicowania uczniów – krzywe stają się lokalnie w każdym punkcie θ mniej strome. Jednocześnie, z wprowadzeniem parametru c_i traci moc bezpośrednia interpretacja wartości parametrów a_i oraz b_i, jaka miała miejsce w modelu 2PLM. Parametr nie jest już punktem, w którym uczniowie uzyskują odpowiedź poprawną z prawdopodobieństwem 0,5 (dla $c_i > 0{,}5$, taki punkt w ogóle nie istnieje). Przełożenie wartości parametru a_i na stromość wykresu w punkcie $\theta = b_i$ również przestaje być tak bezpośrednie jak w 2PLM – aby uzyskać takie samo



Rysunek 2. Przykładowe krzywe charakterystyczne w modelu 3PLM; parametry dyskryminacji oraz trudności ustalone odpowiednio na wartościach: $a_i = 1$ oraz $b_i = 0$.

nachylenie w $\theta = b_i$ przy zwiększającym się c_i trzeba zwiększyć a_i. W związku z tym analizowanie właściwości zadania na podstawie parametrów 3PLM staje się o wiele trudniejsze niż w przypadku 2PLM – trzeba trójkę a_i, b_i, c_i rozpatrywać łącznie. O wiele łatwiej ocenić jakość zadania w 3PLM, patrząc na krzywą charakterystyczną i analizować jej lokalną stromość w zależności od wartości θ – im krzywa jest bardziej nachylona w danym rejonie umiejętności, tym lepiej uczniów w tym rejonie różnicuje (ta uwaga odnosi się oczywiście również do 2PLM).

Modele dla zadań ocenianych politomicznie

Przedstawiając modele dla zadań ocenianych dychotomicznie, dla każdego zadania wprowadzono tylko jedną krzywą charakterystyczną, która opisywała prawdopodobieństwo udzielenia odpowiedzi zakodowanej jako „1", czyli odpowiedzi poprawnej. Dla kategorii odpowiedzi ocenionej jako „0" można również wykreślić krzywą informującą o prawdopodobieństwie udzielenia tej odpowiedzi, jednak jest ona pomijana, gdyż dla zadania ocenianego zerojedynkowo jest redundantna: $P(u_i = 0) = 1 - P(u_i = 1)$. Inaczej jest w przypadku zadań ocenianych na szerszej niż zerojedynkowa skali punktowej. Do opisu zadań ocenianych wielopunktowo konieczne jest przedstawienie krzywych opisujących prawdopodobieństwo udzielenia odpowiedzi ocenianej dla każdej z możliwych m kategorii oceny.

Dla zadania ocenianego na skali $0–m$ w modelu odpowiedzi stopniowanej (*graded response model*, GRM), dokonuje się tego, szacując dla każdej z kategorii punktowej $x \epsilon \{0, \ldots, -1\}$ krzywe zgodne z modelem 2PLM (a dokładniej: z przeciwieństwem 2PLM):

$$P_x(u_i \leq x | \theta, a_i, b_{i,x}) = \frac{-1}{1 + e^{-a_i(\theta - b_{i,x})}} \quad (5)$$

Krzywe określone wzorem (5) mówią o prawdopodobieństwie udzielenia odpowiedzi punktowanej na co najwyżej x, różnią się

parametrem trudności $b_{i,x}$, ale mają wspólny parametr dyskryminacji, więc są względem siebie równolegle przesunięte (por. przykład z lewej na Rysunku 1). Następnie, dla wyznaczenia krzywej opisującej uzyskanie konkretnej wartości punktowej, oblicza się:

- dla kategorii 0 punktów:

$$P(u_i = 0|\theta) = P_0(u_i \leq 0|\theta),$$

- dla kategorii pośrednich $:x\epsilon\{1, \dots, m-1\}$:

$$P(u_i = x|\theta) = P_x(u_i \leq x|\theta) - P_{x-1}(u_i \leq x-1|\theta),$$

- dla kategorii m punktów:

$$P(u_i = m|\theta) = 1 - P_{m-1}(u_i \leq m-1|\theta).$$

.

Uzyskujemy, zatem, dla zadania ocenianego na skali 0–m komplet $m+1$ krzywych, przy czym:

- pierwsza krzywa ma kształt krzywej logistycznej 2PLM z ujemnym parametrem dyskryminacji (funkcja malejąca) oraz z parametrem trudności $b_{i,0}$;
- krzywe dla kategorii pośrednich x mają kształt dzwonowaty, przy czym dla wyższych kategorii punktowych maksimum funkcji przypada bardziej na prawo niż dla niższych kategorii punktowych; konkretnie, dla kategorii x, maksimum przypada w punkcie $\theta = (b_{i,x-1} + b_{i,x})/2$;
- ostatnia krzywa, dla maksymalnej liczby punktów dla danego zadania, ma kształt krzywej logistycznej 2PLM z parametrem trudności $b_{i,x\text{-}1}$.

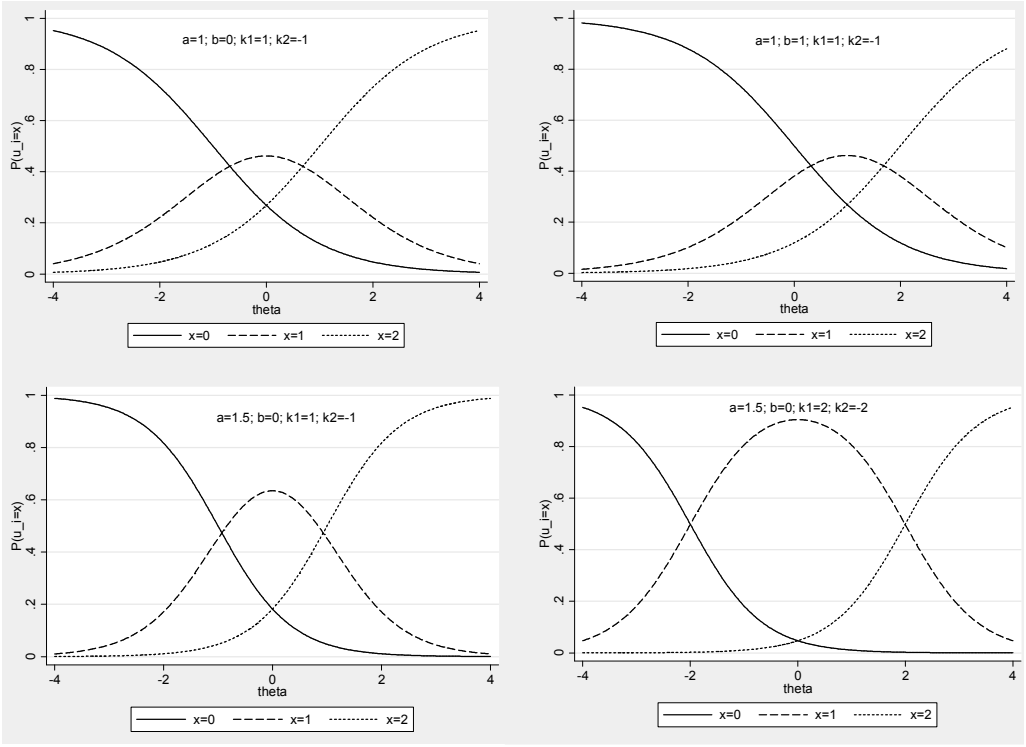
Dla raportowania parametrów modelu GRM powszechnie przyjęto konwencję, w której zamiast podawania poszczególnych $b_{i,x}$ występujących we wzorze na $P_{x,}$ podaje się jeden wspólny parametr położenia b_i, będący średnią z parametrów $b_{i,x}$, oraz parametry kategorii będące odchyleniami od b_i: $k_{ix+1} = b_i - b_{i,x}$.

Przykład czterech krzywych dla zadania ocenianego na skali 0–2 znajduje się na Rysunku 3. Wyjściowo w lewym górnym rogu podano krzywe dla zadania z dyskryminacją równą $a_i = 1$, parametrem położenia $b_i = 0$ i parametrami kategorii ±1. W prawym górnym rogu mamy zadanie, w którym zmieniono jedynie parametr b_i – jego zmiana powoduje analogiczne przesunięcie krzywych, jakie można zaobserwować w modelu 2PLM. W lewym dolnym rogu w wyjściowych parametrach zwiększono jedynie parametr dyskryminacji a_i – spowodowało to zwiększenie stromości krzywych dla skrajnych kategorii oraz zagęszczenie prawdopodobieństwa uzyskania pośredniej kategorii punktowej. Zatem zadanie po zwiększeniu a_i stało się bardziej dyskryminujące – uzyskanie przez ucznia każdej kategorii punktowej niesie ze sobą bardziej precyzyjną informację o jego poziomie umiejętności. W prawym dolnym rogu, oprócz zwiększenia dyskryminacji, zwiększono odchylenia od parametru b_i do ±2. Spowodowało to, że najczęściej uzyskiwaną kategorią punktową przez uczniów w dość szerokim zakresie umiejętności (od –2 do 2) będzie 1 punkt.

Oprócz opisanego modelu GRM zaproponowanego przez Samejimę (1969), innym popularnym modelem dla zadań ocenianych wielopunktowo jest model odpowiedzi częściowej (*partial credit model*, PCM) Geoffa Mastersa (1982), będący uogólnieniem modelu Rascha dla zadań ocenianych kategorialnie. Model PCM został dalej zmodyfikowany przez Eiji'ego Murakiego (1992) tak, aby dopuścić zróżnicowany parametr dyskryminacji. W efekcie powstały model GPCM (*generalized PCM*) stał się bardziej elastyczny, jednak wykroczył poza konserwatywne ramy modeli Rascha. Kompleksowe porównanie modeli dla zadań ocenianych wielokategorialnie można znaleźć u Davida Thissena i Lynne Steinberg (1986).

