Bartosz Kondratek
Instytut Badań Edukacyjnych

Przegląd statystyk dopasowania modeli IRT na poziomie zadania

Wprowadzenie

Parametryczne modele teorii odpowiedzi na zadanie (*item response theory*, IRT) odgrywają coraz to większą rolę zarówno przy konstrukcji testów i kwestionariuszy, podczas samego badania (np. przy komputerowym adaptatywnym testowaniu, *computerized adaptive testing,* CAT), jak i przy interpretacji wyników badań. Modele IRT wypierają na wielu polach dotychczasowe rozwiązania oparte na klasycznej teorii testów (KTT), jednocześnie umożliwiając rozwój aplikacji wcześniej niedostępnych w ramach podejścia klasycznego.

Jakość wnioskowania statystycznego na temat mierzonej umiejętności z wykorzystaniem IRT jest jednak silnie związana z tym, na ile adekwatnie model IRT jest do danych dopasowany. Fakt ten znajduje swoje odzwierciedlenie w *Standardach dla testów stosowanych w psychologii i pedagogice* wydawanych przez AERA, APA oraz NCME (2014), gdzie w przypadku stosowania modeli IRT rekomenduje się przedstawienie dowodów odpowiedniego ich dopasowania do danych. Wielu badaczy wskazuje na konsekwencje, jakie niedopasowanie modelu IRT do danych ma dla późniejszego wnioskowania (Wainer i Thissen, 1987; Woods, 2008; Bolt, Deng i Lee, 2014).

Problematyka oceny dopasowania modeli IRT na poziomie zadań od samych narodzin podejścia stanowiła przedmiot intensywnych badań. Większość rozwiązań polega na przeprowadzeniu podziału badanej testem umiejętności na przedziały (grupy) i dokonania porównania przewidywanych przez model proporcji odpowiedzi do zaobserwowanych w danej grupie proporcji odpowiedzi. Różnice między metodami sprowadzają się do:

1. sposobu, w jaki dokonuje się podziału na grupy;
2. sposobu, w jaki liczone są obserwowane oraz przewidywane proporcje odpowiedzi poprawnych;
3. postaci statystyki testowej;
4. ujęcia w ramach podejścia częstościowego lub bayesowskiego.

W artykule zostanie przeprowadzony krytyczny przegląd wybranych rozwiązań dla zadań ocenianych dychotomicznie wraz z nakreśleniem autorskiej propozycji ujęcia problemu. Ze względu na oszczędność miejsca same modele IRT nie zostaną opisane, Czytelnik zainteresowany wprowadzeniem do IRT może skorzystać z opracowania Bartosza Kondratka i Artura Pokropka (2015).

Statystyki $S\text{-}X^2$ *oraz* $S\text{-}G^2$ Orlando i Thissena

Za punkt wyjścia przy dokonywaniu przeglądu statystyk dopasowania jednowymiarowych modeli IRT dla zadań ocenianych dychotomicznie przyjmiemy parę statystyk $S\text{-}X^2$ oraz $S\text{-}G^2$, zaproponowanych przez Marię Orlando i Davida Thissena (2000). Ich podejście zostało opracowane jako alternatywa dla wcześniej zaproponowanych rozwiązań i, jak wykazali w swojej pracy, charakteryzuje się od nich lepszymi właściwościami statystycznymi. Statystyki $S\text{-}X^2$ lub $S\text{-}G^2$ są jednocześnie najczęściej wybierane jako punkt odniesienia przy porównaniach z innymi współcześnie rozwijanymi podejściami (np. Stone i Zhang, 2003; Glas i Suarez-Falcon, 2003; Toribio i Albert, 2011). W związku z powyższym, zaburzając chronologię, zaczniemy właśnie od omówienia tej pary statystyk, dzięki czemu łatwiej będzie można zaznaczyć wady i zalety wszystkich innych omawianych dalej rozwiązań.

Statystyki Orlando i Thissena mają następującą postać:

$$S - \mathrm{X}^2 = \sum_{g=0}^{G} N_g \left(\frac{\left(O_{ig} - E_{ig}\right)^2}{E_{ig}(1 - E_{ig})} \right), \tag{1}$$

$$S - G^2 = 2 \sum_{g=0}^{G} N_g \left(O_{ig} \ln\left(\frac{O_{ig}}{E_{ig}}\right) + \left(1 - O_{ig}\right) \ln\left(\frac{1 - O_{ig}}{1 - E_{ig}}\right) \right). \tag{2}$$

We wzorach (1) oraz (2) N_g jest liczbą obserwacji w grupie g, natomiast E_{ig} oraz O_{ig} są odpowiednio przewidywaną przez model IRT oraz obserwowaną proporcją odpowiedzi poprawnych na zadanie i w grupie g. Charakterystyczne dla statystyk Orlando i Thissena jest przeprowadzanie grupowania badanych osób ze względu na ich wyniki obserwowane w teście (stąd indeksacja od g=0). Statystyka $S\text{-}X^2$ odwołuje się do testu χ^2 Pearsona, natomiast statystyka $S\text{-}G^2$ jest oparta na ilorazie wiarygodności. Przyjmuje się, że przy prawdziwości hipotezy zerowej rozkład obu statystyk jest przybliżony rozkładem χ^2 z $G\text{-}m$ stopniami swobody, gdzie m jest liczbą szacowanych parametrów dla funkcji charakterystycznej zadania i.

