Rozdział 1
Teoria odpowiedzi na pozycje testowe: jednowymiarowe modele dla cech ukrytych o charakterze ciągłym[1]

Bartosz Kondratek, Instytut Badań Edukacyjnych
Artur Pokropek, Instytut Badań Edukacyjnych

Pod nazwą „teoria odpowiedzi na pozycje testowe" (*item response theory*, IRT) kryje się rodzina narzędzi statystycznych wykorzystywanych do modelowania cech ukrytych o charakterze ciągłym na podstawie dyskretnych wskaźników. Mogą one mieć różny charakter. Najczęściej wskaźnikami są odpowiedzi udzielone na zadania testowe oraz odpowiedzi na pytania kwestionariuszowe, rzadziej zaobserwowane cechy respondentów. Modele IRT wiążą cechę ukrytą ze wskaźnikami dzięki zastosowaniu parametryzacji, która określa właściwości wskaźników i rozkład cech respondentów. W tym rozdziale przedstawimy ogólny opis jednowymiarowego modelu IRT, przybliżymy modele najczęściej stosowane dla wskaźników dychotomicznych (2PLM, 3PLM, 1PLM) oraz porządkowych (GRM, GPCM). Omówimy model wielogrupowy i zarysujemy problematykę estymacji poziomu umiejętności. Opisując model będziemy posługiwać się terminami zaczerpniętymi z pomiaru edukacyjnego – dziedziny, w której IRT jest najbardziej popularne, i w obrębie której zostało stworzone. Badana cecha ukryta nazywana będzie „poziomem umiejętności" a poszczególne wskaźniki –„zadaniami". Są to terminy utrwalone w literaturze odnoszącej się do IRT, lecz w zależności od zastosowania modeli, można je zastępować bardziej dogodnymi, na przykład umiejętność zastąpić poziomem cechy a zadanie można odczytywać bardziej ogólnie – jako pozycję testową, pytanie lub stwierdzenie.

1.1. Statystyczna charakterystyka modeli IRT

1.1.1. Jednowymiarowy model IRT w ujęciu ogólnym

Model IRT opisuje rozkład prawdopodobieństwa wektora odpowiedzi na zadania $\mathbf{U} = (U_1, U_2, \ldots, U_n)$ dla jednostki obserwacji, którą wylosowano z pewnej populacji k. W najogólniejszej postaci jednowymiarowy model IRT można przedstawić jako:

$$P(\mathbf{U} = \mathbf{u} \mid \mathrm{k}) = \int f(\mathbf{u}, \theta, \boldsymbol{\beta}) \psi_k(\theta) d\theta \,, \tag{1.1}$$

gdzie: θ jest losową zmienną ukrytą opisującą poziom mierzonej umiejętności (lub innej cechy); $\psi_k(\theta)$ jest funkcją gęstości prawdopodobieństwa określającą rozkład zmiennej θ w populacji k; $f(\mathbf{u}, \theta, \boldsymbol{\beta})$ jest funkcją, która określa prawdopodobieństwo zaobserwowania konkretnej

[1] W rozdziale wykorzystano fragmenty artykułu: Kondratek, B. i Pokropek, A. (2013). IRT i pomiar edukacyjny. *Edukacja*, *124*(4), 42–66.

wartości **u** wektora odpowiedzi **U**, w zależności od poziomu umiejętności θ oraz wektora parametrów $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$, gdzie parametry zadania $\boldsymbol{\beta}_i$ również mogą być wektorami (np. dla dwuparametrycznego modelu logistycznego $\boldsymbol{\beta}_i = (a_i, b_i)$).

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

$$f(\mathbf{u}, \theta, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(u_i, \theta, \boldsymbol{\beta}_i) . \tag{1.2}$$

Założenie (1.2) nosi nazwę lokalnej niezależności i ma bardzo istotne, techniczne znaczenie dla szacowania parametrów modelu. Samo w sobie stanowi również ważną teoretyczną przesłankę dotyczącą testu złożonego z zadań *i*. Mianowicie (1.2) stanowi, że, gdy poziom umiejętności θ jest znany, to odpowiedzi na zadania testu są względem siebie statystycznie niezależne – poziom umiejętności θ wystarcza do wyjaśnienia wszystkich obserwowalnych współzależności między zadaniami. Tym samym z założenia o lokalnej niezależności (1.2) wynika założenie o jednowymiarowym charakterze testu, którego odpowiedzi są modelowane za pomocą IRT. Zarówno model (1.1), jak i założenie (1.2) można uogólnić do postaci wielowymiarowego poziomu umiejętności[2].

Z wzoru (1.1) wynika, że parametry modelu IRT to zestawy: parametrów zadań $\boldsymbol{\beta}_i$ oraz parametrów określających rozkład umiejętności ψ_k. Zazwyczaj przyjmuje się, że $\psi_k = N(\mu_k, \sigma_k^2)$, czyli że rozkład umiejętności jest określony przez rozkład normalny o średniej μ_k oraz wariancji σ_k^2. Oszacowanie wartości parametrów $\boldsymbol{\beta}_i$ oraz parametrów rozkładu umiejętności na podstawie zebranych danych nosi nazwę kalibracji testu.

1.1.2. Podstawowe modele IRT

Różnica między jednowymiarowymi modelami IRT sprowadza się do postaci funkcji pojawiających się w równaniu (1.2), które określają prawdopodobieństwa uzyskania poszczególnych kategorii 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), Fumiko Samejimy (1969) oraz Eijiego Murakiego (1992). Wszystkie przedstawione modele będą się odwoływały do funkcji logistycznej. Dostępne są również ich wersje opierające się na krzywej skumulowanego rozkładu normalnego, które są historycznie i teoretycznie pierwotne względem rozwiązań opartych na funkcji logistycznej (zob. Lord i Novick, 1968). Ze względu na bardzo przyjazne matematyczne właściwości funkcji logistycznej modele *normal ogive* (NOM) zostały w dużej mierze wyparte z praktycznych zastosowań modeli jednowymiarowych (stosowane są za to w modelach wie-

[2] Szczegółowe omówienie wielowymiarowych modeli IRT można znaleźć w publikacji Marka Reckase'a (2009).