Wybór modelu i liczebność próby kalibracyjnej

Istotnym czynnikiem, jaki należy rozważyć przy podejmowaniu decyzji o wyborze, czy skorzystać z modelu 1-, 2- czy 3PM, jest



Rysunek 3. Przykład krzywych charakterystycznych dla modelu GRM dla czterech zadań ocenianych na skali 0–2.

dostępna liczebność próby, na której będzie przeprowadza kalibracja testu. W warunkach nieograniczonych przez liczebność badanej grupy model trójparametryczny będzie w zdecydowanej większości przypadków najlepszym rozwiązaniem. Model posiadający największą liczbę parametrów będzie gwarantował najlepsze dopasowanie do danych, a co za tym idzie – najwyższą precyzję pomiaru. Jednak w realnych sytuacjach badacz rzadko dysponuje nieograniczoną możliwością wyboru liczebności badanej grupy. Im liczebność próby mniejsza, tym oszacowania parametrów zadań są mniej dokładne, co w konsekwencji pogarsza oszacowanie poziomu umiejętności uczniów. Frederic Lord (1980) wskazywał na to, że przy mało licznych próbach badawczych uzyskuje się bardziej precyzyjne pomiary, używając modelu jednoparametrycznego, niż modeli bardziej złożonych nawet, gdy proces odpowiedzi na zadania wyraźnie odzwierciedla strukturę dwu- lub trójparametryczną. Obarczone wyniki estymacji parametru dyskryminacji lub zgadywania stanowią bowiem większy problem dla szacowania poziomu umiejętności ucznia, niż błędy spowodowane niedopasowaniem zadań do modelu IRT.

Liczne badania symulacyjne pokazują, że do oszacowania, z zadowalającą dokładnością, prostszych modeli IRT (tj. modeli z mniejszą liczbą parametrów) potrzeba znacznie mniej licznych prób. Dla testów złożonych z zadań dychotomicznych model jednoparametryczny, z dobrą dokładnością, daje się szacować na próbach rzędu 100–200 uczniów (Ayala, 2009). Model

dwuparametryczny jest bardziej wymagający, a liczba 500 uczniów wydaje się tu być bardziej odpowiednia (Stone, 1992). Według niektórych autorów model trójparametryczny zachowuje dostateczną precyzję szacowania parametrów przy 1000 uczniów, ale z zastrzeżeniem, że zadania mają wysoką dyskryminację, a rozkład umiejętności w grupie, na której jest przeprowadzana estymacja, zawiera stosunkowo dużo uczniów zarówno o wysokich, jak i niskich umiejętnościach (DeMars, 2010). W innych sytuacjach liczebność próby niezbędna do oszacowania modelu trójparametrycznego przekracza 2000 uczniów (Woods, 2008).

Ze zwiększeniem liczebności próby potrzebnej do estymacji modelu trzeba również liczyć się w sytuacji, gdy zadania punktowane są na skali dłuższej niż dwukategorialna. Seung Choi, Karon Cook i Barbara Dodd (1997) pokazują, że w przypadku modelu odpowiedzi częściowej (PCM) dla trzech kategorii odpowiedzi potrzeba przynajmniej 250 uczniów. Gdy liczba kategorii wzrasta do sześciu, wymagania co do próby rosną do 1000 uczniów. Modele dla zadań o większej liczbie kategorii oraz z parametrem dyskryminacji (GPCM, GRM) zwiększają wymagania co do liczebności. Na przykład przy trzech kategoriach odpowiedzi, rozsądną dolną granicą jest 500 uczniów (Reise i Yu, 1990).

Oczywiście, przytoczone liczebności próby są tylko wartościami orientacyjnymi. Liczebność próby zależy od długości testu, jakości zadań, rozkładu umiejętności uczniów w badanej populacji, rodzaju estymacji i wreszcie od precyzji pomiaru, na jakiej zależy badaczowi. Przedstawione wartości odwołują się do sytuacji typowych, czyli do testów dłuższych niż 20 zadań, charakteryzujących się dobrze dopasowanymi zadaniami i prób uczniów losowanych z populacji o rozkładzie umiejętności zbliżonym do normalnego.

Szacowanie poziomu umiejętności, funkcja informacji

Często zachodzi potrzeba, żeby oprócz parametrów określających jednowymiarowy model IRT (1), czyli parametrów zadań oraz parametrów rozkładu umiejętności w całej populacji, oszacować również poziom umiejętności θ pojedynczych uczniów, na podstawie zaobserwowanych dla nich wektorów odpowiedzi $\boldsymbol{U} = \boldsymbol{u}$. Szacowania poziomu umiejętności pojedynczych uczniów dokonuje się poprzez odwołanie się do twierdzenia Bayesa:

$$\mathbb{P}(parametr|dane) = \\ = \frac{\mathbb{P}(dane|parametr)\mathbb{P}(parametr)}{\mathbb{P}(dane)}.$$

Uzyskawszy oszacowania parametrów modelu (1) w wyniku kalibracji, można twierdzenie Bayesa zastosować w celu przedstawienia rozkładu a posteriori, parametru θ, pod warunkiem zaobserwowania wektora odpowiedzi $\boldsymbol{u}$ w następujący sposób:

$$\mathbb{P}(\theta|\boldsymbol{U} = \boldsymbol{u}) = \frac{f(\boldsymbol{u}, \theta, \beta)\psi_{\mathcal{P}}(\theta)}{\int f(\boldsymbol{u}, \theta, \beta)\psi_{\mathcal{P}}(\theta)\, d\theta}. \qquad (6)$$

Wzór (6) określa zatem nie punktowe oszacowanie poziomu umiejętności ucznia, ale cały rozkład prawdopodobieństwa dla poziomu umiejętności ucznia, dzięki czemu dostarcza również informacji o niepewności, z jaką umiejętność została oszacowana. W przypadku konieczności uzyskania punktowego oszacowania poziomu umiejętności ucznia mamy do dyspozycji dwa podstawowe rozwiązania:

- Estymator EAP (*expected a posteriori*), będący wartością oczekiwaną dla rozkładu (6), rozwiązanie to wymaga całkowania numerycznego całego (6) po rozkładzie $\psi_{\mathcal{P}}$;
- Estymator MAP (*maximum a posteriori*), będący maksimum funkcji gęstości rozkładu (6) – to rozwiązanie jest o tyle prostsze, że sprowadza się do znalezienia maksimum funkcji $f(\boldsymbol{u}, \theta, \beta)\ \psi_{\mathcal{P}}(\theta)$.

We wczesnych zastosowaniach IRT korzystano także ze zwykłego estymatora największej wiarygodności (*maximum likelihood estimator*, MLE) umiejętności ucznia, który nie uwzględniał rozkładu umiejętności w populacji. Znalezienie takiego estymatora dla ucznia o wektorze odpowiedzi $\boldsymbol{U} = \boldsymbol{u}$ sprowadza się do znalezienia maksimum funkcji wiarygodności $f(\boldsymbol{u}, \theta, \beta)$. Rozwiązanie to jednak ma dużo wad, w szczególności nie istnieją punktowe estymatory umiejętności dla uczniów o najniższym i najwyższym możliwym do uzyskania wyniku, gdyż funkcja $f(\boldsymbol{u}, \theta, \beta)$ w takich przypadkach nie osiąga maksimum. Ten przykład pokazuje również rolę, jaką odgrywa rozkład poziomu umiejętności θ w całej populacji podczas estymacji poziomu umiejętności pojedynczego ucznia – funkcjonuje on w pewnym sensie jak dodatkowe zadanie i jego wkład jest tym większy, im mniej informacji dostarczają odpowiedzi udzielone na zadania testu. W szczególności, jeżeli uczeń nie odpowiedział na żadne zadanie testu, jego poziom umiejętności będzie równy średniej w populacji, a błąd standardowy będzie równy odchyleniu standardowemu w populacji.

Różnice między wynikami estymatora EAP oraz MAP będą zależały od kształtu rozkładu (6) i będą tym większe, im bardziej będzie on niesymetryczny. Generalnie estymator EAP jest obciążony ujemnie (w stronę zera – tzw. *bayesian shrinkage*), a estymator MAP jest obciążony dodatnio. Punktowy estymator poziomu umiejętności z poprawką na obciążenie zaproponował Thomas Warm (1989). Jego estymator WMLE (*weighted maximum likelihood*) stanowi modyfikację MLE i również nie odwołuje się do rozkładu a posteriori (6), czyli nie uwzględnia informacji a priori o rozkładzie umiejętności w populacji $\psi_{\mathcal{P}}$. Estymator WMLE ma wiele zastosowań, na przykład w testowaniu adaptatywnym (Cheng i Liou, 2000; Wang i Wang, 2001).

Dużą rolę w badaniach edukacyjnych odgrywają również wygenerowane losowo wartości z rozkładu a posteriori ucznia (6), które noszą nazwę *plausible values* (PV). Przeprowadzanie wtórnych analiz na PV, zamiast na punktowych oszacowaniach poziomu umiejętności, pozwala na uwzględnienie błędu pomiarowego związanego z nierzetelnością narzędzi badających umiejętności przy wykorzystaniu standardowych narzędzi statystycznych. Nieuwzględnienie błędu pomiarowego w takich analizach, np. poprzez przeprowadzanie ich na punktowych oszacowaniach umiejętności, prowadzi do obciążenia badanych statystyk oraz ich błędów standardowych. Więcej o roli PV w badaniach edukacyjnych można znaleźć w publikacji Margaret Wu (2005), a także w technicznych raportach z międzynarodowych badań, np. PISA (OECD, 2009).