Przy grupowaniu ze względu na wyniki obserwowane w całym teście wartość O_{ig} jest wyliczana bezpośrednio z danych, natomiast uzyskanie oczekiwanej proporcji odpowiedzi poprawnych, E_{ig}, stanowi pewne wyzwanie, którego sprytne rozwiązanie stanowi istotny element omawianego podejścia. Parametry modelu IRT wprost określają prawdopodobieństwo zaobserwowania dowolnego wektora odpowiedzi, tj. model IRT określa funkcję wiarygodności dla każdej unikalnej konfiguracji odpowiedzi na wszystkie zadania testu. Do wyliczenia E_{ig} potrzebna jest natomiast funkcja wiarygodności dla konkretnej sumy poprawnych odpowiedzi na wszystkie zadania testu, która to suma może powstać poprzez realizację różnych wektorów odpowiedzi. Rekursywny algorytm pozwalający na skonstruowanie funkcji wiarygodności dla dowolnej sumy punktów w teście został zaproponowany przez Fredericka Lorda i Marilyn Wingersky (1984).

Taką funkcję wiarygodności określającą prawdopodobieństwo uzyskania wyniku k oznacza się jako as $S_k(\theta)$, stąd też prefiks "S-" przy statystykach podanych w (1) oraz (2).

Oczekiwana proporcja odpowiedzi poprawnych na zadanie i w grupie o wyniku obserwowanym g przy obliczaniu statystyk $S\text{-}X^2$ oraz $S\text{-}G^2$ jest ostatecznie wyliczana jako:

$$E_{ig} = \frac{\int f_i(\theta)\tilde{S}^i_{g-1}(\theta)\varphi(\theta)d\theta}{\int S_g(\theta)\varphi(\theta)d\theta} \quad (3)$$

gdzie $f_i(\theta)$ jest oszacowaną funkcją charakterystyczną dla zadania i (prawdopodobieństwem odpowiedzi poprawnej na zadanie w zależności od θ), $\tilde{S}^i_{g-1}(\theta)$ jest funkcją wiarygodności określającą prawdopodobieństwo uzyskania wyniku g -1 przy usunięciu zadania i z testu, natomiast $\varphi(\theta)$ jest rozkładem *a priori* badanej umiejętności w populacji. W oryginalnym rozwiązaniu Orlando i Thissena wielkość (3) była obliczana za pomocą kwadratury. Warto jednak zauważyć, zgodnie z uwagą poczynioną przez Glasa i Béguina (2010) w kontekście zrównywania wyników obserwowanych opartych na wykorzystaniu IRT, że wielkość (3) również dobrze może być oszacowana także poprzez metodę Monte Carlo, w której generuje się wystarczająco dużo wyników obserwowanych z wykorzystaniem oszacowanych parametrów modelu IRT.

W praktycznych aplikacjach metody Orlando i Thissena, jak i innych metod opartych na grupowaniu ze względu na wyniki obserwowane, często konieczne jest łączenie niektórych grup, aby oczekiwana liczba obserwacji przypadających na pojedynczą grupę osiągała pewną minimalną wartość dla przybliżenia statystyki testowej przez rozkład χ^2. Ogólnie akceptowaną granicą dla minimalnej oczekiwanej liczebności obserwacji w komórce tabeli dla testów typu X^2 oraz G^2 jest wartość 5, choć są znane przypadki, gdy nawet oczekiwana liczebność obserwacji wynosząca 1 skutkuje dobrym przybliżeniem (Agresti, 1990, s. 49). Trzeba jednak mieć na względzie, że łączenie kategorii ma wpływ na moc wnioskowania statystycznego (Chon, Lee i Dunbar, 2010). Zależność właściwości statystyk $S\text{-}X^2$ oraz $S\text{-}G^2$ od algorytmu zastosowanego przy łączeniu mało licznych kategorii można uznać za pewną ich wadę.

Klasyczne statystyki oparte na podziale na grupy ze względu na oszacowania punktowe

Pierwotne podejścia do testowania dopasowania modeli IRT na poziomie zadań korzystały również ze statystyk typu X^2 oraz G^2 , tak jak przytoczone we wzorach (1) oraz (2) statystyki Orlando i Thissena. Istotną różnicą, będącą jednocześnie głównym elementem ich krytycznej oceny, jednak był fakt, że zamiast dokonywać podziału na grupy umiejętności ze względu na wyniki obserwowane, grupowały badane osoby na podstawie oszacowań ich umiejętności na skali zmiennej ukrytej θ.

Pierwszą spośród takich statystyk jest Q_1 zaproponowana przez Yena (1981):

$$Q_1 = \sum_{g=1}^{10} N_g \left(\frac{\left(O_{ig} - E_{ig}\right)^2}{E_{ig}(1 - E_{ig})} \right). \quad (4)$$

Oprócz ograniczenia do 10 kategorii wzór na Q_1 ma identyczną postać jak $S\text{-}X^2$ podany w (1), jednak sposób podziału na grupy oraz obliczania wartości oczekiwanych oraz obserwowanych proporcji odpowiedzi poprawnych jest diametralnie odmienny od podejścia Orlando i Thissena. Przy obliczaniu statystyki Q_1 badane osoby są dzielone na mniej więcej równoliczne (decylowe) grupy na podstawie punktowych oszacowań $\hat{\theta}$. Wartość O_{ig} jest wyliczana jako średni wynik w zadaniu *i* w grupie *g*, natomiast E_{ig} jest wartością krzywej charakterystycznej dla zadania *i* w punkcie równym średniej z oszacowań punktowych θ w grupie *g*. Przyjmuje się, że statystyka Q_1 ma rozkład χ^2 z 10-m stopniami swobody, gdzie *m* liczbą szacowanych parametrów dla funkcji charakterystycznej zadania *i*.