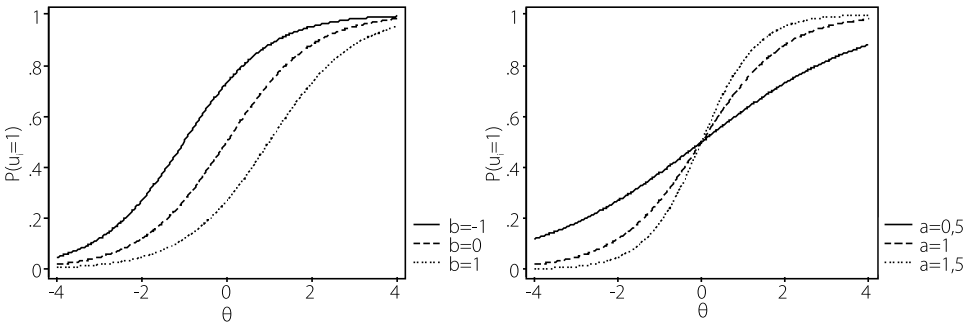
lowymiarowych) przez dające bardzo zbliżone wyniki modele logistyczne i nie zostaną w tym rozdziale opisane. Relacja między modelami IRT opartymi na funkcji logistycznej a modelami opartymi na krzywej skumulowanego rozkładu normalnego może być sprowadzona do relacji między logitową a probitową funkcją wiążącą w uogólnionych modelach liniowych lub nieliniowych, ponieważ modele IRT stanowią ich szczególny przypadek (Boeck i Wilson, 2004).

1.1.2.1. Modele dla zadań ocenianych dychotomicznie

W modelu dwuparametrycznym (*two-parameter logistic model*, 2PLM) 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:

$$P(u_i = 1 \mid \theta, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}. \tag{1.3}$$

Parametr a_i nosi nazwę parametru dyskryminacji (*item discrimination*), natomiast parametr b_i to parametr trudności (*item difficulty*)[3]. W IRT wykres funkcji określającej prawdopodobieństwo udzielenia odpowiedzi ocenianej na daną liczbę punktów w zależności od poziomu umiejętności ucznia nosi nazwę krzywej charakterystycznej zadania (*item characteristic curve*, ICC). Zależność między wartościami parametrów modelu dwuparametrycznego a kształtem krzywej charakterystycznej zadania modelującej prawdopodobieństwo udzielenia poprawnej odpowiedzi zilustrowano na Rysunku 1.1.



Rysunek 1.1. Przykładowe krzywe charakterystyczne w 2PLM; z lewej strony zróżnicowany parametr trudności (gdy parametr dyskryminacji a_i, został ustalony na poziomie 1), z prawej zróżnicowany parametr dyskryminacji (gdy parametr trudności b_i, został ustalony na poziomie 0).

[3] W klasycznej teorii testu (KTT) trudnością zadania bywa nazywany odsetek poprawnych odpowiedzi udzielonych na zadanie (lub w przypadku zadań ocenianych na szerszej skali niż zero-jedynkowa – średni wynik w zadaniu podzielony przez maksymalną liczbę punktów możliwych do zdobycia w tym zadaniu). Należy zauważyć, że taki wskaźnik informuje nie tyle o trudności zadania, ile o jego łatwości, ponieważ wzrost jego wartości oznacza, że zadanie jest coraz częściej rozwiązywane poprawnie przez badane osoby. W dalszej części książki, jeżeli będzie mowa o tym klasycznym wskaźniku, to będzie on nazywany łatwością zadania, zgodnie z jego faktyczną interpretacją, co pozwoli odróżnić go od parametru trudności w IRT.

Na Rysunku 1.1 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 wzięła się nazwa tego parametru. W 2PLM parametr trudności wyznacza na skali umiejętności punkt $\theta = b_i$, w którym prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie wynosi 0,5. Parametr trudności można zatem w 2PLM bezpośrednio odnieść do skali wyznaczonej przez zmienną umiejętności θ. Dodatkowo w tym modelu $\theta = b_i$ jest punktem przegięcia krzywej charakterystycznej i wskazuje na otoczenie punktu θ, w którym krzywa jest najbardziej stroma.

Parametr dyskryminacji w 2PLM określa natomiast wspomnianą stromość krzywej charakterystycznej. Na Rysunku 1.1 obserwujemy, że im wyższa 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 krzywa ICC jest stroma w pewnym punkcie θ, tym większa zdolność zadania do rozróżniania poziomu umiejętności uczniów znajdujących się po dwóch stronach 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 dwuparametryczny przedstawiony wzorem (1.3) do postaci, w której wartość dyskryminacji wszystkich zadań byłaby równa jedności, to powstałby model jednoparametryczny (*one-parameter logistic model*, 1PLM), nazwany na cześć duńskiego matematyka Georga Rascha jego nazwiskiem. Krzywe w modelu Rascha są względem siebie równoległe, tak jak na wykresie z lewej strony Rysunku 1.1. 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. Jest jedynym z omawianych modeli, w którym wynik sumaryczny w teście jest statystyką dostateczną dla oszacowania poziomu umiejętności ucznia (Wright i Stone, 1979). W praktyce umożliwia to na przykład prostą 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 przedstawionego na Rysunku 1.1 przykładu z trzema zadaniami o różnej mocy dyskryminacyjnej, zauważamy, że dla uczniów, których odpowiedzi osiągają wartość $\theta = 0$, zadania mają taką samą trudność. Jednak dla uczniów najsłabszych, zadanie o najniższej dyskryminacji jest najłatwiejsze, zadanie o dyskryminacji równej 1 jest od niego trudniejsze, a zadanie o najwyższej mocy dyskryminacyjnej jest najtrudniejsze. Gdy popatrzymy na uczniów o poziomie umiejętności większym niż 0, porządek trudności zadań odwraca się: najłatwiejszym jest zadanie najbardziej dyskryminacyjne a najtrudniejszym – zadanie najmniej dyskryminacyjne. Opisana interakcja relatywnej (względem innych zadań) trudności zadania z poziomem umiejętności budzi pewne zastrzeżenia zwolenników modelu Rascha. W sposób