W zależności od rodzaju estymatora zastosowanego do punktowego oszacowania poziomu umiejętności ucznia, w różny sposób szacowany jest błąd standardowy tego estymatora. Dla estymatora EAP błąd standardowy jest liczony jako pierwiastek z wariancji rozkładu (6), co wymaga przeprowadzenia ponownie całkowania numerycznego. Dla estymatora MAP natomiast błąd standardowy liczy się przez odwołanie się do koncepcji informacji Fishera. Jeżeli poprzez $L(\theta)$ oznaczymy logarytm z funkcji wiarygodności (w przypadku (6) jest to Bayesowska funkcja wiarygodności: $f(\boldsymbol{u}, \theta, \beta)\,\psi_{\mathcal{P}}(\theta)$), to informacja Fishera jest dana wzorem (Lehmann, 1991):

$$I(\theta) = E\left(\left(\frac{dL(\theta)}{d\theta}\right)^2 \middle| \theta\right). \tag{7}$$

Informacja Fishera jest zatem miarą krzywizny funkcji wiarygodności w zależności od θ. Jeżeli dane dostarczają dużo informacji o szukanym parametrze θ, to maksimum funkcji wiarygodności będzie strome, a i wartość $I(\hat{\theta})$ w punkcie maksimum $\hat{\theta}$ będzie

duża, i odwrotnie – przy małej ilości informacji z danych maksimum funkcji wiarygodności będzie bardziej rozmyte i wartość funkcji $I(\hat{\theta})$ będzie mniejsza. Stąd intuicyjnie można wywnioskować odwrotną zależność między funkcją informacji w punkcie $I(\theta)$ a precyzją oszacowania poziomu umiejętności. W rzeczywistości, asymptotycznie zachodzi równość pomiędzy wariancją estymatora a odwrotnością funkcji informacji (Deutsch, 1969). Mimo że ta równość nie będzie ściśle prawdziwa dla skończonej liczby zadań, pozwala dość dobrze i łatwo oszacować błąd standardowy $\hat{\theta}$:

$$SE(\hat{\theta}) \approx \frac{1}{\sqrt{I(\hat{\theta})}}. \quad (8)$$

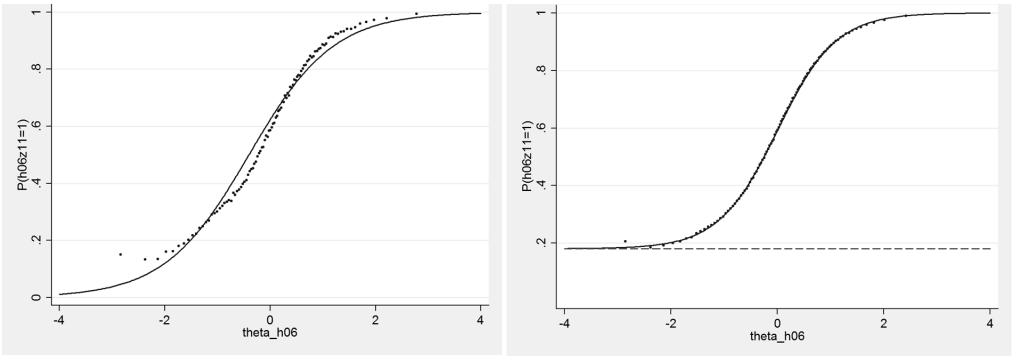
Informacja Fishera ma kilka istotnych właściwości, które są bardzo cenne w kontekście IRT. Po pierwsze, nie będzie zależała od tego, jakich konkretnych odpowiedzi $\boldsymbol{U}=\boldsymbol{u}$ udzielił uczeń. Po drugie, jest addytywna (Rao, 1982), co pozwala na rozbicie jej na sumę wkładów informacji poszczególnych zadań testu. W związku z tym $I(\theta)$ stanowi miarę lokalnej precyzji pomiaru umiejętności zarówno dla całego testu, jak i dla poszczególnych zadań, będącą doskonałą alternatywą do klasycznych miar, jak współczynnik rzetelności dla całego testu oraz na poziomie pojedynczego zadania – współczynnik dyskryminacji. Koncepcja funkcji informacji testu oraz zadania ma bezpośrednie zastosowanie zarówno przy konstrukcji narzędzia, jak i przy testowaniu adaptatywnym. W dalszej części artykułu, właśnie przy opisie testowania adaptatywnego, pojawią się przykłady funkcji informacji dla modelu 2PLM.

Dopasowanie modelu

Każdy model statystyczny jest próbą opisania złożonej struktury danych za pomocą ograniczonej liczby parametrów. Aby wnioski wyciągane na podstawie modelu statystycznego były trafne, musi być on dobrze dopasowany do danych. Oceny dopasowania modelu IRT można dokonywać na różnych poziomach. Na przykład dwa modele IRT oszacowane dla tych samych danych można porównywać ze sobą w całości za pomocą testu ilorazu funkcji wiarygodności, żeby podjąć decyzję, który z nich lepiej opisuje dane. Można też oceniać dopasowanie modelu do wektorów odpowiedzi pojedynczych uczniów (np. dla zidentyfikowania zgadywania lub ściągania odpowiedzi). W końcu można analizować dopasowanie do modelu na poziomie poszczególnych zadań. W tym miejscu zajmiemy się ostatnim przypadkiem.

Analizy stopnia dopasowania zadania do modelu można dokonać na podstawie wizualnej oceny wykresów, na których – obok krzywych charakterystycznych zadań – naniesiono empiryczne wyniki uzyskane przez uczniów, a także przez odwołanie się do odpowiednich statystyk dopasowania.

Przykład metody graficznej zilustrowano na Rysunku 4, który przedstawia dwie krzywe charakterystyczne dla tego samego zadania z egzaminu gimnazjalnego z 2006 r. Wykres z lewej uzyskano dla pierwszej próby dopasowania modelu IRT do danych, w której wszystkie zadania oceniane dychotomicznie modelowano za pomocą 2PLM. Widać systematyczne różnice w przebiegu krzywej wyznaczonej przez dopasowany model na „krzywej" sugerowanej przez empiryczne proporcje odpowiedzi poprawnych. W szczególności, układanie się punktów empirycznych powyżej zera dla uczniów o niskim poziomie umiejętności oraz większa dyskryminacja zadania w rejonie łatwości 0,5 dla punktów empirycznych niż w porównaniu do krzywej 2PLM sugeruje badaczowi, że zadanie to lepiej byłoby opisane modelem trójparametrycznym. W istocie, dla tego samego zadania po prawej uzyskano praktycznie idealne dopasowanie, gdy w ostatecznie użytym dla egzaminu gimnazjalnego modelu IRT było ono modelowane przez 3PLM.



Rysunek 4. Przykład graficznej analizy dopasowania modelu IRT do danych; punkty odpowiadają empirycznym proporcjom odpowiedzi poprawnej w poszczególnych centylach umiejętności.

Warto tutaj zauważyć, że niedopasowanie jednego z zadań całego testu przekłada się na trafność oszacowania rozkładu umiejętności uczniów i wtórnie może się przełożyć na niedopasowanie pozostałych zadań. Jest to w pewnym sensie sytuacja naczyń połączonych. Tak dobre dopasowanie zadania do modelu 3PLM na Rysunku 4 jest nie tylko konsekwencją zmiany modelu dla tego zadania, ale też zmian modelu dla wybranych innych zadań tego testu, jakiej dokonano między pierwszą próbą kalibracji testu (z lewej), a ostateczną kalibracją. W praktyce kalibracja testu w modelu IRT polega właśnie na wielokrotnym, metodą prób i błędów, dopasowywaniu różnego typu modeli dla poszczególnych zadań tak, aby uzyskać jak najmniejsze odchylenia empirycznych wyników od przewidywań modelu. W przypadku skrajnych niedopasowań, badacz może również podjąć decyzję o wykluczeniu zadania z kalibracji.

Wizualna ocena „regularności" odchylania się empirycznych wyników od przewidywań modelu może w pewnych przypadkach przysparzać większych kłopotów, w szczególności dla mniejszych prób. W pewnym momencie granica między tym co systematyczne, a tym co losowe, zaciera się i staje się trudna do obiektywnej oceny na podstawie wizualnej inspekcji wykresów takich jak na Rysunku 4. Obiektywną kwantyfikacją stopnia dopasowania są odpowiednie statystyki. Ogólne podejście, leżące u podstawy większości statystyk dopasowania dla zadań, jest bardzo proste i polega na ocenie rozkładu odchyleń wyników rzeczywistych od wyników szacowanych przez model w arbitralnie wyznaczonych przez umiejętności uczniów grupach.

Podstawową i najprostszą statystyką tego rodzaju jest χ^2 dla IRT opisany przez R. Darrella Bocka (1972):

$$\chi^2 = \sum_{g=1}^{G} \frac{N_g (O_{ig} - E_{ig})^2}{E_{ig}(1 - E_{ig})}. \qquad (9)$$

Aby obliczyć tę statystykę, należy podzielić grupę badanych obserwacji na G grup (zazwyczaj 10) na podstawie szacowanych umiejętności. We wzorze na wartość tej statystyki N_g oznacz liczbę uczniów w grupie g, G to całkowita liczba grup, O_{ig} to proporcja poprawnych odpowiedzi udzielonych na zadanie i w grupie g, natomiast E_{ij} to przewidywana wartość prawdopodobieństwa poprawnej odpowiedzi za pomocą parametrów i. Im wartość statystyki większa, tym zadanie jest gorzej dopasowane. Statystyka ma rozkład χ^2 o ilości stopni swobody równej liczbie

grup pomniejszonej o liczbę parametrów estymowanych dla zadania.