Bardzo podobna do Q_1 statystyka została zaproponowana przez Bocka (1972), różniąca się od Q_1, jedynie brakiem ograniczenia do 10 równolicznych kategorii oraz wykorzystaniem mediany z oszacowań punktowych θ w grupie g zamiast średniej. Analizując konstrukcję statystyki Q_1, nasuwa się skojarzenie z bardzo popularną przy ocenie dopasowania modelu regresji logistycznej statystyką Hosmera i Lemeshowa (1980). Jednak w przypadku regresji logistycznej, w odróżnieniu od modeli IRT, zmienna niezależna jest obserwowana i nie zależy od parametrów modelu podlegających estymacji.

McKinley i Mills (1985) zaproponowali statystykę, w ramach której podział na grupy oraz wyliczanie obserwowanej oraz oczekiwanej proporcji odpowiedzi poprawnych (O_{ig} oraz E_{ig}) jest taki sam jak w Q_1 Yena (4), jednak postać statystyki testowej odwołuje się do ilorazu wiarygodności:

$$G^2 = 2 \sum_{g=1}^{10} N_g \left(O_{ig} \ln\left(\frac{O_{ig}}{E_{ig}}\right) + \left(1 - O_{ig}\right) \ln\left(\frac{1 - O_{ig}}{1 - E_{ig}}\right) \right). \quad (5)$$

Przyjmuje się, podobnie jak we wcześniejszych przypadkach, że przy prawdziwości hipotezy zerowej G^2 ma rozkład χ^2 z 10-m stopniami swobody. Można zauważyć, że statystyka (5) jest swoistym protoplastą statystyki $S\text{-}G^2$ (2), tak samo jak Q_1 Yena. (4) służy za punkt wyjścia dla $S\text{-}X^2$ (1)

Zastosowany w (4) oraz (5) podział na w przybliżeniu równoliczne grupy oparte na decylach pozwala uniknąć kłopotów związanych z niewystarczającą liczebnością oczekiwanych obserwacji w grupach, z czym mamy do czynienia w przypadku statystyk Orlando i Thissena. Jak się jednak okazuje, dokonanie podziału badanych osób na grupy umiejętności ze względu na oszacowania punktowe zmiennej ukrytej odbija się bardzo negatywnie na właściwościach tak skonstruowanych statystyk. W swoich symulacyjnych badaniach Orlando i Thissen (2000) wykazali, że proponowane przez nich statystyki (1) oraz (2)

mają wielkość błędów I rodzaju zbliżoną do wartości nominalnej, natomiast statystyki (4) oraz (5) oparte na podziale na grupy ze względu na punktowe oszacowania θ drastycznie ją zawyżają. Efekt zawyżonych błędów I rodzaju („fałszywych alarmów") jest tym większy, im mniejsza liczba obserwacji oraz im krótsze są rozpatrywane testy.

Przyczyny zawyżonych błędów I rodzaju dla statystyk są upatrywane w dwóch źródłach. Po pierwsze, tworzenie przedziałów grupujących ze względu na oszacowania ukrytej zmiennej umiejętności powoduje, że obserwowane proporcje odpowiedzi poprawnych są zależne od oszacowań parametrów modelu IRT. Jak to ujął Sandip Sinharay (2006), taka sytuacja jest niezgodna „z duchem" tradycyjnego testowania dopasowania odwołującego się do rozkładu χ^2, w którym obserwowane proporcje są… obserwowane, czyli niezależne od parametrów modelu i możliwe do wyliczenia przed przystąpieniem do szacowania parametrów modelu. To zastrzeżenie było jedną z głównych przyczyn, dla których Orlando i Thissen rozwinęli swoje statystyki $S\text{-}X^2$ oraz $S\text{-}G^2$ oparte na podziale na grupy ze względu na wyniki obserwowane. Kolejne zastrzeżenie jest związane z tym, że przy konstrukcji przedziałów oszacowania punktowe umiejętności są traktowane jakby były prawdziwymi wartościami umiejętności badanych osób. W rzeczywistości są one obarczone błędem i w części przypadków klasyfikacja badanej osoby do danej kwantylowej grupy umiejętności na podstawie oszacowań θ jest niezgodna z jej faktycznym poziomem umiejętności. Przy obliczaniu statystyk (4) oraz (5) stochastyczna natura oszacowań poziomu umiejętności jest, zatem, zupełnie zignorowana.

Warto zauważyć, że zastrzeżenie o nieuwzględnianiu stochastycznej natury oszacowań parametrów modeli IRT w pewnej mierze dotyczy również miar dopasowania Orlando i Thissena. Problem ten podnieśli np. Glas oraz Suárez Falcón (2003). W przypadku statystyk $S\text{-}X^2$ oraz $S\text{-}G^2$ zarówno przyporządkowanie badanych do grup, jak i wyliczanie obserwowanej proporcji odpowiedzi poprawnych, O_{ig}, jest niezależne od jakichkolwiek oszacowań parametrów modelu IRT. Oszacowania parametrów modelu IRT są - jednak - wciąż wykorzystywane przy wyliczaniu oczekiwanej proporcji odpowiedzi poprawnych, E_{ig}. Sinharay (2006) przeprowadził symulacje, w których wykazał, że statystyki $S\text{-}X^2$ oraz $S\text{-}G^2$ mają nieznacznie zawyżone błędy I rodzaju, przy niewielkiej liczbie zadań i małych liczebnościach badanych osób. Zawyżona częstość fałszywych alarmów tych statystyk może być powiązana właśnie z korzystaniem z punktowych oszacowań parametrów zadań przy obliczaniu E_{ig}, zgodnie ze wzorem (3). Mimo znacznego polepszenia właściwości statystyk, jaki nastąpił między Q_1 Yena oraz G^2 McKinley'a i Millsa a parą $S\text{-}X^2$ oraz $S\text{-}G^2$ Orlando i Thissena, zagadnienie poprawnej i efektywnej oceny dopasowania zadań było w następnych latach ponownie podejmowane przez wielu badaczy.

