HENRYK SZALENIEC
OKE Kraków

WYKORZYSTANIE PROBABILISTYCZNYCH MODELI ZADANIA TESTOWEGO DO ZRÓWNYWANIA WYNIKÓW SPRAWDZIANU 2003–2005 I BUDOWANIA BANKU ZADAŃ

Wstęp

Jednym z celów wprowadzenia egzaminów zewnętrznych było oczekiwanie, że ich wyniki będą przydatne do śledzenia zmian edukacyjnych w skali całego kraju i na przestrzeni wielu lat. Statystyczne wskaźniki wyników uczniów wyrażone w skali punktów surowych (liczba punktów za cały test) uzyskanych poprzez zastosowanie tych arkuszy w poszczególnych latach, mogą zmieniać się z powodu wahań poziomu badanej cechy w populacji. Z drugiej strony, wyniki uzyskane za pomocą różnych arkuszy egzaminacyjnych, zbudowanych nawet najbardziej porównywalnie, jak to jest możliwe do badania tych samych umiejętności, są różne, chociaż pomiar był przeprowadzony dla tej samej populacji. Oznacza to, że wyniki dwóch różnych testów nie są bezpośrednio porównywalne.

Względne zmiany poziomu osiągnięć uczniów w województwach, powiatach gminach i szkołach, a nawet i klasach nauczanych w tej samej szkole przez różnych nauczycieli są stosunkowo łatwo porównywalne w standardowej skali staninowej. W skali normowanej corocznie dla każdego egzaminu na krajowej populacji uczniów piszących dany egzamin, zapewniając układ odniesienia dający każdego roku średnią dla kraju równą 5 staninów. Niestety nie można w ten sposób dokonać porównań podłużnych (pomiędzy kolejnymi latami) dla całego kraju, gdyż średni wynik dla całej populacji każdego roku jest taki sam i wynosi 5 staninów.

Aby wyniki dwóch egzaminów mogły być stosowane zamiennie do wnioskowania o osiągnięciach kolejnych populacji uczniów, musimy dokonać kalibrowania arkuszy zastosowanych w kolejnych latach, czyli takiego wyjustowania narzędzi pomiaru, aby można było powiedzieć, że uzyskane za ich pomocą wyni-

ki egzaminów z różnych lat są równoważne. Niestety, w pomiarze dydaktycznym nie dysponujemy punktami stałymi dla skali, jak to ma miejsce w przypadku pomiaru temperatury. Empiryczna skala temperatury Celsjusza została określona przez przyjęcie 0^oC dla temperatury topnienia lodu i 100^oC dla temperatury wrzenia wody pod ciśnieniem 101325 N/m^2. Skala Fahrenheita zaś, stosowana w angielskim czy amerykańskim systemie miar, też jest oparta na dwóch punktach stałych, ale 0^oF odpowiada temperaturze mieszaniny śniegu z salmiakiem, a 100^oF – temperaturze normalnej ludzkiego ciała. Słuchając w Krakowie prognozy pogody z Waszyngtonu, musimy dokonać przeliczenia na podstawie przeprowadzonego kiedyś zrównania (ang. *equating*) odczytów temperatury, wiedząc, że 0^oC to 32^oF, a 100^oC to 212^oF. Dwa termometry, jeden ze skalą Celsjusza, a drugi ze skalą Fahrenheita, to tak, jak dwa różne arkusze sprawdzianu z 2003 i 2005 r. – z tą różnicą, że dla poziomu umiejętności nie możemy określić punktów stałych, jak w przypadku pomiaru temperatury. Dlatego też zrównywania musimy dokonywać za każdym razem, gdy zbudujemy nowe narzędzie pomiaru – kolejny arkusz egzaminacyjny, wykorzystując albo wspólne zadania dla dwóch sprawdzianów (zadania kotwiczące), albo próbę uczniów rozwiązujących dodatkowy test zawierający reprezentację zadań z obydwu testów.

Mislevy podkreśla[1], że proces zrównywania wyników egzaminu jest ściśle związany z procesem tworzenia arkuszy egzaminacyjnych. Jeżeli zrównanie dobrze funkcjonuje, to najczęściej dlatego, że przestrzegano procedury dobrego konstruowania narzędzi egzaminacyjnych. Warunkiem koniecznym do przeprowadzenia procesu zrównywania wyników dwóch egzaminów, niezależnie od wybranej metody, jest to, aby zastosowane na egzaminach narzędzia były zbudowane do pomiaru tego samego zakresu umiejętności i według tych samych specyfikacji statystycznych opisujących charakterystyki testu (liczba i forma zadań, łatwość, moc różnicująca zadań, rzetelność, odchylenie standardowe). W przypadku egzaminów zewnętrznych warunek ten jest spełniony, jeżeli zespoły autorskie arkuszy egzaminacyjnych poszczególnych okręgowych komisji egzaminacyjnych w pełni przestrzegają ustaleń Centralnej Komisji Egzaminacyjnej.