Kolejną statystyką, stosowaną w popularnych programach do estymacji modeli IRT (np. BILOG-MG i PARSCALE), jest G^2 (Muraki i Bock, 2003). Jest to statystyka analogiczna do statyk opartych na ilorazie wiarygodności:

$$G^2 = 2\sum_{g=1}^{G}\left[r_{ig} \log \frac{r_{ig}}{N_g P_i(\bar{\theta}_g)} + \right.$$
$$\left. + (N_g - r_{ig}) \log \frac{N_g - r_{gi}}{N_g \left(1 - P_i(\bar{\theta}_g)\right)} \right]. \quad (10)$$

gdzie: r_{gi} oznacza liczbę poprawnych odpowiedzi na zadanie *i* w grupie *g*, N_g to liczba jednostek w grupie *g*, $P_i(\theta_g)$ oznacza prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie *i* w grupie *k* dla ucznia o średniej wartości θ z grupy *g*, w której θ szacowana jest za pomocą estymatora EAP. Podobnie jak w przypadku poprzedniej statystyki im większa wartość G^2, tym zadanie gorzej dopasowane. G^2 ma rozkład χ^2 o liczbie stopni swobody równej liczbie grup bez korekty na liczbę szacowanych parametrów (dla szczegółowej dyskusji patrz: Muraki i Bock, 2003).

Stosując testy statystyczne do oceny dopasowania zadania, należy pamiętać, że ze wzrostem wielkości próby zwiększa się moc każdego testu statystycznego. Mając na względzie, że częstokroć modelowanie IRT jest stosowane dla licznych prób uczniów, trzeba się liczyć z tym, że nawet niewielkie, z praktycznego punktu widzenia, odstępstwa między przewidywaniem modelu a obserwowanymi wynikami, mogą skutkować pozytywnym wynikiem testu dopasowania. Przy odpowiednio dużej próbie każde analizowane zadanie okaże się istotnie różne od modelowej odpowiedzi. Oprócz istotności statystycznej należy także analizować uzyskaną wartość statystyki dopasowania oraz wspierać się graficzną analizą, o czym wspomniano wcześniej.

Zastosowania IRT

Złożone schematy badawcze

Badacz edukacyjny może spojrzeć na problem testowania umiejętności tak jak statystyk na badanie sondażowe. Analogia przywołana przez Daniela Koretza (2008) uwydatnia jeden z najważniejszych problemów związanych z trafnością i rzetelnością pomiaru. Jeżeli chcemy mieć wiarygodne wyniki badania sondażowego, czyli takie, które w sposób trafny i rzetelny oddają opinie badanej populacji, warunkiem koniecznym jest właściwie dobrana próba badawcza: musi być losowa i odpowiednio duża. Analogicznie w testowaniu edukacyjnym, jeżeli chcemy uzyskać rzetelne i trafne informacje na temat badanych umiejętności, musimy dysponować odpowiednio dużą i dobrze dobraną próbą zadań. W teście, w którym badanych jest pięć umiejętności, daje to niebagatelną liczbę około 100–150 zadań. W praktyce badawczej postulat jak największej liczby zadań stwarza poważne przeszkody natury technicznej. Uczeń podczas jednej sesji testowania może rozwiązać zazwyczaj niezbyt długą listę zadań (aby inne czynniki, np. zmęczenie, nie miały zasadniczego wpływu na wyniki). Testowanie kilkudniowe jest kosztowne i trudne do zrealizowania pod względem organizacyjnym i logistycznym.

Zamiast wydłużać test, można przyjąć inną strategię: maksymalizować liczbę zadań nie dla poszczególnych uczniów, ale dla badanej populacji, tj. żeby różni uczniowie rozwiązywali częściowo różne zadania. Koniecznego instrumentarium statystycznego dostarcza nam IRT, gdyż aby oszacować poziom umiejętności na tej samej skali, nie jest konieczne, aby wszyscy uczniowie odpowiadali na wszystkie zadania badające daną umiejętność.

Jednym ze schematów badawczych posiadających wiele pożądanych charakterystyk jest tzw. BIB7 Wiliama Youdena (*Balanced Incomplete 7-Block Design*). Początkowo używany był w badaniach prowadzonych przez biologów (Preece, 1990), w edukacyjnym kontekście został wykorzystany po raz pierwszy w badaniu NAEP (*National Assessment of Educational Progress*) w roku szkolnym 1983/1984 (Aitkin i Aitkin, 2011; Rutkowski, Gonzales, Joncas i von Davier, 2010). W Tabeli 1 przedstawiono omawiany schemat. Składa się on z siedmiu wersji testu i siedmiu zestawów zadań. Każda z tych wersji składa się z trzech zestawów zadań. Każdy zestaw zadań pojawia się raz w każdej części testu (jako pierwszy, drugi bądź trzeci). Każdy blok zadań rozwiązywany jest przez 43% uczniów. Każdy zestaw zadań pojawia się raz z każdym innym zestawem zadań tak, że dowolną parę zadań z badania rozwiązuje przynajmniej 14% uczniów. Taki schemat badawczy zapewnia ponad dwukrotne zwiększenie liczby wykorzystanych do pomiaru zadań przy stałości liczebności próby badawczej. Zwiększona liczba zadań pozwala na zwiększenie trafności i rzetelności pomiaru.

Podobne schematy badawcze są wykorzystywane w międzynarodowych edukacyjnych badaniach porównawczych. Dobrym przykładem jest tutaj choćby PISA. Dla przykładu, w edycji przeprowadzonej w 2009 r. standardowo testowano trzy dziedziny umiejętności uczniów: czytanie (*reading*), umiejętności matematyczne (*mathematics*), rozumowanie w naukach przyrodniczych (*science*). Wykorzystano 53 pytania dla testu z nauk przyrodniczych, 35 dla testu z matematyki i 131 pytań mierzących umiejętność czytania ze zrozumieniem (OECD, 2012). Do pomiaru umiejętności wykorzystano takie formy zadań, jak: pytania wielokrotnego wyboru, krótkie pytania otwarte oraz pytania otwarte wymagające dłuższej wypowiedzi. Krótkie pytania otwarte wymagały wstawienia odpowiedzi liczbowej, słowa lub krótkiego zdania. Dłuższe pytania otwarte wymagały bardziej rozbudowanej wypowiedzi, często popartej wyjaśnieniem w zaprezentowanej opinii. Wśród pytań wielokrotnego wyboru znalazły się zarówno standardowe, wymagające wybrania najlepszej odpowiedzi spośród czterech propozycji, jak i pytania bardziej złożone, wymagające prawidłowej oceny każdej z podanych alternatyw (na zasadzie „tak/nie", „prawda/fałsz" itp.). Znaczna część zadań wymagała udzielenia krótkiej odpowiedzi, część zadań zamkniętych, mierzących umiejętność czytania ze zrozumieniem, wymagała przeczytania dłuższego tekstu.

Tabela 1
Przykład niekompletnego zbalansowanego schematu badania (tzw. BIB7)

Wersja testu	Układ bloków zadań			Rozkład bloków zadań*						
				A	B	C	D	E	F	G
1	A	B	D	1	1	0	1	0	0	0
2	B	C	E	0	1	1	0	1	0	0
3	C	D	F	0	0	1	1	0	1	0
4	D	E	G	0	0	0	1	1	0	1
5	E	F	A	1	0	0	0	1	1	0
6	F	G	B	0	1	0	0	0	1	1
7	G	A	C	1	0	1	0	0	0	1

* 1 – występuje; 0 – nie występuje

Żaden uczeń nie byłby w stanie rozwiązać w ciągu jednego dnia 219 zadań pojawiających się w badaniu PISA 2009. Narzędzie pomiarowe zostało zatem skonstruowane za pomocą złożonego schematu badawczego. Zadania pogrupowane były w 13 zestawów (7 dla czytania, 3 dla przyrody i 3 matematyczne), przy czym na rozwiązanie każdego z nich przeznaczone było 30 minut. Każdy zestaw zadań składał się z 4 bloków, stworzonych zgodnie ze schematem rotacyjnym przedstawionym w Tabeli 2. Symbole od S_1 do S_3 oznaczają bloki przyrodnicze, R_1 do R_7 bloki czytania ze zrozumieniem, natomiast M_1 do M_3 to bloki matematyczne. Przyporządkowanie bloków opiera się także na zrównoważonym, niekompletnym schemacie blokowym. Każdy z bloków występuje raz na każdej z czterech pozycji w zestawie. Niemożliwe jest również znalezienie takiej pary bloków, która nie występowałaby razem w dwóch różnych zestawach.

Tabela 2
Rozlokowanie bloków w zeszytach testowych PISA

Zestaw	Blok			
1	M_1	R_1	R_3	M_3
2	R_1	S_1	R_4	R_7
3	S_1	R_3	M_2	S_3
4	R_3	R_4	S_2	R_2
5	R_4	M_2	R_5	M_1
6	R_5	R_6	R_7	R_3
7	R_6	M_3	S_3	R_4
8	R_2	M_1	S_1	R_6
9	M_2	S_2	R_6	R_1
10	S_2	R_5	M_3	S_1
11	M_3	R_7	R_2	M_2
12	R_7	S_3	M_1	S_2
13	S_3	R_2	R_1	R_5

Testowanie polegało na losowym przydzieleniu uczniom zestawów pytań, na rozwiązanie których przewidziano dwie godziny. Po upływie pierwszej godziny przewidziana była krótka przerwa. Taki schemat badania pozwolił na testowanie szerokiego wachlarza zadań, przy minimalnym obciążeniu czasowym uczniów.

Przegląd schematów badawczych opartych na niekompletnym schemacie doboru zadań testowych można znaleźć u Andreasa Freya Johannesa Hartiga i André Ruppa (2009). Dla bardziej technicznego opisu wykorzystania niekompletnych schematów doboru zadań odsyłamy czytelnika do prac Roberta Misevy'ego (1991) oraz zespołu Wima van der Lindena (Van der Linden, Veldkamp i Carlson, 2004). Czytelnicy zainteresowani niekompletnymi schematami badawczymi stosowanymi nie tylko w zastosowaniach edukacyjnych, mogą sięgnąć do publikacji Williama Cochrana i Gertrudy Cox (1957) lub Artura Pokropka (2011).