Podejście Glasa i Suárez Falcóna

W znacznym stopniu odrębnym, na tle wszystkich innych, rozwiązaniem problemu analizy dopasowania zadań jest podejście zaproponowane przez Ceesa Glasa (1999), rozwinięte dalej wraz z Juanem Carlosem Suárez Falcónem (2003). Wyjątkowość ich podejścia polega na tym, że statystyka testowa

służąca do oceny dopasowania nie jest zbudowana na porównywaniu obserwowanych i przewidywanych proporcji poprawnych odpowiedzi (standaryzowanych reszt) w grupach umiejętności. Zamiast tego, zaproponowane przez nich rozwiązanie odwołuje się do przeprowadzenia testu dla zagnieżdżonych parametrycznych modeli – modelu zerowego, będącego dopasowanym do danych modelem IRT, oraz modelu alternatywnego, w którym wprowadza się dodatkowe parametry umożliwiające zamodelowanie specyficznego dla grup uczniów niedopasowania zadania.

Podobnie jak w statystykach Orlando i Thissena, grupowanie uczniów jest przeprowadzane ze względu na wyniki obserwowane. Zadanie poddawane testowaniu ze względu na dopasowanie jest wyłączane przy obliczaniu wyniki sumarycznego, więc grupowanie jest specyficzne dla każdego zadania *i*, co znajduje swoje odzwierciedlenie w indeksacji grup: $g^{(i)} \in \{1,\ldots,G^{(i)}\}$. Glas i Suárez Falcón (2003) w swoich symulacjach zastosowali podział, w którym tak wyliczone wyniki obserwowane łączono w $G^{(i)} = 4$ oraz $G^{(i)} = 9$ w przybliżeniu równolicznych grup. Dla grup $s \in \{1,\ldots,G^{(i)}\}$-1} w modelu alternatywnym wprowadzany jest dodatkowy, specyficzny dla grupy, parametr modyfikujący trudność zadania β_{is} (ostatnia grupa jest pominięta, gdyż dodając w niej parametr, uzyskalibyśmy nieidentyfikowalny model). Hipotezy zerowa i alternatywna dla trójparametrycznego modelu logistycznego są określone poprzez parę:

$$P(x_i = 1|\theta, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}} \quad (6.a)$$

$$P\left(x_i = 1|\theta, a_i, b_i, c_i, g^{(i)} = s\right) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - (b_i + \beta_{is}))}} \quad (6.b)$$

Jak widać, model z H_1 jest zagnieżdżony w modelu H_0. Typowo stosowanym rozwiązaniem dla zagnieżdżonych modeli jest przeprowadzenie testu ilorazu wiarygodności. Wymagałoby to jednak przeprowadzenia kompletnej estymacji modelu alternatywnego dla każdego z analizowanych zadań, co byłoby czasochłonne ze względu na wykorzystanie algorytmu EM w modelach IRT (zob. Kondratek, 2015). Aby uniknąć konieczności przeprowadzenia pełnej estymacji parametrów modelu alternatywnego, Glas (1999) zaproponował wykorzystanie testu mnożnika Lagrange'a, który wymaga przeprowadzenia tylko pojedynczej dodatkowej iteracji algorytmu EM dla modelu rozszerzonego o parametry wprowadzone w H_1. Postać statystyki testowej jest następująca:

$$LM(\beta_i) = h_i' \, \Sigma_{(i)}^{(-1)} \, h_i, \quad (7)$$

gdzie h_i jest wektorem pochodnych cząstkowych z funkcji wiarygodności ze względu na parametry β_{is} włączone w modelu H_1, natomiast $\Sigma_{(i)}$ jest macierzą kowariancji dla h_i. Statystyka (7) ma rozkład χ^2 z $G^{(i)}$ -1 stopniami swobody.

Dzięki temu, że dopasowanie zadania jest rozpatrywane jako problem zweryfikowania, czy pewne dodatkowe parametry modelujące niedopasowanie różnią się w sposób istotny statystycznie od zera, podejście Glasa i Suárez Falcóna jest wolne od zastrzeżenia podnoszonego do wcześniej omówionych metod. Mianowicie, taki sposób postępowania uwzględnia niepewność pomiaru

parametrów modelu IRT. Wyniki przeprowadzonych przez autorów symulacji jednak nie potwierdziły lepszych właściwości statystycznych proponowanego podejścia w porównaniu ze statystykami Orlando i Thissena. Statystyka $LM(\beta_i)$ miała w niektórych przypadkach zawyżony poziom błędów I rodzaju w porównaniu z $S\text{-}X^2$ oraz $S\text{-}G^2$, a także okazała się, jak podkreśla Sinharay (2006), silnie zależna od zastosowanej liczby grup, $G^{(i)}$. Swaminathan, Hambleton oraz Rogers (2007) zauważają dodatkowo, że parametryzacja wprowadzona celem modelowania niedopasowania może ogranicza użyteczność metody jedynie do niektórych sytuacji. Ze względu na silną teoretyczną podbudowę podejścia Glasa i Suárez Falcóna ich rozwiązanie wydaje się warte podjęcia dalszych badań.

Ocena dopasowania zadań w podejściu Bayesowskim

Sinharay (2006), wykazawszy nieznacznie zawyżone błędy pierwszego rodzaju dla $S\text{-}X^2$ oraz $S\text{-}G^2$ Orlando i Thissena zaproponował analogiczne miary dopasowania, jednak osadził je w metodologii bayesowskiej. Wykorzystał w tym celu metodę PCC (*posterior predictive checks*) rozwiniętą przez Rubina (1984).