intuicyjny ilustruje również, dlaczego sumaryczny wynik uzyskany 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, że miary umiejętności ucznia konstruowane za pomocą tego modelu lokują wyniki na skali przedziałowej, podczas gdy dla modeli o większej liczbie parametrów nie jest to możliwe (DeMars, 2010; Wright, 1983). Stosunek „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 rozdziału. Warto w tym punkcie zaznaczyć, że teza mówiąca o tym, że model Rascha umożliwia pomiary na skali przedziałowej w rozumieniu Stanleya Stevensa, nie jest ogólnie podzielana i wzbudza wiele kontrowersji od momentu jej sformułowania, aż po dzień dzisiejszy[4].

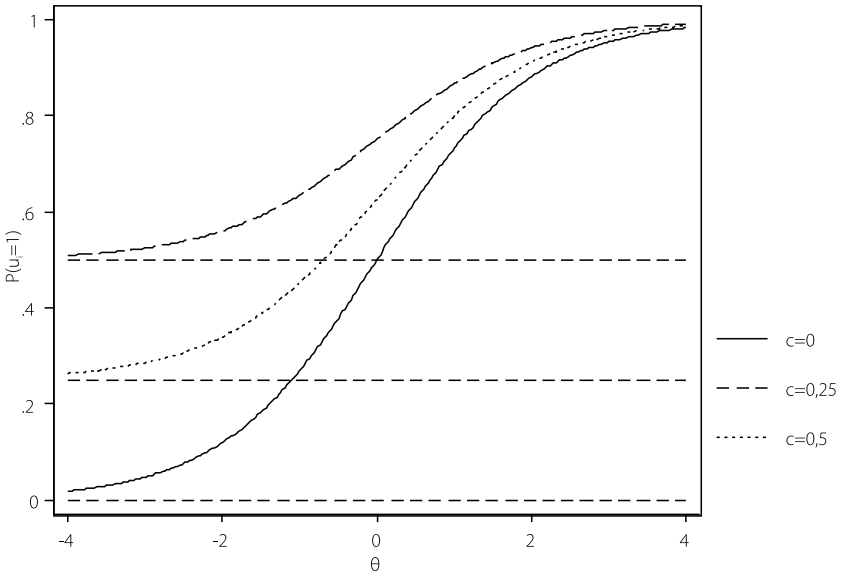
Trzyparametryczny model logistyczny (*three-parameter logistic model*, 3PLM) powstaje natomiast poprzez uogólnienie modelu 2PLM, wyrażonego wzorem (1.3), w taki sposób, aby dolna asymptota przypadała powyżej zera. Uzyskuje się to przez wprowadzenie dodatkowego parametru c_i w następujący sposób:

$$P(u_i = 1 \mid \theta, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}}. \tag{1.4}$$

Krzywą charakterystyczną w 3PLM można traktować jako średnią ważoną z dwóch prawdopodobieństw udzielenia poprawnej odpowiedzi: wynoszącego 1 na całym zakresie umiejętności θ (z wagą c_i) oraz zadanego zgodnie z modelem 2PLM (z wagą $(1 - c_i)$). W konsekwencji uzyskujemy krzywe, których dolna asymptota jest równa parametrowi c_i. Na Rysunku 1.2 widać również, że 2PLM można traktować 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 poprawnej odpowiedzi na dane zadanie. 3PLM często okazuje się przydatny do modelowania odpowiedzi na zadania wyboru, w których istnieje możliwość odgadnięcia poprawnej odpowiedzi. Z tego powodu parametr c_i bywa nazywany parametrem zgadywania (*guessing*). Jednak interpretacja odgadywania poprawnej odpowiedzi nie zawsze jest w pełni uzasadniona do wyjaśnienia konkretnego poziomu c_i, sugerowanego przez model. Dlatego c_i określa się również mianem parametru pseudozgadywania (*pseudo-guessing*).

[4] Przegląd krytycznej dyskusji na temat przedziałowości skal powstałych w wyniku zastosowania modeli IRT, rozumianej 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). Omówienie modeli Rascha można znaleźć również w rozdziale 2.



Rysunek 1.2. Przykładowe krzywe charakterystyczne w modelu 3PLM; parametry dyskryminacji i trudności ustalone odpowiednio na wartościach: $a_i = 1$ oraz $b_i = 0$.

Krzywe przedstawione na Rysunku 1.2 pokazują, 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 mniej strome w każdym punkcie θ. Jednocześnie, z wprowadzeniem parametru c_i traci moc bezpośrednia interpretacja wartości parametrów a_i i b_i, jaka była możliwa w modelu 2PLM. Parametr b_i 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 nachylenie w punkcie $\theta = b_i$, przy zwiększającym się c_i, musi wzrosnąć wartość a_i. W związku z tym analizowanie właściwości zadania na podstawie parametrów modelu 3PLM staje się o wiele trudniejsze niż w przypadku 2PLM, ponieważ wszystkie trzy parametry: a_i, b_i i c_i trzeba rozpatrywać łącznie. O wiele łatwiej jest ocenić jakość zadania w 3PLM, analizując krzywą charakterystyczną i 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 modelu 2PLM).

1.1.2.2. Modele dla zadań ocenianych wielopunktowo

Aby przedstawić 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 odpo-

wiedzi ocenionej jako „0" można również wykreślić krzywą informującą o prawdopodobieństwie udzielenia tej odpowiedzi, jednak jest ona pomijana, ponieważ dla zadania ocenianego zero-jedynkowo jest redundantna: $P(u_i = 0) = 1 - P(u_i = 0)$. Inaczej jest w przypadku zadań ocenianych na szerszej niż dychotomiczna skali punktowej. Do opisu zadań ocenianych wielopunktowo konieczne jest przedstawienie krzywych opisujących prawdopodobieństwo udzielenia odpowiedzi ocenianej dla każdej z możliwych *m* kategorii oceny.