Zrównywanie i łączenie wyników testowych

Poprzez łączenie wyników testowych (*test linking*) rozumiemy szeroką rodzinę metod, w których wyniki uzyskiwane w jednym teście są przekształcane na wyniki uzyskiwane w innym/innych testach. Za najprostszą formę łączenia wyników testowych można uznać przewidywanie wyników w teście *Y* na podstawie wyników uzyskanych w teście *X* za pomocą regresji liniowej. Na drugim końcu można znaleźć tak zwane zrównywanie wyników testowych (*test equating*), w którym nakłada się bardzo ostre ograniczenia na zrównywane testy (m.in. muszą mierzyć tę samą zmienną umiejętności i z taką samą precyzją), jak i na postać funkcji zrównującej testy. Mimo istotnych metodologicznych różnic między wieloma typami łączenia wyników testowych, częstokroć statystyczne rozwiązania wykorzystywane do przeprowadzenia łączenia wyników testowych nie różnią się. Szczegółową typologię metod łączenia wyników

można znaleźć u Paula Hollanda (2007) lub u Michaela Kolena (2004). W naszym artykule ograniczymy się jedynie do przykładu problemu umiejscowienia na wspólnej skali wyników z dwóch testów dla planu nierównoważnych grup z testem kotwiczącym (*nonequivalent groups with anchor test design*, *NEAT*), który może dotyczyć zarówno zrównywania testów, jak i prostszych typów umieszczania na wspólnej skali wyników testów mierzących takie same lub zbliżone umiejętności.

Plan NEAT schematycznie można przedstawić w następujący sposób:

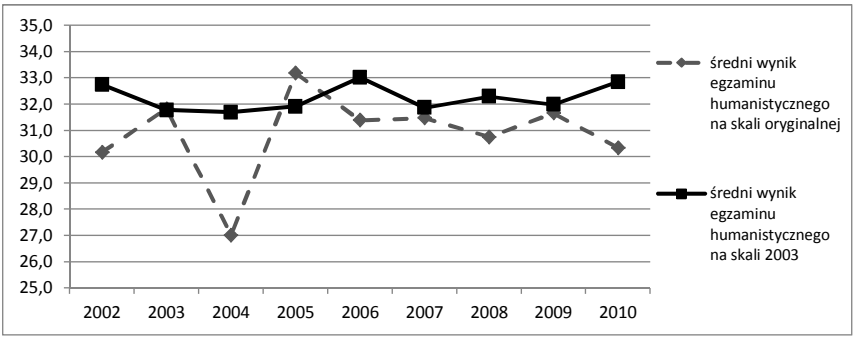
Populacja	Próba	X	Y	A
$\mathcal{P}$	S_1	✓		✓
$\mathcal{Q}$	S_2		✓	✓

Zasadniczym problemem jest tutaj oddzielenie różnic w trudności dwóch różnych testów *X* oraz *Y* od różnic w poziomie umiejętności między populacjami $\mathcal{P}$ i $\mathcal{Q}$. Koniecznej w tym celu informacji dostarczają nam odpowiedzi uzyskane od uczniów pochodzących z obu populacji na dodatkowy test *A*, noszący nazwę testu kotwiczącego (*anchor test*) lub w skrócie – kotwicy (*anchor*).

Modele IRT, poprzez uwzględnienie wprost parametrów rozkładu populacji (1) oraz bezproblemowe radzenie sobie z niekompletnymi schematami zbierania danych (co opisano pierwszej części artykułu) stanowią bardzo dobre narzędzie do rozwiązania takiego problemu. W obrębie IRT wypracowano wiele metod do umieszczania na wspólnej skali rozkładów θ dla populacji $\mathcal{P}$ i $\mathcal{Q}$ oraz parametrów β_i dla testów *X, Y, A*. Wyróżnić można następujące metody (Kolen i Brennan, 2004):

- Łączna kalibracja (*concurrent calibration*) wszystkich trzech testów;
- Oddzielna kalibracja (*separate calibration*) par testów (*X, A*) oraz (*Y, A*), po której stosuje się sprowadzające do wspólnej skali przekształcenia oparte na:
  a) liniowej funkcji parametrów kotwicy – metody: średnia/średnia lub średnia/sigma (*mean/mean*, *mean/sigma*);
  b) krzywych charakterystycznych kotwicy – metoda Stockinga–Lorda lub Haebary;
- Metoda ustalonych parametrów (*fixed parameters method*) dla kotwicy *A*;
- Metoda przekształcania umiejętności (*proficiency transformation*).

Po uzyskaniu rozkładu umiejętności na wspólnej skali θ można również dokonać ich transformacji do rozkładów wyników obserwowanych testów *X* oraz *Y* w populacji docelowej. Przedstawienie rezultatów zrównywania lub łączenia testów na surowych



Rysunek 5. Średnie wyników obserwowanych dla humanistycznej części egzaminu gimnazjalnego dla oryginalnego testu oraz na skali wyników egzaminu z 2003 r.

skalach wyników w teście *X* oraz *Y* jest podyktowane tym, że te skale najczęściej stanowią podstawę raportowania wyników.

Na Rysunku 5 przedstawiono przykładowy praktyczny rezultat przeprowadzenia zrównania wyników obserwowanych humanistycznej części egzaminu gimnazjalnego. Zrównanie przeprowadzono z wykorzystaniem metody łącznej kalibracji modelu IRT. Wyniki obserwowane uczniów z lat 2002–2010 zostały przekształcone na wyniki w teście z roku 2003. Na wykresie widać, że istotne wahania w średnim wyniku testu między latami należy w największej mierze przypisać zmianom w trudności testu, a nie zmianom w poziomie umiejętności uczniów. Więcej o wynikach zrównania egzaminów gimnazjalnych można znaleźć w publikacji zespołu Henryka Szaleńca (2012).

Testowanie adaptatywne

Obszarem, którego rozkwit i rozwój w praktyce w całości opiera się na metodologii związanej z IRT, jest komputerowe testowanie adaptatywne (*Computer Adaptive Testing*, CAT). Temat ten potraktujemy bardziej szczegółowo, ponieważ ukazuje on pełną gamę możliwości, jakie wynikają z lokalnej oceny jakości zadań oraz z błędu oszacowania umiejętności ucznia, a jednocześnie skutkują bardziej rzetelnym, trafnym i ekonomicznym pomiarem umiejętności.

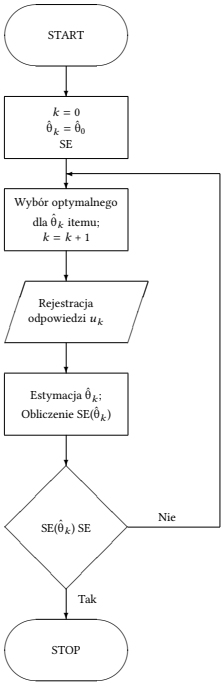
Z pomiarowego punktu widzenia najbardziej optymalnym narzędziem do oceny umiejętności ucznia jest test, który mierzy umiejętności zadaniami z przedziału trudności jak najbardziej zbliżonego do umiejętności ucznia. Innymi słowy, dobry pomiarowo test nie jest ani za łatwy, ani za trudny. Intuicyjnie, jeżeli test jest zbyt łatwy nie pozwoli na rozróżnienie między uczniem średnim a bardzo dobrym. W obu wypadkach wyniki będą bliskie maksymalnym. Jeżeli test jest zbyt trudny, rozróżnienie między uczniem o bardzo niskich umiejętnościach a uczniem o umiejętnościach średnich staje się niemożliwe, gdyż obydwie grupy uzyskują wyniki bliskie minimalnym. Psychometrycznie, zakładając brak zgadywania (parametr c_i w 3PLM równy 0), największą informację dla ucznia niosą zadania, których prawdopodobieństwo poprawnej odpowiedzi wynosi 0,5. W teście trudnym dla danego ucznia średnie prawdopodobieństwo poprawnej odpowiedzi dla puli zadań jest znacznie niższe od 0,5, w teście łatwym znacznie wyższe. W związku z tym, informacja, jaką niesie ze sobą trudny lub łatwy test będzie zawsze mniejsza od testu o przeciętnej trudności (zakładając, że zadania z różnych testów mają zbliżone dyskryminacje i parametry zgadywania).

Z praktycznej strony, poddawanie pomiarowi uczniów narzędziem niedostosowanym trudnością do umiejętności skłania do zachowań niepożądanych i może w sposób zasadniczy wpłynąć na jakość pomiaru. Testy zbyt trudne mogą skłaniać do zgadywania, ściągania, wywołują frustrację, która prowadzi do osłabienia motywacji i często rezygnacji z próby rozwiązania kolejnych zadań. Testy zbyt łatwe mogą szybko prowadzić do znużenia, osłabienia uwagi, która wiąże się ze zwiększeniem liczby przypadkowych błędów, a także doszukiwaniem się dodatkowych treści (tzw. drugiego dna), co może prowadzić uczniów do błędnych odpowiedzi.

W klasycznej technologii testowania każdy uczeń z testowanej grupy dostaje test o takiej samej trudności. Jeżeli jest to klasyczny test umiejętności, konstruktorzy starają się zaprojektować go w taki sposób, aby pod względem trudności dostosowany był do jak najszerszej grupy uczniów. Praktycznie oznacza to, że test dostosowany jest najbardziej do uczniów średnich – zawiera najwięcej zadań, których prawdopodobieństwo odpowiedzi przez średniego ucznia w badanej grupie wynosi około 0,5. Tym

samym, dla uczniów średnich za pomocą takiego testu otrzymujemy najwięcej informacji, a co za tym idzie największą precyzję pomiaru. Dla uczniów bardzo słabych i bardzo dobrych zadowalamy się stosunkowo niską precyzją pomiaru.