W metodzie PCC ocena dopasowania modelu jest przeprowadzana poprzez badanie predyktywnego rozkładu *a posteriori* (*posterior predictive distribution*), który jest zreplikowanym rozkładem obserwowanych danych na podstawie modelu bayesowskiego dopasowanego do faktycznych danych:

$$P(x^{rep}|x)=\int P(x^{rep}|\xi)P(\xi|x)d\xi, \tag{8}$$

gdzie x to prawdziwe dane, x^{rep} to dane zreplikowane, natomiast ξ to parametry modelu. Aby ocenić dopasowanie modelu, wprowadzana jest pewna miara rozbieżności $D(x,\xi)$ i jej rozkład *a posteriori* jest porównany z rozkładem $D(x^{rep},\xi)$. Rozbieżności pomiędzy tymi dwoma rozkładami wskazują na niedopasowanie modelu, które opisuje się poprzez predyktywne p-wartości *a posteriori (posterior predictive p-values)*, będące bayesowskim odpowiednikiem klasycznych p-wartości. Jednostronna bayesowska p-wartość ma postać:

$$p=P(D(x^{rep},\xi)\geq D(x,\xi|x)). \tag{9}$$

Główną korzyścią z zastosowania bayesowskich p-wartości (9) do oceny statystycznej istotności niedopasowania modelu jest brak założeń co do postaci rozkładu wykorzystanych miar rozbieżności. Glas (1988), a także Glas i Verhelst (1989) przedstawili teoretyczne argumenty, które podważają słuszność założenia o zgodności statystyk $S\text{-}X^2$ oraz $S\text{-}G^2$ w formie zaproponowanej przez Orlando i Thissena z deklarowanym rozkładem χ^2. Sinharay (2006), stosując funkcje analogiczne do $S\text{-}X^2$ oraz $S\text{-}G^2$, ale w ramach opisanej bayesowskiej procedury, zatem broni się przed tymi zastrzeżeniami. Przeprowadzone przez Sinharaya symulacje wykazały, że jego podejście nie przejawia zawyżonych błędów I rodzaju w sytuacjach, w których takie zachowanie zaobserwował dla $S\text{-}X^2$ oraz $S\text{-}G^2$. Jednocześnie okazało się jednak, że błąd I rodzaju w jego podejściu jest nieznacznie niższy od nominalnej wartości 0,05. Skomentował ten fakt, zauważając, że zaniżona czułość procedury stanowi mniejszą wadę niż błędne wskazywanie zadania jako niedopasowane do modelu.

Statystyki grupujące ze względu na zmienną θ uwzględniające niepewność pomiaru

Jak wspomniano, pierwotnie stosowane grupowanie ze względu na punktowe oszacowania zmiennej ukrytej wiązało się z drastycznie zawyżonym odsetkiem fałszywych alarmów. W związku z tym późniejsze podejścia skupiały się rozwoju metod wykorzystujących grupowanie badanych osób ze względu na wyniki obserwowane. Część badaczy podjęła jednak próbę rozwiązania problemów pojawiających się przy grupowaniu ze względu na zmienną ukrytą poprzez uwzględnienie niepewności jej pomiaru przy obliczaniu.

W wyniku estymacji parametrów IRT algorytmem EM, gdzie następuje całkowanie względem rozkładu umiejętności z wykorzystaniem kwadratury, rozkład umiejętności każdej osoby jest przybliżony wagami przypisanymi do skończonej liczby punktów umiejętności (zob. Kondratek, 2015). Stone (2000) zaproponował podejście, w którym klasyczne statystyki typu X^2 lub G^2 są budowane z wykorzystaniem „pseudo-częstości" (*pseudocounts*), $\tilde{N}_g$, wyliczanych na punktach kwadratury, *g*. Stone mnoży odpowiedzi poprawne udzielone przez każdą osobę przez wspomniane wagi, tak aby wyliczyć ich udział na każdym punkcie kwadratury. Następnie, uśredniając po wszystkich badanych osobach, uzyskuje „pseudo-proporcje" odpowiedzi poprawnych, $\tilde{O}_{ig}$, dla analizowanego zadania *i* w każdym punkcie kwadratury *g*. Dzięki temu odpowiedź poprawna pojedynczej osoby zostaje uwzględniona jednocześnie we wszystkich punktach kwadratury *g*, a jej wkład w obliczanie konkretnej $\tilde{O}_{ig}$, jest proporcjonalny do gęstości rozkładu umiejętności tych osób w punktach *g*. Dodając do tego oczekiwane proporcje odpowiedzi poprawnych, E_{ig}, obliczone w tych samych punktach kwadratury, uzyskujemy wszystkie składniki gotowe do skonstruowania statystyk dopasowania typu X^2 lub G^2, które ze względu na korzystanie z pseudo-częstości oznaczymy „falką":

$$\tilde{X}^2 = \sum_{g=1}^{G} \tilde{N}_g \left(\frac{\left(\tilde{O}_{ig} - E_{ig}\right)^2}{E_{ig}(1 - E_{ig})} \right) \tag{10}$$

$$\tilde{G}^2 = 2 \sum_{g=1}^{G} \tilde{N}_g \left(\tilde{O}_{ig} \ln\left(\frac{\tilde{O}_{ig}}{E_{ig}}\right) + \left(1 - \tilde{O}_{ig}\right) \ln\left(\frac{1 - \tilde{O}_{ig}}{1 - E_{ig}}\right) \right). \tag{11}$$

Korzystanie z opisanych powyżej pseudo-częstości powoduje jednak, że nie jest spełnione założenie o niezależności obserwacji (obserwacja należy jednocześnie do wielu komórek tabeli) konieczne do uzyskania rozkładu X^2. Kolejnym ograniczeniem przed przyjęciem założenia o rozkładzie X^2są również bardzo niewielkie liczebności (poniżej 1) na skrajnych punktach kwadratury. W związku z tym Stone zaproponował techniki wielokrotnego losowania dla wyznaczenia rozkładu statystyk testowych, co wiąże się jednak ze znacznym obciążeniem obliczeniowym jego metody.