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

$$P_x(u_i \le x \mid \theta, a_i, b_{i,x}) = \frac{-1}{1+e^{-a_i(\theta - b_{i,x})}}. \tag{1.5}$$

Krzywe określone wzorem (1.5) 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 strony Rysunku 1.1). Następnie, dla wyznaczenia krzywej opisującej uzyskanie konkretnej wartości punktowej, oblicza się:

- dla kategorii 0 punktów: $P(u_i=0|\theta) = P_0(u_i \le 0|\theta)$;
- dla kategorii pośrednich: $x \in \{0, ..., m-1\}: P_x(u_i = x \mid \theta) = P_x(u_i \le x|\theta) - P_{x-1}(u_i \le x-1|\theta)$;
- dla kategorii *m* punktów: $P_x(u_i = m|\theta) = 1 - P_{m-1}(u_i \le m-1|\theta)$.

Dla zadania ocenianego na skali 0–*m* uzyskujemy zatem komplet $m+1$ krzywych, przy czym:

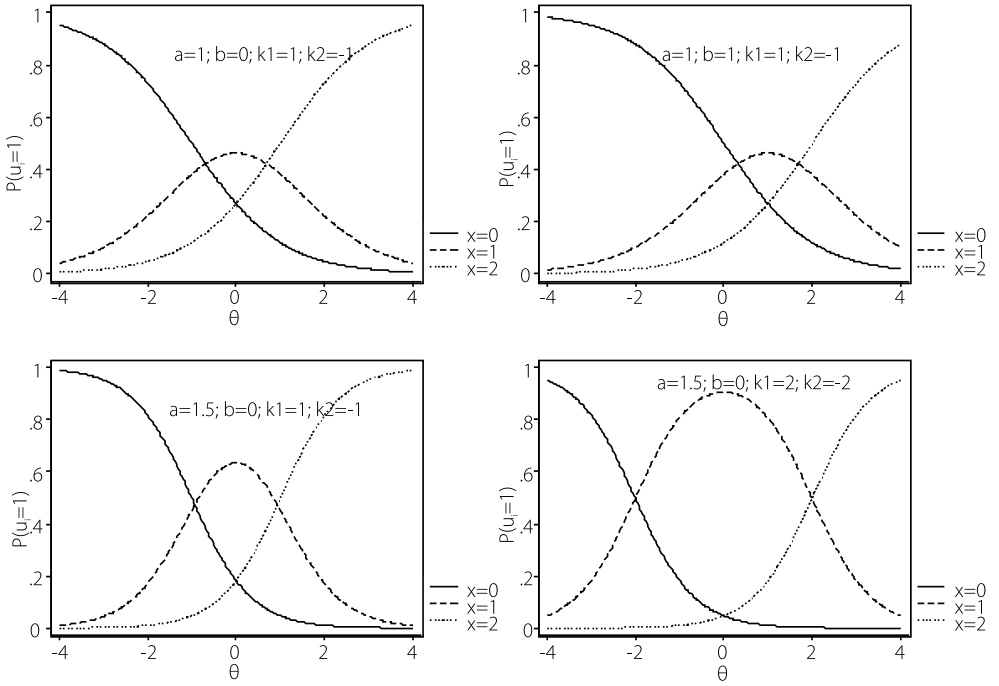
- pierwsza krzywa ma kształt krzywej logistycznej 2PLM z ujemnym parametrem dyskryminacji (funkcja malejąca) oraz z parametrem trudności $b_{i,0}$;
- krzywe dla kategorii pośrednich *x* mają kształt dzwonowaty, przy czym dla wyższych kategorii punktowych maksimum funkcji przypada bardziej na prawo niż dla niższych kategorii punktowych, a 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 $b_{i,x-1}$.

W ramach raportowania parametrów 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, oraz parametry kategorii, będące odchyleniami od b_i: $k_{i,x+1} = b_i - b_{i,x}$.

Przykład czterech krzywych dla zadania ocenianego na skali 0–2 znajduje się na Rysunku 1.3. W lewym górnym rogu podano krzywe dla zadania z dyskryminacją równą $a_i = 1$, trudnością $b_i = 0$ i parametrami kategorii ±1. W prawym górnym rogu mamy zadanie, w którym zmieniono jedynie wartość parametru b_i – jego zmiana powoduje takie samo przesunięcie krzywych, jakie można zaobserwować w 2PLM. W lewym dolnym rogu w wyjściowych parametrach zwiększono jedynie parametr dyskryminacji a_i, co spowodowało zwiększenie stromości krzy-

wych dla skrajnych kategorii oraz zagęszczenie prawdopodobieństwa uzyskania pośredniej kategorii punktowej. Zatem zadanie po zwiększeniu a_i stało się bardziej dyskryminacyjne – uzyskanie przez ucznia każdej kategorii punktowej niesie ze sobą bardziej precyzyjną informację o jego poziomie umiejętności. W prawym dolnym rogu, oprócz zwiększenia mocy dyskryminacyjnej zwiększono odchylenia od parametru b_i do ±2. Wskutek tego najczęściej uzyskiwaną kategorią punktową przez uczniów w dość szerokim zakresie umiejętności (od –2 do 2) będzie jeden punkt.

Obok opisanego 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. PCM został dalej zmodyfikowany przez Murakiego (1992) tak, aby dopuścić zróżnicowany parametr dyskryminacji. Powstały w efekcie uogólniony model odpowiedzi częściowej (GPCM, *generalized PCM*) stał się bardziej elastyczny, jednak wykroczył poza konserwatywne ramy modelu Rascha.