Jedynym wyjściem, które zapewniałoby podobną precyzję pomiaru dla uczniów ze skraju rozkładu umiejętności i tych ze środka, byłoby podanie różnych pod względem trudności testów uczniom o różnych umiejętnościach. Uczniowie o niskich umiejętnościach otrzymywaliby relatywnie łatwe testy, uczniowie o średnich umiejętnościach – relatywnie średniej trudności zadania, a uczniowie o wysokich umiejętnościach – testy z zadaniami relatywnie trudnymi. W efekcie każdy uczeń rozwiązywałby inny, dostosowany do jego umiejętności test maksymalizujący informację i minimalizujący błąd pomiaru.



Rysunek 6. Schemat badania CAT na podstawie (Gruijter i van der Kamp, 2005, s. 139).

Taka procedura w kontekście klasycznej teorii testów jest praktycznie niemożliwa do zaimplementowania. Natomiast IRT jest wprost stworzone dla takiego zastosowania – daje możliwość szacowania poziomu umiejętności ucznia na podstawie odpowiedzi na dowolny podzbiór zadań.

W komputerowym adaptatywnym testowaniu postępuje się według następującego algorytmu: najpierw określamy błąd standardowy (SE), z jakim co najwyżej pragniemy uzyskać pomiar poziomu umiejętności ucznia lub najmniejszą wartości informacji (7), jaką chcemy dysponować dla każdego ucznia. Zaczynamy od pewnej startowej wielkości początkowej, szacującej poziom cechy badanej osoby $\hat{\theta}_0$ (np. średnia z populacji) i za każdą udzieloną odpowiedzią uaktualniamy wartość $\hat{\theta}_k$ oraz SE ($\hat{\theta}_k$), a tym samym również $I(\hat{\theta}_k)$. Jeżeli w kroku *k*-tym badanie nie osiągnęło odpowiedniej precyzji, to w następnym kroku przydzielane jest takie zadanie, które w zbiorze pozostałych pozycji w punkcie $\hat{\theta}_k$ ma możliwie największą wartość funkcji informacji. W ten sposób można znacznie skrócić czas badania, ponieważ osoba w miarę udzielania odpowiedzi dostaje zadania coraz bardziej dostosowane do poziomu jej umiejętności i nie musi odpowiadać na wiele zadań, mających małą wartość informacji, które znajdowałyby się prawdopodobnie w tradycyjnej, papierowej wersji testu. Schemat badania CAT przedstawiono na Rysunku 6.

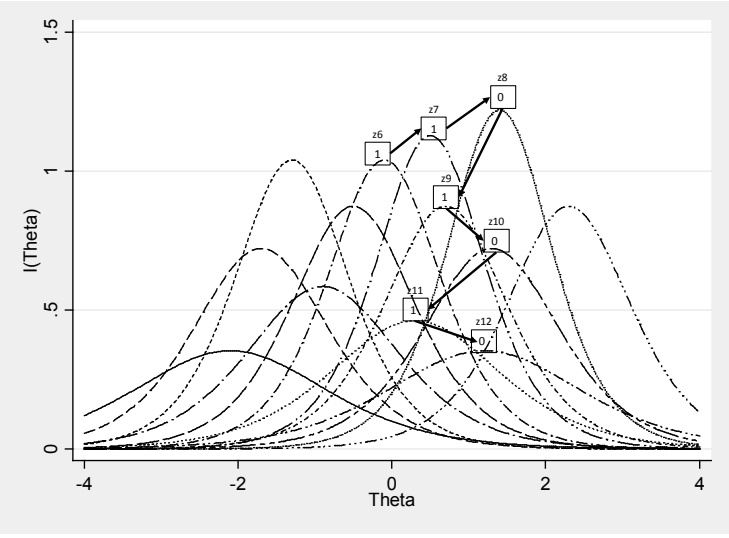
Prześledźmy to na prostym przykładzie. Załóżmy, że mamy 13 zadań o znanych parametrach, które mogą zostać użyte w testowaniu adaptatywnym. Zadania te są dobrze dopasowane do modelu dwuparametrycznego i skalowane są za jego pomocą. Parametry tych zadań przedstawione zostały w Tabeli 3.

Jako że nie jest znany poziom umiejętności ucznia przed rozpoczęciem testowania, pierwsze przypisane mu zadanie powinno mieć trudność jak najbardziej zbliżoną

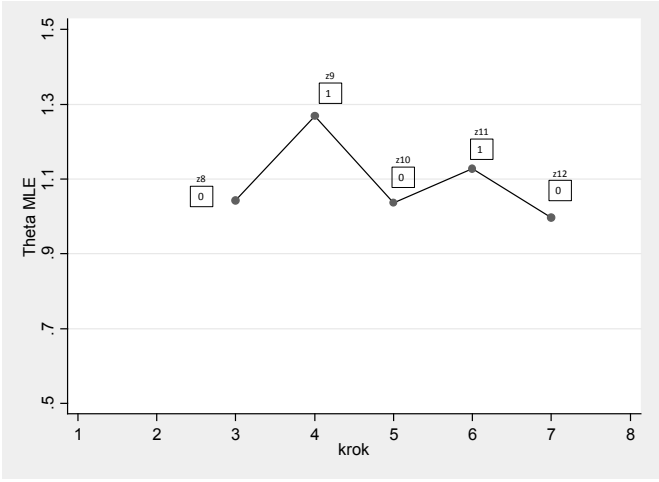
Tabela 3
Parametry zadań przykładowego testu adaptatywnego (a – dyskryminacja; b – trudność)

Zadanie	z1	z2	z3	z4	z5	z6	z7	z8	z9	z10	z11	z12	z13
a	0,70	1,00	1,20	0,90	1,10	1,20	1,25	1,30	1,10	1,00	0,80	0,70	1,10
b	-2,10	-1,70	-1,30	-0,90	-0,50	-0,10	0,50	1,40	0,70	1,30	0,30	1,20	2,30

do średniej w populacji. Zapewnia to najwyższe prawdopodobieństwo uzyskania największej wartości informacji w pierwszych krokach testowania. Takim jest zadanie 6, którego trudność na skali logitowej wynosi -0,1. Jeżeli uczeń rozwiąże je poprawnie, kolejnym wybranym przez algorytm powinno być zadanie trudniejsze, gdyż przewidywana umiejętność ucznia po poprawnym wykonaniu pierwszego zadania nie powinna być mniejsza od trudności tego zadania. Następne zadanie powinno być trudniejsze, lecz bliskie przewidywanemu poziomowi umiejętności ucznia, tzn. nie może być zbyt trudne. Czwarte zadanie, oprócz zwiększonej trudności, powinno również mieć największą z możliwych wartość informacyjną. Natomiast jeżeli uczeń rozwiąże źle zadanie z pierwszego kroku, kolejne zadanie powinno być łatwiejsze oraz o jak najwyższej wartości informacyjnej. Załóżmy, że nasz przykładowy uczeń rozwiązał prawidłowo zadanie 6. Algorytm podaje mu kolejne zadanie: 7. Oprócz zadania 7 rozpatrywane mogłoby być zadanie 11, bardziej zbliżeone trudnością do zadania 6 niż zadanie 7. Zadanie 11 nie zostało wybrane w tym kroku, ponieważ miało znacznie mniejszą wartość informacyjną (przede wszystkim z powodu niższej dyskryminacji). Ilustracyjnym uzasadnieniem dla takiego wyboru jest Rysunek 7. Przedstawiono na nim funkcje informacyjne dla każdego zadania z testu



Rysunek 7. Funkcje informacyjne zadań oraz kroki testowania dla wybranego ucznia.



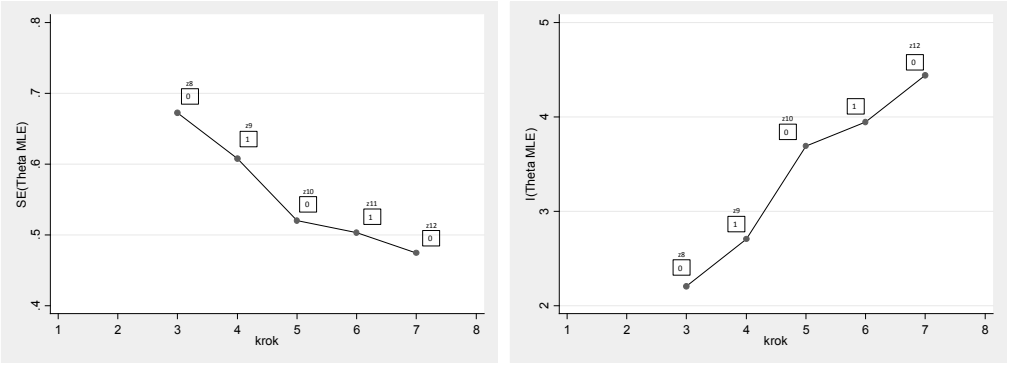
Rysunek 8. Wartości estymowanych umiejętności dla kolejnych kroków testowania dla wybranego ucznia dla estymatora MLE.

oraz kolejne kroki procedury CAT dla jednego ucznia. W kwadratach oznaczono, czy uczeń rozwiązuje zadanie poprawnie (1), czy niepoprawnie (0).

Po poprawnym rozwiązaniu zadania 6 rozwiązanie zadania 7 powinno przynieść największą ilość informacji. Załóżmy, że zadanie 7 zostało rozwiązane poprawne. Kolejnym zadaniem, które zostanie wybrane przez algorytm CAT będzie zadanie 8. trudniejsze niż poprzednie i potencjalnie przynoszące największą ilość informacji. Jeżeli tym razem zadanie nie zostanie rozwiązane poprawnie, algorytm powinien wybrać zadanie łatwiejsze o jak największej wartości funkcji informacyjnej – w naszym wypadku jest to zadanie 9. Procedura przydzielania kolejnych zadań trwa aż do uzyskania zadowalającego $SE(\hat{\theta})$ dla ucznia. W naszym przykładzie procedura testowania została zakończona po 7 zadaniach.