Stone oraz Zhang (2003) porównali właściwości omawianej metody ze statystykami *S*-X^2 oraz *S*-G^2 z Orlando i Thissena. Obie metody okazały się mieć zbliżoną do nominalnej wartości wielkość błędów I rodzaju, przy czym podejście Stone'a miało

niższą moc statystyczną niż podejście Orlando i Thissena. W obliczu tych wyników, wydaje się, że stosowanie wymagającej pod względem obliczeniowym procedury Stone'a jest zasadne w wypadkach, gdy oparte na grupowaniu ze względu na wyniki obserwowane statystyki Orlando i Thissena nie mogą być zastosowane. Ma to miejsce na przykład przy komputerowym testowaniu adaptatywnym (CAT), gdzie wynik sumaryczny nie jest wspólną dla wszystkich badanych osób miarą umiejętności.

Propozycja nowych statystyk uwzględniających stochastyczną naturę zmiennej

Zgodnie z dokonanym przeglądem, mimo licznych opracowań i wielu proponowanych statystyk dopasowania, wciąż jednak nie wypracowano rozwiązania, które byłoby powszechnie uznane. Poniżej nakreślona zostanie autorska propozycja nowego ujęcia problemu, które potencjalnie może być wolne od wielu wad wskazywanych przy okazji omawiania dotychczas opracowanych metod.

Postulowane rozwiązanie polega na podziale badanych osób na równoliczne przedziały ze względu na kwantyle (centyle lub decyle) rozkładu θ z prawdopodobieństwem proporcjonalnym do gęstości rozkładu *a posteriori* θ danej osoby w każdej z grup. Rozkład *a posteriori* mierzonej testem umiejętności określa gęstość prawdopodobieństwa tego, jaki jest poziom θ badanej testem osoby *j*, $j \in \{1, \ldots, J\}$, przy uwzględnieniu informacji o udzielonych przez nią odpowiedziach:

$$P(\theta|\mathbf{x}_j) = \frac{P(\mathbf{x}_j|\theta)P(\theta)}{P(\mathbf{x}_j)}, \tag{12}$$

gdzie $P(x_j|\theta)$ jest funkcją wiarygodności opisującą zależność między parametrami modelu IRT a prawdopodobieństwem zaobserwowania wektora odpowiedzi x_j, $P(\theta)$ jest rozkładem *a priori* badanej cechy w populacji, natomiast $P(x_j)$ jest brzegowym, bezwarunkowym, prawdopodobieństwem zaobserwowania wektora odpowiedzi x_j w populacji (zob. Kondratek i Pokropek, 2015).

Rozkład *a posteriori* (12) podsumowuje całą informację o poziomie θ badanej osoby na podstawie przeprowadzonego pomiaru. Wartość oczekiwana z tego rozkładu stanowi popularny estymator punktowy EAP (*expected a posteriori*), jednak szczególnie istotna w wielu kontekstach jest również informacja o niepewności pomiaru, jaką zawiera (12). Przykładowo, zgodnie z metodą rozwiniętą przez Donalda Rubina (2004) analizy wykorzystujące wielokrotne realizacje z rozkładu (12) dają nieobciążone oszacowania statystyk będących funkcją θ. Jest to powszechnie wykorzystywane we współczesnych badaniach edukacyjnych (Wu, 2004), gdzie takie wielokrotne imputacje noszą nazwę wartości możliwych (*plausible values*, PV).

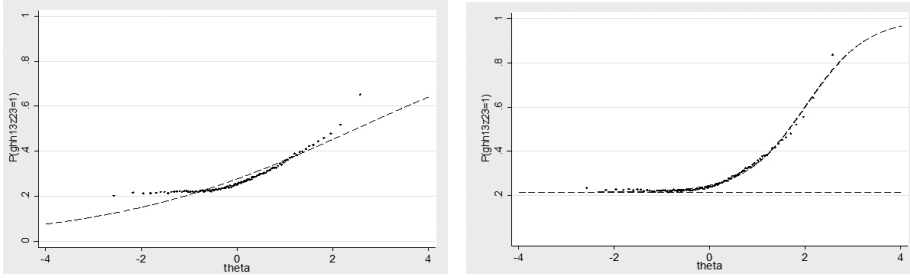
W proponowanej metodzie ogólny algorytm postępowania w celu wyliczenia proporcji obserwowanych wartości jest następujący:

1. dla każdej badanej osoby wylosuj *M* PV, łącznie uzyskując ich *MJ*; pojedynczą PV oznaczmy: θ^*_{mj};

2. na podstawie uzyskanych PV skonstruuj *G* (100 dla centyli, 10 dla decyli) przedziałów kwantylowych rozkładu umiejętności:
$$\{\theta\}_g = \left\{\theta^*_{mj}: \frac{g-1}{G} < P\left(\theta^*_{st} \leq \theta^*_{mj}\right) \leq \frac{g}{G}\right\},$$
co sprowadza się do posortowania θ^*_{mj} i **podzielenia na zbiory o liczebności** w zaokrągleniu równych $\frac{MJ}{G}$;
3. w obrębie każdego przedziału $\{\theta\}_g$ oszacuj proporcję obserwowanych w tym przedziale odpowiedzi poprawnych na zadanie *i*:
$$O^*_{ig} = \frac{G}{MJ}\sum_{m=1}^{M}\sum_{j=1}^{J}\mathbf{1}_{\{\theta\}_g}\left(\theta^*_{mj}\right)x_{ij}, \tag{13}$$

gdzie $\mathbf{1}_{\{\theta\}_g}$ jest indykatorem zwracającym wartość 1, gdy argument znajduje się w zbiorze $\{\theta\}_g$, 0 w przeciwnym wypadku.