Rysunek 1.3. Przykład krzywych charakterystycznych w modelu GRM dla czterech zadań ocenianych na skali 0–2.

Dla zadania ocenianego na skali 0–*m* w modelu GPCM prawdopodobieństwo zaobserwowania odpowiedzi dla kategorii punktowej x ∈ {0,...,*m*} jest określone funkcją:

$$P(u_i = x \mid \theta, a_i, b_{i,\cdot}) = \frac{e^{a_i\left(x\theta - \sum_{r=0}^{x} b_{i,r}\right)}}{\sum_{c=0}^{m} e^{a_i\left(c\theta - \sum_{r=0}^{c} b_{i,r}\right)}}, \tag{1.6}$$

przy czym parametr $b_{i,0} = 0$. W GPCM mamy zatem, podobnie jak w modelu GRM, pojedynczy parametr dyskryminacji zadania (a_i) oraz *m* podlegających estymacji parametrów $b_{i,m}$, skojarzonych z trudnością poszczególnych kategorii.

Istotną, z praktycznego punktu widzenia, różnicą między omawianymi dwoma modelami dla wielopunktowych odpowiedzi, jest brak konieczności monotonicznego wzrastania wartości parametru $b_{i,m}$ wraz ze wzrostem kategorii *m* w GPCM. W modelu Samejimy kolejne wartości $b_{i,m}$ są matematycznie ograniczone przed przyjmowaniem niższych wartości niż parametry dla wcześniejszych kategorii. W GPCM takiego ograniczenia nie ma, co wprowadza elastyczność, która w niektórych sytuacjach może skutkować lepszym dopasowaniem do danych niż w przypadku GRM. Brak wspomnianych ograniczeń wartości $b_{i,m}$ w GPCM również ułatwia interpretację błędów standardowych dla oszacowań tego parametru[5].

1.1.2.3. Modele wielogrupowe

Przy założeniu warunkowej niezależności odpowiedzi uczniów modele IRT umożliwiają warunkowe określenie prawdopodobieństwa uzyskania wektora odpowiedzi $\mathbf{U} = (U_1, U_2, \ldots, U_n)$ ze względu na poziom umiejętności θ i parametry zadań, za pomocą iloczynu funkcji charakterystycznych dla określonej liczby zadań, opisanych w równaniach (1.1) i (1.2).

Poprzestając na tej formule, łatwo można zapomnieć, że w prostym modelu IRT zakładamy, że jednostki pochodzą z jednej grupy (populacji), którą charakteryzuje funkcja gęstości prawdopodobieństwa określająca rozkład θ dla danej próby. Aby rozszerzyć prosty model IRT o możliwość szacowania parametrów dla kilku grup jednocześnie, należy wprowadzić wektor parametrów charakteryzujących rozkład θ w *K* różnych grupach $\eta = (\eta_1, \ldots, \eta_k, \ldots \eta_K)$ (najczęściej $\boldsymbol{\eta}_k$ jest średnią oraz odchyleniem standardowym rozkładu normalnego). Model wielogrupowy powstaje poprzez włączenie do modelu opisanego wzorem (1.1) wspomnianych parametrów rozkładu umiejętności, zależnych od przynależności grupowej:

$$P(\mathrm{U} \mid \beta, \eta_k) = P(\mathrm{U} \mid \beta, \theta)\, \Psi(\theta \mid \eta_k) d\theta. \tag{1.7}$$

Jest on rozszerzeniem klasycznego modelu IRT, w literaturze jest nazywany wielogrupowym modelem IRT (*multiple group IRT*, MG-IRT; Bock i Zimowski, 1997). Oprócz szacowania

[5] Szczegółowe porównanie modeli dla zadań ocenianych wielokategorialnie można znaleźć u Davida Thissena i Lynne Steinberg (1986).

parametrów zadań, tak jak w prostych modelach IRT, umożliwia on szacowanie parametrów rozkładów θ w różnych grupach, dając jednocześnie nieobciążone estymacje θ dla poszczególnych jednostek. W równaniu (1.7) przedstawiamy ogólny wzór na model wielogrupowy. Modele wielogrupowe charakteryzowane mogą być oczywiście zarówno przez jedno- dwuczy trzyparametryczne modele odpowiedzi. Mogą być również stosowane dla zmiennych dychotomicznych lub wielokategorialnych.

Z modelem wielogrupowym wiążą się jednak trudne kwestie natury technicznej. Do jego oszacowania (tak jak i innych modeli IRT) niezbędna jest procedura estymacji największej wiarygodności, jednak – w przeciwieństwie do prostego modelu IRT – w modelu wielogrupowym funkcja wiarygodności (a precyzyjniej: całka stanowiąca jej część) nie ma analitycznego rozwiązania (szczegóły w publikacjach: Baker i Kim, 2004; Bock i Zimowski, 1997). Estymacja jest zatem trudna i wymaga wprowadzenia procedur całkowania numerycznego, które jest dostępne tylko w niektórych programach służących do modelowania IRT.

Stosowanie modelu wielogrupowego jest konieczne w sytuacjach, gdy uczniowie pochodzą z różnych prób o różnych rozkładach poziomu umiejętności. Wykorzystanie prostego modelu IRT w takich wypadkach może prowadzić do znacznych błędów w szacowaniu parametrów grupowych. Modele wielogrupowe stosowane są także w analizach zrównujących. Zakłada się w nich brak równoważności grup, wykorzystuje test kotwiczący (złożony z zadań wspólnych dla różnych sesji testu lub egzaminu), a estymacja jest przeprowadzana z wykorzystaniem kalibracji łącznej – tak jak w przypadku zrównywania polskich egzaminów (Szaleniec, Grudniewska, Kondratek, Kulon i Pokropek, 2012; zob. też rozdział 16). Innym przypadkiem, w którym konieczne jest wykorzystanie wielogrupowego modelowania IRT, są analizy zróżnicowanego funkcjonowania pozycji testowej (*differential item functioning*, DIF), podczas których sprawdza się czy zadania funkcjonują w taki sam sposób między wybranymi grupami (rozdział 3).