W testowaniu adaptatywnym, gdy używamy estymatora MLE, nie jest możliwe oszacowanie umiejętności ucznia, gdy ten odpowiada tylko poprawnie lub tylko niepoprawnie. W przypadku pierwszych kroków uznaje się, że uczeń ma umiejętność nie mniejszą niż trudność rozwiązanego poprawnie zadania, lub nie większą od zadania rozwiązanego niepoprawnie. W prezentowanym przykładzie algorytm najpierw wybiera zadanie 6 – najbliższe średniemu poziomowi umiejętności z wysoką wartością funkcji informacyjnej. Po poprawnym rozwiązaniu zadania uczniowi przydzielane jest zadanie trudniejsze o najwyższej wartości funkcji informacyjnej. To zadanie zostało poprawnie rozwiązane, algorytm postępuje dalej, następnym trudniejszym zadaniem o najwyższej wartości funkcji dyskryminacyjnej jest zadanie 8, w tym wypadku uczeń odpowiada niepoprawnie. Od tego kroku można oszacować wstępnie poziom umiejętności ucznia. Opis procedur wyboru zadań oraz szacowania poziomu umiejętności dla innych estymatorów umiejętności niż MLE w kompleksowy sposób opisano u Wima van der Lindena i Petera Pashley'a (2010).

Na Rysunku 8 przedstawiono wartości estymowanych umiejętności dla kolejnych kroków testowania, począwszy od kroku 3. Z każdym krokiem estymacji błąd oszacowania poziomu umiejętności maleje,



Rysunek 9. Wartości błędów standardowych oraz informacji dla wybranego ucznia w kolejnych krokach estymacji.

a wartość informacji rośnie, co z kolei przedstawione zostało na Rysunku 9.

Tak jak w przypadku estymacji poziomu umiejętności w pierwszych krokach, błąd standardowy i wartość informacji nie są szacowane dopóty, dopóki uczeń zachowuje maksymalną lub minimalną liczbę punktów możliwych do uzyskania w *k*-tym kroku. Gdy tylko uczeń błędnie rozwiąże zadanie po serii poprawnych rozwiązań lub odpowie poprawnie po serii błędnych rozwiązań, wartości $\hat{\theta}_k$, a wraz z nią wartości błędu standardowego (*SE*) i informacji zostają szacowane. W każdym kolejnym kroku błąd standardowy zmniejsza się, a wartość informacji rośnie.

W prezentowanym przykładzie zakończono testowanie po siedmiu zadaniach, gdy dla estymatora MLE osiągnięty został błąd standardowy mniejszy niż 0,5, korespondujący z wartością informacji powyżej 4. Oczywiście w zastosowaniach praktycznych wymaga się większej precyzji, co skutkuje większą liczbą kroków CAT, a co za tym idzie z większą liczbą zadań. W praktyce zazwyczaj nawet po osiągnięciu odpowiedniej precyzji zachowuje się minimalną długość testu (która oczywiście jest znacznie krótsza niż w wersji klasycznej), by uchronić się przed ewentualnymi zarzutami niesprawiedliwości procedury testowania (Linacre, 2000).

Warto podkreślić, że testowanie za pomocą procedury CAT jest niezwykle efektywne. Odwołując się do podawanego przykładu po zadaniu 7 z 13 zadań dla estymatora MLE otrzymujemy wartość równą 1,00, a błąd standardowy tego oszacowania wynosi 0,47. Gdyby rozpatrywany przez nas uczeń rozwiązał również dodatkowe zadania zgodnie z najbardziej prawdopodobnym wzorcem odpowiedzi (czyli zadania 1–5 poprawnie, zadanie 13 niepoprawnie), poziom jego umiejętności oszacowany za pomocą estymatora MLE wyestymowany zostałby na poziomie 0,9, a błąd standardowy wyniósłby 0.45. Innymi słowy, zadając dodatkowo niemalże drugie tyle zadań, poprawilibyśmy oszacowanie poziomu umiejętności nie więcej niż o 10% i zredukowalibyśmy standardowy błąd pomiaru o nie więcej niż 5%.

CAT w dobie rozwijających się technologii informatycznych jest kuszącą alternatywą wobec klasycznego testowania, obok indywidualnego dopasowania testu do ucznia. Wiąże się z krótszym pomiarem, przy zachowaniu założonej precyzji i minimalizacji niepożądanych zachowań testowych. Testowanie CAT pozwala na kontrolę tajności zadań (praktycznie żaden uczeń nie rozwiązuje takiego samego zestawu zadań). Stosowanie CAT pozwala również na

szybsze zbieranie danych i bieżącą poprawę właściwości psychometrycznych testu, poprzez standaryzacje zadań w naturalnych warunkach testowania (Deville, 1993).

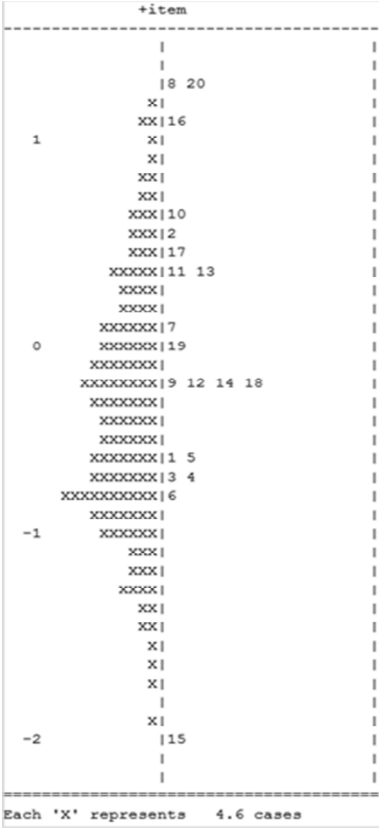
Wadą CAT jest to, że do testowania niezbędny jest komputer, a w większości wypadków odbywa się ono bezpośrednio za pomocą komputera. Umiejętności uczniów, którzy nie mieli wcześniej dostępu do komputera, mogą być w takiej sytuacji szacowane z wyraźnym błędem. Na szczęście dla CAT z każdym rokiem komputery stają się nieodzowną częścią życia, o czym świadczą np. wyniki badania PISA 2009, w którym 94% polskich gimnazjalistów potwierdziło, że posiadają komputer w domu, a 99,5% że używało komputera przynajmniej raz (OECD, 2012). Nie zmienia to faktu, że do przeprowadzenia testów CAT potrzebna jest odpowiednia liczba komputerów i oprogramowania w dyspozycji realizatora badania, co przy wprowadzaniu tej technologii może wiązać się z jednorazowym dużym kosztem.

Oprócz zaplecza sprzętowego potrzebna jest odpowiednio duża liczba zadań o znanych parametrach, czyli zadań przetestowanych pilotażowo. Reguła kciuka mówi o tym, że liczba zadań w banku powinna być przynajmniej dziesięciokrotnie większa niż przewidywana liczba zadań wykorzystanych podczas testowania dla jednego ucznia, czyli przy teście w którym jeden uczeń przeciętnie ma rozwiązać około 30 zadań, liczba zadań w banku nie powinna być mniejsza niż 300 (Weiss, 2004).

Szczegółowa problematyka dotycząca CAT znacznie wykracza poza ramy tego artykułu, warto jednak nadmienić, że procedury stosowane w praktyce w CAT bywają znacznie bardziej złożone niż przedstawiony przez nas poglądowy algorytm. Mowa tutaj choćby o algorytmach, w których dokonuje się rekalibracji trudności zadań podczas testowania, kontroli pokrycia nie tylko odpowiedniego poziomu trudności zadań, ale i treści programowych, wprowadzanie adaptatywnego testowania wielowymiarowego, wprowadzanie algorytmów, które umożliwiają powrót uczniom do wcześniej rozwiązanych zadań (por. van der Linden i Glass, 2000).

Mapowanie zadań

Bardzo popularnym narzędziem usprawniającym interpretację skali umiejętności oraz właściwości zadań są tzw. mapy zadań. Mapa taka przedstawia na jednej skali rozkład umiejętności uczniów oraz rozkład trudności zadań. Przykład przedstawiony został na Rysunku 10: z lewej strony znakiem X zaznaczeni zostali uczniowie (średnio jeden X odzwierciedla 4,6 uczniów), a po



Rysunek 10. Przykładowa mapa zadań stworzona w programie ConQuest.

prawej stronie numerami zostały zaznaczone kolejne zadania. Mapa testu może służyć ocenie trafności narzędzia pomiarowego (Wilson, 2005). Uczeń, nauczyciel czy rodzic może odnieść dzięki temu wynik ucznia do poziomu trudności konkretnego zadania i skupić pracę dydaktyczną wokół nich oraz zaplanować pracę od nauki zdań łatwych, poprzez średnie, aż do najtrudniejszych. Wynik ucznia skonfrontowany z konkretnym zadaniem nadaje znaczenia abstrakcyjnej liczbie. Analitycy, korzystający z takich map, mogą porównywać rozkład uczniów i zadań, aby uzyskać szczegółową informację na temat poziomu umiejętności.

Stworzenie mapy zadań jest najłatwiejsze dla modelu 1PLM. Dla modeli o większej liczbie parametrów staje się to o tyle problematyczne, że wymaga wyboru kryterium określającego trudność zadania, czyli poziomu prawdopodobieństwa poprawnej odpowiedzi, co do którego będzie się odnosiła miara trudności zadań. W omawianym już przykładzie, odwołującym się do Rysunku 1, dla modelu dwuparametrycznego, gdyby za punkt określający trudność wybrać 50%, wszystkie 3 zadania miałyby taki sam jej poziom. Jeżeli wybrany zostanie punkt powyżej 50% trudności, zadania będą miały odwrotną hierarchię niż w przypadku, gdy wybierzemy punkt poniżej 50%. Położenie zadań względem siebie na mapie w modelu 1PLM będzie takie samo, niezależnie od tego wybranego kryterium określającego jego trudność. Przy konstrukcji map zadań dla modeli o większej liczbie parametrów zazwyczaj postępuje się za badaniem NAEP. Trudność zadań określa się w nim na poziomie 65% poprawnej odpowiedzi dla zadań modelowanych za pomocą modelu dwuparametrycznego. W przypadku zadań modelowanych za pomocą modelu trójparametrycznego dobiera się poziom w zależności od wartości parametru c_i, przy czym wynosi on 72 lub 74% (DeMars, 2010, s. 79–80).