Dzięki opisanej procedurze odpowiedzi x_{ij} są wykorzystywane podczas **obliczania** procentu poprawnych odpowiedzi w danym przedziale $\{\theta\}_g$ w sposób proporcjonalny do gęstości rozkładu *a posteriori* (12) w tym przedziale. Opisany algorytm wyliczania O^*_{ig} został już zaimplementowany w autorskim programie UIRT[1], przy czym jego wykorzystanie jest ograniczone jedynie do przeprowadzania graficznej analizy dopasowania zadań (rys. 1). Natomiast problem zbudowania statystyki dopasowania na podstawie tak szacowanych obserwowanych proporcji odpowiedzi poprawnych nie został jeszcze zbadany.



Rysunek 1. Przykład graficznej analizy dopasowania modelu IRT do danych z wykorzystaniem proponowanego algorytmu dla zadania z części humanistycznej egzaminu gimnazjalnego z 2013 roku (Kondratek i Pokropek, 2015)

Oszacowanie proporcji oczekiwanych odpowiedzi poprawnych w przedziałach $\{\theta\}_g$, konieczne do zbudowania statystyk dopasowania, można w opisanym podejściu uzyskać analogicznie do (13):

[1] Program UIRT (*Unidimensional Item Response Theory models)* działający w środowisku Stata operacyjnie wykorzystano w kilku badaniach (np. Kondratek i Szaleniec (2015), Kondratek, Szaleniec i Legutko (2013)). Uzyskanie kopii programu jest możliwe poprzez kontakt z autorem.

$$E_{ig}^{*} = \frac{G}{MJ} \sum_{m=1}^{M} \sum_{j=1}^{J} \mathbf{1}_{\{\theta\}_g}(\theta_{mj}^{*}) f_i(\theta_{mj}^{*}) \tag{14}$$

gdzie $f_i(\theta)$ jest oszacowaną postacią funkcji charakterystycznej dla zadania *i*.

Zestawienie wielkości (13) oraz (14) otwiera drogę do budowy statystyk dopasowania typu χ^2 oraz G^2:

$$\theta^{*} - \mathrm{X}^2 = J \sum_{g=1}^{G} \left(\frac{\left(O_{ig}^{*} - E_{ig}^{*}\right)^2}{E_{ig}^{*}(1 - E_{ig}^{*})} \right) \tag{15}$$

$$\theta^{*} - G^2 = 2J \sum_{g=1}^{G} \left(O_{ig}^{*} \ln\left(\frac{O_{ig}^{*}}{E_{ig}^{*}}\right) + \left(1 - O_{ig}^{*}\right) \ln\left(\frac{1 - O_{ig}^{*}}{1 - E_{ig}^{*}}\right) \right). \tag{16}$$

Para θ^{*}- χ^2 oraz θ^{*}- G^2 jest analogiczna do S - χ^2 (1) oraz S - G^2 (2) Orlando and Thissena (2000). Zasadnicza różnica między statystykami jest odzwierciedlona w oznaczeniach – u Orlando i Thissena grupowanie odbywa się po wynikach obserwowanych (*S*), natomiast tutaj poprzez wykorzystanie wartości możliwych (θ^{*}), które determinuje też obliczanie proporcji O^{*}_{ig} oraz E^{*}_{ig}. Zauważalny jest również brak specyficznych dla grup liczebności N_g, obecnych w (1) oraz (2), który został zastąpiony poprzez całkowitą liczbę badanych osób, *J*, przed znakiem sumowania. Wynika to z faktu, że w proponowanej metodzie każda grupa jest zdefiniowana przez kwantylowe punkty odcięcia – podobnie jak w statystykach Q_1 Yena (4) czy G^2 McKinley'a i Millsa (5).

Statystyki θ^{*}- χ^2 oraz θ^{*}- G^2 odznaczają się potencjalnie następującymi zaletami:

1. Uwzględniona jest w nich stochastyczna natura oszacowań parametrów IRT.
2. Oczekiwane liczebności w każdej z grup nie osiągają bardzo niskich wartości (ze względu na podział w oparciu o kwantyle grupy będą równoliczne).
3. Możliwość podziału na znaczną liczbę grup (np. centylowych) pozwala na detekcję zaburzeń dopasowania o bardziej lokalnym charakterze.
4. Możliwość zastosowania w sytuacji braków odpowiedzi na część zadań (np. CAT).
5. Możliwość wsparcia analizy poprzez informatywną graficzną analizę dopasowania.

Żadna pojedyncza metoda analizy dopasowania modelu IRT do zadań z omówionych w niniejszym opracowaniu nie łączy w sobie tych wszystkich zalet.

Zaproponowane statystyki θ^{*}- χ^2 oraz θ^{*}- G^2 muszą jednak zostać poddane weryfikacji w badaniach symulacyjnych pod kątem ich rozkładu przy prawdziwości hipotezy zerowej. Mają względem nich bowiem zastosowanie zastrzeżenia sformułowane przy okazji omawiania podejścia Stone'a, tj. zaburzenie założenia o niezależności obserwacji w komórkach tabeli ze względu

na wykorzystanie pojedynczej odpowiedzi przy wyliczaniu O^*_{ig} dla wielu grup jednocześnie. Możliwość ustalenia rozkładu tych statystyk w sposób obliczeniowo złożony nie bardziej niż ma to miejsce w podejściu Stone'a będzie kluczowym kryterium dla określenia ich faktycznej użyteczności. Wątek ten jest przedmiotem planowanych badań.