1.2. Wybór modelu i liczebność próby kalibracyjnej

Istotnym czynnikiem, jaki należy rozważyć przy podejmowaniu decyzji o wyborze modelu: 1PLM, 2PLM czy 3PLM, jest liczebność dostępnej próby, na której będzie przeprowadzana kalibracja testu. W warunkach nieograniczonych liczebnością badanej próby model trzyparametryczny będzie w zdecydowanej większości przypadków najlepszym rozwiązaniem. Model posiadający największą liczbę parametrów będzie gwarantował najlepsze dopasowanie do danych, a w związku z tym, najwyższą precyzję pomiaru. Jednak w rzeczywistości badacz rzadko dysponuje nieograniczoną możliwością wyboru liczebności grupy. Im mniejsza próba, tym oszacowania parametrów zadań są mniej dokładne, co w konsekwencji pogarsza oszacowanie poziomu umiejętności uczniów. Frederic Lord (1980) wskazywał na to, że bardziej precyzyjne pomiary dla mało licznych prób uzyskuje się za pomocą modelu jednoparametrycznego, a nie modeli bardziej złożonych, nawet gdy proces odpowiedzi na zadania wyraźnie odzwierciedla strukturę dwu- lub trzyparametryczną. Obciążone wyniki estymacji parametru dyskryminacji lub pseudozgadywania stanowią bowiem większy problem w kontekście sza-

cowania poziomu umiejętności ucznia, niż błędy spowodowane niedopasowaniem zadań do modelu IRT.

Liczne badania symulacyjne pokazują, że do szacowania parametrów z zadowalającą dokładnością na prostszych modelach IRT (tj. modelach z mniejszą liczbą parametrów) wystarczą małe próby. W przypadku testów złożonych z zadań dychotomicznych model jednoparametryczny pozwala szacować z dobrą dokładnością na próbach złożonych ze 100–200 uczniów (Ayala, 2009). Model dwuparametryczny jest nieco bardziej wymagający, a próba 500 uczniów wydaje się dla niego bardziej odpowiednia (Stone, 1992). Według niektórych autorów model trzyparametryczny zachowuje dostateczną precyzję szacowania parametrów przy 1000 uczniów, ale z zastrzeżeniem, że zadania mają wysoką moc dyskryminacyjną a grupa, dla której przeprowadzana jest estymacja, jest stosunkowo silnie zróżnicowana pod względem umiejętności (DeMars, 2010). W innych sytuacjach liczebność próby, niezbędna do oszacowania modelu trzyparametrycznego, powinna przekraczać 2000 uczniów (Woods, 2008).

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

Oczywiście są to tylko wartości orientacyjne. Liczebność próby zależy od długości testu, jakości zadań, rozkładu umiejętności uczniów w badanej populacji, sposobu doboru próby, 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 oraz do prób uczniów losowanych z populacji o rozkładzie umiejętności zbliżonym do normalnego.

W przypadkach niestandardowych (np. ze skomplikowanym schematem doboru próby lub skomplikowanym schematem przydziału zadań do uczniów) badacz może przeprowadzić na własną rękę odpowiedni eksperyment Monte Carlo, aby oszacować jakość wnioskowania statystycznego przy założonych warunkach. Do przeprowadzenia takich symulacji jest jednak konieczne przyjęcie pewnych założeń dotyczących spodziewanego zakresu wartości przyjmowanego przez parametry zadań oraz dotyczących rozkładu poziomu umiejętności uczniów w próbie.

1.3. Szacowanie poziomu umiejętności, funkcja informacji

Częstokroć zachodzi potrzeba, aby oprócz parametrów określających jednowymiarowy model IRT (1.1), czyli parametrów zadań oraz parametrów rozkładu poziomu umiejętności w całej populacji, oszacować również poziom umiejętności, θ, pojedynczych uczniów, na

podstawie zaobserwowanych dla nich wektorów odpowiedzi **U = u**. Dokonuje się tego poprzez odwołanie do twierdzenia Bayesa:

$$P(parametr \mid dane) = \frac{P(dane \mid parametr)P(parametr)}{\mathrm{P}(dane)}.$$

Uzyskawszy oszacowania parametrów modelu (1.1) w wyniku kalibracji, można twierdzenie Bayesa zastosować do przedstawienia rozkładu a posteriori parametru θ, pod warunkiem zaobserwowania wektora odpowiedzi **u** w następujący sposób:

$$P(\theta \mid \mathbf{U} = \mathbf{u}) = \frac{f(u,\theta,\beta)\psi_k(\theta)}{\int f(u,\theta,\beta)\psi_k(\theta)d\theta}. \tag{1.8}$$

Wzór (1.8) określa zatem nie tyle punktowe oszacowanie poziomu umiejętności ucznia, ile cały rozkład prawdopodobieństwa dla poziomu umiejętności ucznia. Dzięki temu dostarcza również informacji o niepewności, z jaką dana umiejętność została oszacowana. W przypadku konieczności uzyskania punktowego oszacowania poziomu umiejętności ucznia mamy do dyspozycji dwie podstawowe alternatywy:

- Estymator EAP (*expected a posteriori*), będący wartością oczekiwaną dla rozkładu (1.8) – rozwiązanie to wymaga całkowania numerycznego całego (1.8) po rozkładzie $\psi_K(\theta)$;
- Estymator MAP (*maximum a posteriori*), będący maksimum funkcji gęstości rozkładu (1.8) – to rozwiązanie jest nieco prostsze: sprowadza się do znalezienia maksimum funkcji $f(u, \theta, \beta)\, \psi_P(\theta)$.