Podsumowanie

Niniejszy artykuł miał na celu przybliżyć i uporządkować czytelnikowi zagadnienia związane z tematem modelowania IRT, które dotychczas w piśmiennictwie krajowym nie doczekały się systematycznego opracowania odzwierciedlającego współczesny stan badań i jednocześnie odnoszącego się do kwestii związanych z pomiarem edukacyjnym.

Autorzy wyrażają nadzieję, że przedstawione przykłady zastosowań IRT zachęcą do częstszego sięgania do tych modeli przy rozwiązywaniu praktycznych problemów, w projektowaniu badań edukacyjnych oraz analizowaniu pochodzących z takich badań danych. Godna odnotowania w tym kontekście jest wciąż rosnąca liczba krajowych badań, w których analizy oparte na IRT stanowią istotny komponent. Wystarczy wspomnieć kilka najnowszych projektów: Henryk Szaleniec z zespołem (Szaleniec i in., 2012) wykorzystał modelowanie IRT do zrównywania wyników egzaminacyjnych. Modelowanie IRT wykorzystywane jest w konstrukcji wskaźnika EWD (Pokropek, 2008, Pokropek i Żółtak, 2012). Wielowymiarowe modelowanie IRT wraz z generowaniem PV do wtórnych analiz regresyjnych z szeregiem zmiennych kontekstowych zastosowano w projekcie badającym umiejętności trzecioklasistów (Kondratek, 2012). Zastosowania IRT w polskiej literaturze można znaleźć w badaniu efektu egzaminatora (Dubiecka, Szaleniec i Węziak 2006), w testowaniu adaptatywnym (Kaczan i Rycielski, 2012). Polscy badacze coraz częściej sięgają po modelowanie IRT podczas konstrukcji narzędzi pomiarowych (Jasińska i Modzelewski, 2012), jak również wykorzystują możliwości IRT do tworzenia złożonych schematów badań (IBE, 2011)

Jednocześnie warto nadmienić, że większość z przedstawionych w tym artykule informacji może zostać przeniesiona na

pole innych dyscyplin badawczych, w których następuje inferencja o poziomie ukrytych cech na podstawie odpowiedzi zebranych w narzędziach w formie testów lub kwestionariuszy, jak np. w psychologii czy w badaniach klinicznych.

Literatura

Aitkin, I. i Aitkin, M. (2011). *Statistical modeling of the National Assessment of Educational Progress*. New York: Springer.

Ayala, R. J. de (2009). *The theory and practice of Item Response Theory*. New York – London: The Guilford Press.

Birnbaum, A. (1968). Some latent trait models. W: F. M. Lord i M. R. Novick (red.), *Statistical theories of mental test scores*. Reading: Addison – Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.

Cheng, P. E. i Liou, M. (2000). Estimation of trait level in Computerized Adaptive Testing. *Applied Psychological Measurement*, *24*(3), 257–265.

Choi, S. W., Cook, K. F. i Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement*, *1*(2), 114–142.

Cochran, W. G. i Cox, G. M. (1957). *Experimental designs*. New York: John Willey & Sons.

De Boeck, P. i Wilson, M. (red.). (2004). *Explanatory item response models: a generalized linear and nonlinear approach*. New York: Springer.

DeMars, C. (2010). *Item Response Theory*. Oxford – New York: Oxford University Press.

Deutsch, R. (1969). *Teoria estymacji*. Warszawa: Państwowe Wydawnictwa Naukowe.

Deville, C. (1993) Flow as a testing ideal. *Rasch Measurement Transactions*, *7*(3), 308.

Dubiecka, A, Szaleniec, H. i Węziak, D. (2006). Efekt egzaminatora w egzaminach zewnętrznych. W: B. Niemierko i M. K. Szmigel (red.), O wyższą jakość egzaminów szkolnych, cz. I, Etyka egzaminacyjna i zagadnienia ogólne (s. 526–355). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Frey, A., Hartig, J. i Rupp, A. A. (2009). An NCME Instructional module on booklet designs in large-scale assessments of student achievement: theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53.

Gruijter, D. N. M. i Kamp, L. J. van der (2005). *Statistical test theory for education and psychology.* Pobrano z: http://irt.com.ne.kr/data/test_theory.pdf

Holland, P. W. (2007). A framework and history for score linking. W: N. J. Dorans, M. Pommerich i P. W. Holland (red.), *Linking and aligning scores and scales* (s. 5–30). New York: Springer.

IBE (2011). «Laboratorium myślenia». Diagnoza umiejętności gimnazjalistów w zakresie przedmiotów przyrodniczych. Raport z badań. Pobrano z: http://eduentuzjasci.pl/pl/publikacje-ee-lista/162-raport/raport-z-badania/laboratorium-myslenia/812-laboratorium-mysleniа-raport-z-badania.html

Jasińska, A., i Modzelewski M. (2012). Można inaczej. Wykorzystanie IRT do konstrukcji testów osiągnięć szkolnych. W: B. Niemierko i M. K. Szmigel (red.), *Regionalne i lokalne diagnozy edukacyjne* (s. 178–187). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Kaczan, R. i Rycielski, P. (2012). Diagnoza umiejętności dzieci 5-, 6- i 7-letnich za pomocą Testu Umiejętności na Starcie Szkolnym (TUnSS). Referat wygłoszony na konferencji Polskiego Towarzystwa Diagnostyki Edukacyjnej, Wrocław. Pobrano z: http://www.ptde.org/file.php/1/Archiwum/XVIII_KDE/XVIII%20KDE%20-%20referaty/Kaczan,Rycielski.pdf

Kolen, M. J. (2004). Linking assessments: concept and history. *Applied Psychological Measurement*, *28*(4), 219–226.

Kolen, M. J. i Brennan R. L. (2004). *Test equating, scaling, and linking: method and practice* (wyd. 2). New York: Springer.

Kondratek, B. (2012). Konteksty osiągnięć uczniów. W: M. Żytko (red.), *Badanie umiejętności podstawowych uczniów trzecich klas szkoły podstawowej. Uczeń, szkoła, dom. Raport z badań*. (s. 187–217). Warszawa: Instytut Badań Edukacyjnych.

Koretz, D. (2008). *Measuring up: what educational testing really tells us*. Cambridge: Harvard University Press.

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology, 64*(3), 478–497.

Lehmann, E. L. (1991). *Teoria estymacji punktowej*. Warszawa: Wydawnictwo Naukowe PWN.

Linacre, M. (2000). *Computer Adaptive Testing: a methodology whose time has come*. MESA Memorandum No. 69.

Linden, W. J. van der i Glas, C. A. W. (2000). *Computerized Adaptive Testing: theory and practice*. Norwell: Kluwer Academic.

Linden, W. J. van der i Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. W: W. J. van der Linden i C. A. W. Glas (red.), *Elements of adaptive testing* (s. 3–30). New York: Springer.

Linden, W. J. van der, Veldkamp, B. P. i Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, *28*(5), 317–331.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates.

Lord, F. M. i Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison – Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Muraki, E. i Bock, D. (2003). PARSCALE 4.0 [Instrukacja programu komputerowego]. Lincolnwood: Scientific Software International.

OECD (2009). *PISA 2006. Technical Raport*. Paris: OECD Publishing.

OECD (2012). *PISA 2009. Technical Raport*. Paris: OECD Publishing.

Pokropek, A. (2008). *Metody obliczania edukacyjnej wartości dodanej dla szkół kończących się egzaminem maturalnym.* W: B. Niemierko i M. K. Szmigel (red.), *Uczenie się i egzamin w oczach nauczyciela* (s. 237–247). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Pokropek, A. (2011). Missing by design: planned missing-data designs in social science. *ASK. Research&Methods*, 20, 81–105.

Pokropek, A. i Żółtak, T. (2012). *Nowe modele jednorocznej EWD. W:* B. Niemierko i M. K. Szmigel (red.), *Regionalne i lokalne diagnozy edukacyjne (*s. 178–187*)*. Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Preece, D. A. (1990). Fifty years of Youden squares: a review. *Bulletin of the Institute of Mathematics and Its Applications*, *26*(4), 65–75.

Rao, C. R. (1982). *Modele liniowe statystyki matematycznej*. Warszawa: Państwowe Wydawnictwo Naukowe.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Reise, S. P. i Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133–144.

Rutkowski, L., E. Gonzalez, M. Joncas i Davier, M. von (2010). International large-scale assessment data. *Educational Researcher*, *39*(2), 142–151.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* [Psychometric Monograph No. 17]. Richmond: Psychometric Society.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: an evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1–16.

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F. i Pokropek, A. (2012). Wyniki egzaminu gimnazjalnego 2002–2010 na wspólnej skali. *Edukacja*, *119*(3), 9–30.

Thissen, D. J. i Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.

Wang, S. i Wang. T, (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in Computerized Adaptive Testing. *Applied Psychological Measurement*, *25*(4), 317–331.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika*, *54*(3), 427–450.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 70–84.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah: Lawrence Erlbaum Associates.

Woods, C. M. (2008). Consequences of ignoring guessing when estimating the latent density in item response theory. *Applied Psychological Measurement*, *32*(5), 371–384.

Wright, B. D. (1983). *Fundamental measurement in social science and education*. Research Memorandum No. 33a MESA Psychometric Laboratory. Pobrano z: http://www.rasch.org/memo33a.htm

Wright, B. D. i Stone, M. (1979). *Best test design*. Chicago: MESA Press.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114–128.