Bibliografia

1. *American Educational Research Association, American Psychological Association, & National Council for Measurement in Education* (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.
2. Bolt, D.M., Deng, S., Lee, S. (2014). *IRT model misspecification and measurement of growth in vertical scaling*. Journal of Educational Measurement, 51(2), 141-162.
3. Chon, K.H., Lee, W.-C. and Dunbar, S.B. (2010). *A Comparison of Item Fit Statistics for Mixed IRT Models*. Journal of Educational Measurement, 47: 318-338.
4. Glas, C.A.W. (1988). *The derivation of some tests for the Rasch model from the multinomial distribution.* Psychometrika, 53, 525-546.
5. Glas, C.A.W. (1999). *Modification indices for the 2-pl and the nominal response model.* Psychometrika, 64, 273-294.
6. Glas, C.A.W., Béguin, A.A. (2010). *Robustness of IRT observed-score equating.* [w:] A.A. von Davier (red.), *Statistical Models for Test Equating, Scaling, and Linking*. (297-316). New York, NJ: Springer.
7. Glas, C.A.W., Suarez-Falcon, J.C. (2003). *A comparison of item-fit statistics for the three-parameter logistic model. A*pplied Psychological Measurement, 27(2), 87-106.
8. Glas, C.A.W., & Verhelst, N.D. (1989). *Extensions of the partial credit model.* Psychometrika, 54, 635-659.
9. Hosmer, D.W., Lemeshow, S. (1980). *Goodness-of-fit tests for the multiple logistic regression model*, Communications in Statistics - Theory and Methods. 10, 1043-1069.
10. Kondratek, B. (2015). *Estymacja parametrów modeli ze zmiennymi ukrytymi za pomocą algorytmu EM*. [w:] Pokropek, A. (red.). *Modele cech ukrytych w psychologii, socjologii i badaniach edukacyjnych. Teoria i zastosowania*. (198-211). Warszawa: IBE.
11. Kondratek, B., Pokropek, A. (2015). *Teoria odpowiedzi na pozycje testowe: jednowymiarowe modele dla cech ukrytych o charakterze ciągłym.* [w:] Pokropek, A. (red.). *Modele cech ukrytych w psychologii, socjologii i badaniach edukacyjnych. Teoria i zastosowania*. (15-32). Warszawa: IBE.
12. Kondratek B., Szaleniec, H., Legutko, M. (2013). *Rola diagnozy w Gimnazjalnym Programie Kształcenia Kompetencji Kluczowych*. [w:] Niemierko, B., Szmigel, M.K. (red.). *Polska edukacja w świetle diagnoz prowadzonych z różnych perspektyw badawczych*. Kraków: PTDE.
13. Kondratek, B., Szaleniec, H. (2015). *Sprowadzanie wyników różnych testów do wspólnej skali*. [w:] Pokropek, A. (red.). *Modele cech ukrytych w psychologii, socjologii i badaniach edukacyjnych. Teoria i zastosowania*. (320-334).Warszawa: IBE.
14. Lord, F. M., Wingersky, M.S. (1984). *Comparison of IRT true-score and equipercentile observed-score "equatings"*. Applied Psychological Measurement, 8, 453-461.
15. McKinley, R., Mills, C. (1985). *A comparison of several goodness-of-fit statistics.* Applied Psychological Measurement, 9, 49-57.

16. Orlando, M., Thissen, D. (2000). *Likelihood-based item-fit indices for dichotomous item response theory models.* Applied Psychological Measurement, 24, 50-64.
17. Rubin, D. B. (1984). *Bayesianly justifiable and relevant frequency calculations for the applied statistician*. Annals of Statistics, 12, 1151-1172.
18. Rubin, D.B. (2004). *Multiple imputations for non-response in surveys*. New York: Wiley.
19. Sinharay, S. (2006). *Bayesian item fit analysis for unidimensional item response theory models*. British Journal of Mathematical and Statistical Psychology, 59 , 429-449.
20. Stone, C.A. (2000). *Monte Carlo Based Null Distribution for an Alternative Goodness-of-Fit Test Statistic in IRT Models.* Journal of Educational Measurement, 37, 58-75.
21. Stone, C.A., Zhang, B. (2003). *Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures.* Journal of Educational Measurement 40 (4), 331-352.
22. Swaminathan, H., Hambleton, R.K., Rogers, H.J. (2007). *Assessing the Fit of Item Response Theory Models*. [w:] C.R. Rao. S. Sinharray (red.). *Handbook of Statistics, Volume 26: Psychometrics*. (683-718). Amsterdam: Elsevier Publishing Company.
23. Toribio, S.G., Albert, J.H. (2011). *Discrepancy measures for item fit analysis in item response theory*. Journal of Statistical Computation and Simulation, 81, 1345-1360.
24. Wainer, H., Thissen, D. (1987). *Estimating ability with the wrong model*. Journal of Educational Statistics, 12, 339-368.
25. Woods, C.M. (2008). *Consequences of ignoring guessing when estimating the latent density in item response theory.* Applied Psychological Measurement, 32(5), 371-384.
26. Wu, M. (2005). *The Role of Plausible Values in Large-Scale Surveys.* Elsevier: Studies in Educational Evaluation 31, 114-128.
27. Yen, W. (1981). *Using simulation results to choose a latent trait model.* Applied Psychological Measurement, 5, 245-262.