We wczesnych zastosowaniach IRT korzystano także ze zwykłego estymatora największej wiarygodności umiejętności ucznia (*maximum likelihood estimator*, ML), który nie uwzględniał rozkładu umiejętności w populacji. Znalezienie takiego estymatora dla ucznia o wektorze odpowiedzi **U = u** sprowadza się do znalezienia maksimum funkcji wiarygodności $f(\mathbf{u},\theta,\beta)$. Rozwiązanie to jednak ma dużo wad. Przede wszystkim nie istnieją punktowe estymatory umiejętności dla uczniów o najniższym i najwyższym możliwym do uzyskania wyniku, ponieważ funkcja $f(\mathbf{u},\theta,\beta)$ w takich przypadkach nie osiąga maksimum. Ten przykład pokazuje również rolę, jaką odgrywa rozkład poziomu umiejętności θ w całej populacji ψ_K podczas estymacji poziomu umiejętności pojedynczego ucznia. Funkcjonuje on w pewnym sensie jako dodatkowe zadanie i jego wkład jest tym większy, im mniej informacji dostarczają odpowiedzi udzielone na zadania testu. Szczególnie gdy 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 (1.8) i będą tym większe, im bardziej będzie on niesymetryczny. Ogólnie 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 ważonej największej wiarygodności (*weighted*

maximum likelihood, WML) stanowi modyfikację ML i również nie odwołuje się do rozkładu a posteriori (1.8), czyli nie uwzględnia zakładanej informacji o rozkładzie umiejętności w populacji ψ_K. Estymator WML 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 (1.8), które noszą nazwę „wartości potencjalnych" (*plausible values*, PV). Przeprowadzanie wtórnych analiz na PV, zamiast na punktowych oszacowaniach poziomu umiejętności, pozwala uwzględnić błąd pomiaru, związany z nierzetelnością standardowych narzędzi statystycznych badających umiejętności. Nieuwzględnienie błędu pomiaru w takich analizach (np. przez przeprowadzanie ich na punktowych oszacowaniach umiejętności) prowadzi do obciążenia badanych statystyk oraz ich błędów standardowych[6].

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 (1.8), co wymaga przeprowadzenia ponownie całkowania numerycznego. Dla estymatora MAP natomiast błąd standardowy oblicza się przez odwołanie do koncepcji informacji Fishera. Jeżeli za $L(\theta)$ oznaczymy logarytm z funkcji wiarygodności (w przypadku (1.8) jest to Bayesowska funkcja wiarygodności: $f(\mathbf{u},\theta,\beta)\psi_K(\theta)$, to informacja Fishera jest dana wzorem (Lehmann, 1991):

$$I(\theta) = E\left(\left(\frac{dL(\theta)}{d\theta}\right)^2 \middle| \theta\right). \tag{1.9}$$

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

$$SE(\hat{\theta}) \approx \frac{1}{\sqrt{I(\hat{\theta})}}. \tag{1.10}$$

Informacja Fishera posiada kilka właściwości, bardzo cennych w kontekście IRT. Po pierwsze, błąd oszacowany na jej podstawie nie będzie zależał od konkretnych odpowiedzi $U = u$ udzielonych przez ucznia. Po drugie, informacja Fishera jest addytywna (Rao, 1982), co pozwala na rozbicie jej na sumę wkładów informacji poszczególnych zadań testu. W związku z tym

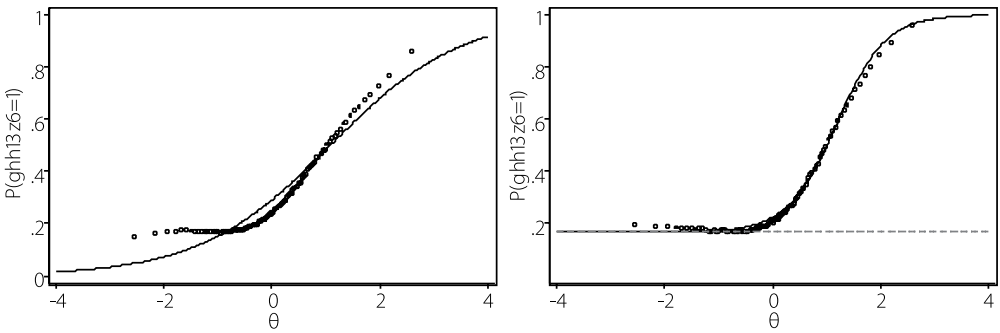
[6] Więcej o roli PV w badaniach edukacyjnych można znaleźć w publikacji Margaret Wu (2005), w technicznych raportach z międzynarodowych badań, np. PISA (OECD, 2009), a także w rozdziale 11.

$I(\theta)$ stanowi miarę lokalnej precyzji pomiaru umiejętności zarówno dla całego testu, jak i dla poszczególnych zadań, i jest doskonałą alternatywą do klasycznych miar, takich jak współczynnik rzetelności dla całego testu, a na poziomie pojedynczego zadania – współczynnik dyskryminacji. Koncepcja funkcji informacji testu oraz zadania ma bezpośrednie zastosowanie w konstrukcji narzędzi testowych i w testowaniu adaptatywnym.

1.4. Dopasowanie modelu do danych

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 dokonać na różne sposoby. Na przykład dwa modele IRT oszacowane dla tych samych danych można porównywać ze sobą w całości za pomocą testu ilorazu funkcji wiarygodności, aby 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 wzrokowej oceny wykresów, na których – obok krzywych charakterystycznych zadań – naniesiono empiryczne wyniki uzyskane przez uczniów, a także przez odwołanie się do odpowiednich statystyk dopasowania.



Rysunek 1.4. Przykład graficznej analizy dopasowania modelu IRT do danych; punkty odpowiadają empirycznym proporcjom odpowiedzi poprawnej w poszczególnych centylach umiejętności.

Przykład metody graficznej zilustrowano na Rysunku 1.4. Przedstawia on dwie krzywe charakterystyczne dla tego samego zadania z egzaminu gimnazjalnego z 2011 roku. Wykres z lewej strony uzyskano dla pierwszej próby dopasowania modelu IRT do danych, w której wszystkie zadania oceniane dychotomicznie modelowano za pomocą 2PLM. Widać systematyczne różni-

ce w przebiegu krzywej wyznaczonej przez dopasowany model a „krzywej" sugerowanej przez empiryczne proporcje odpowiedzi poprawnych. Układanie się punktów empirycznych powyżej zera dla uczniów o niskim poziomie umiejętności oraz większa moc dyskryminacyjna zadania w rejonie łatwości 0,5 dla punktów empirycznych, w porównaniu do krzywej 2PLM, sugeruje badaczowi, że zadanie to lepiej byłoby opisane modelem trzyparametrycznym. W istocie, dla tego samego zadania (po prawej stronie) uzyskano praktycznie idealne dopasowanie, gdy w ostatecznie użytym dla egzaminu gimnazjalnego modelu IRT było ono modelowane za pomocą 3PLM.

Warto tutaj zauważyć, że niedopasowanie jednego z zadań całego testu przekłada się na trafność oszacowania rozkładu poziomu umiejętności uczniów i wtórnie może się przełożyć na niedopasowanie pozostałych zadań. Jest to w pewnym sensie sytuacja naczyń połączonych. Tak dobre dopasowanie zadania do 3PLM na Rysunku 1.4 jest nie tylko konsekwencją zmiany modelu dla tego zadania, lecz także zmian modelu dla wybranych innych zadań tego testu, jakiej dokonano między pierwszą próbą kalibracji testu (z lewej strony) 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.

Wzrokowa ocena „regularności" odchylania się empirycznych wyników od przewidywań modelu może w pewnych przypadkach przysparzać większych kłopotów, zwłaszcza w przypadku mniejszych prób. W pewnym momencie granica między tym, co systematyczne a tym, co losowe zaciera się i staje się trudna do oceny. Obiektywnej kwantyfikacji stopnia dopasowania dostarczają podstawowe statystyki. Podejście to jest bardzo proste – polega na ocenie rozkładu odchyleń wyników rzeczywistych od wyników szacowanych przez model w grupach arbitralnie wyznaczonych na podstawie umiejętności uczniów.

Podstawową i najprostszą statystyką tego rodzaju jest chi-kwadrat, dla IRT opisane przez R. Darrella Bocka (1972):

$$\chi^2 = \sum_{g=1}^{G} \frac{N_g (O_{ig} - E_{ig})^2}{E_{ig}(1 - E_{ig})}. \tag{1.11}$$

gdzie: N_g oznacza 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.

Aby obliczyć tę statystykę, należy podzielić grupę badanych obserwacji na G grup (zazwyczaj 10)[7] na podstawie szacowanych umiejętności. Im wyższa wartość tej statystyki, tym zadanie jest gorzej dopasowane (ma ona rozkład chi-kwadrat o liczbie stopni swobody równej liczbie grup, pomniejszonej o liczbę parametrów estymowanych dla zadania).

[7] Podobnie jak w podejściu Hosmera i Lemeshowa (2000) do testowania dopasowania modelu regresji logistycznej, w którym wyróżniane są decyle ryzyka.

Kolejną statystyką stosowaną w popularnych programach do estymacji modeli IRT (np. Bilog-MG i Parscale) jest G^2 (Muraki i Bock, 2003). Odpowiada ona statystykom opartym na ilorazie wiarygodności:

$$G^2 = 2\sum_{g=1}^{G}\left[r_{ig} \log \frac{r_{ig}}{N_g P_i(\bar{\theta}_g)} + (N_g - r_{ig}) \log \frac{N_g - r_{gi}}{N_g \left(1 - P_i(\bar{\theta}_g)\right)} \right]. \qquad (1.12)$$

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 wyższa wartość G^2, tym zadanie gorzej dopasowane. G^2 ma rozkład chi-kwadrat o liczbie stopni swobody równej liczbie grup bez korekty liczby szacowanych parametrów (szczegółowy opis w: Muraki i Bock, 2003).

Stosując testy statystyczne do oceny dopasowania zadania, należy pamiętać, że wraz ze wzrostem liczebności próby zwiększa się moc każdego testu statystycznego. Mając na względzie fakt, że modelowanie IRT często jest stosowane na dużych próbach uczniów, trzeba się liczyć z tym, że nawet niewielkie – z praktycznego punktu widzenia – odstępstwa między przewidywaniem modelu a obserwowalnymi wynikami, mogą skutkować pozytywnym wynikiem dopasowania testu. W odpowiednio dużej próbie okaże się, że każde analizowane zadanie statystycznie istotnie odbiega od krzywej oszacowanej przez model IRT. Dlatego oprócz obliczania istotności statystycznej, należy także analizować uzyskaną wartość statystyki dopasowania i wspomagać się graficzną analizą uzyskanych wyników.

1.5. Podsumowanie

Niniejszy rozdział miał na celu przybliżenie i uporządkowanie zagadnień związanych z tematem jednowymiarowych modeli IRT. Przybliżone zostały najczęściej stosowane modele dla zadań ocenianych dychotomicznie oraz wielopunktowo), wprowadzono podstawową definicje modelu wielogrupowego, a także zarysowano problematykę estymacji poziomu umiejętności. Jest to elementarz podstawowych pojęć i definicji. Jest to również punkt wyjścia do przedstawienia bardziej złożonych modeli oraz problemów opisywanych w kolejnych rozdziałach tej książki.

Literatura

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

Baker, F. B. i Kim, S. (2004). *Item response theory. Parameter estimation techniques*. New York: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models. W: F. M. Lord I M. R. Novick (red.), *Statistical theories of mental test scores* (397–479). 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.

Bock, R. D. i Zimowski, M. F. (1997). Multiple group IRT. W: W. J. van der Linden i R. K. Hambleton (red.), *Handbook of modern item response theory* (433–448). New York: Springer.

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

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 partialcredit model using MULTILOG. *Journal of Outcome Measurement*, *1*(2), 114–142.

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

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

Hosmer, D. W. i Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

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

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.

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

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.

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 report*. Paris: OECD Publishing.

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.

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), 105–119.

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.

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.