W ogólnych zarysach ustalenia CKE pozwalają na uzyskanie narzędzi mierzących ten sam zakres umiejętności w podobny sposób. Nie obejmują one jednak takich parametrów statystycznych, jak moc różnicująca zadań, łatwości zadań i odchylenie standardowe dla całego testu. W przyszłości warto rozważyć uzupełnienie ustaleń ogólnopolskich również w zakresie wspomnianych parametrów statystycznych arkusza. Zadanie to byłoby znacznie łatwiejsze, gdyby nasz system egzaminów zewnętrznych dysponował odpowiednio zasobnym bankiem zadań dla każdego egzaminu. Banku nie da się jednak utworzyć tylko poprzez zgromadzenie zadań i skatalogowanie ich według pomiarowych parametrów. Parametry zadań, takie jak moc różnicująca i trudność, muszą być wykalibrowane w tej samej przedziałowej skali, z ustalonym punktem „0" i jednostką pomiarową.

[1] Mislevy R.J., *Linking educational assessments: Concepts, issues, methods, and prospects*, ETS Policy Information Center, Princeton, NJ 1992.

Tworzenie banku zadań to drugie obok śledzenia trendów zmian osiągnięć w populacji zadanie, do którego konieczne jest zrównywania arkuszy egzaminacyjnych.

W praktyce pomiarowej stosowane są najczęściej trzy statystyczne metody zrównywania: metoda liniowa, metoda średnich i metoda ekwicentylowa, bazujące na klasycznej teorii pomiaru, oraz szereg metod z wykorzystaniem probabilistycznej teorii wyniku zadania (Item Response Theory – IRT). Każda może być zastosowana dla różnych planów eksperymentalnych powiązania ze sobą arkuszy egzaminacyjnych. Od trzech lat metoda ekwicentylowa z powodzeniem jest stosowana przez Wydział Sprawdzianów CKE i prof. Bolesława Niemierkę do zrównywania wyników sprawdzianu w kolejnych latach, przyjmując za rok bazowy 2003[2]. Wyniki prowadzonych analiz pozwoliły zauważyć zarysowujący się od 2003 r. trend wzrostu poziomu osiągnięć kolejnych roczników szóstoklasistów, spowodowany prawdopodobnie efektem zwrotnym egzaminów zewnętrznych.

Równolegle z analizami prowadzonymi z wykorzystaniem metody ekwicentylowej, począwszy od 2004 r. prowadzone są próby zastosowania do zrównywania arkuszy metod bazujących na probabilistycznej teorii wyniku zadania. Metody te od wielu lat są szeroko stosowane na świecie w instytucjach prowadzących badania porównawcze i egzaminy zewnętrzne. W dalszej części tego artykułu spróbujemy przedstawić doświadczenia uzyskane podczas zrównywania wyników sprawdzianu 2004 i 2005 r. do rezultatów zarejestrowanych w 2003 r.

1. Krótkie wprowadzenie do probabilistycznej teorii zadania

Po egzaminie dysponujemy dla każdego ucznia określoną liczbą punktów, które ten uzyskał za rozwiązanie zadań zawartych w arkuszu egzaminacyjnym. Tę liczbę punktów nazywamy też zaobserwowaną liczbą punktów i jest ona najczęściej podstawą wszelkich analiz w klasycznej teorii pomiaru dydaktycznego. Gdybyśmy mieli możliwość powtórzenia wiele razy tego samego testu dla danej osoby (w taki sposób, aby wykluczyć efekt zapamiętania rozwiązań), to otrzymalibyśmy szereg wyników, które miałyby rozkład normalny. Byłby to indywidualny (ograniczony do danej osoby) rozkład wyników z testu. Średnia arytmetyczna z tego *prywatnego* rozkładu punktów nazywana jest w klasycznej teorii testu *wynikiem prawdziwym*. Wynik prawdziwy nie jest obserwowalny - jest pojęciem statystycznym i nie ma nic wspólnego z pojęciem liczby punktów, która rzeczywiście określa wynik danego ucznia. Wynik zaobserwowany to jedna próba losowa z takiego rozkładu. Różnica pomiędzy wynikiem zaobserwowanym a wynikiem prawdziwym nazywa się błędem pomiaru.

---

[2] Niemierko B., *Zrównywanie wyników sprawdzianu 2004 do wyników sprawdzianu 2003*, [w:] B. Niemierko, H. Szaleniec (red.), *Diagnostyka edukacyjna. Standardy wymagań i normy testowe w diagnostyce edukacyjnej*, Kraków 2004; tenże, *Zrównywanie wyników sprawdzianu 2005 do wyników sprawdzianu 2003*, raport przesłany do CKE.

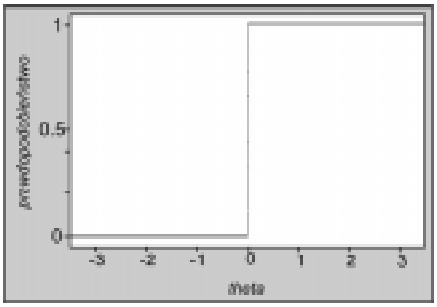
Jeżeli Beata i Romek uzyskali różne wyniki w sprawdzianie 2005 r., np. 32 i 35 pkt, to nie wiemy, czy ta różnica pochodzi z różnicy w poziomie ich umiejętności, czy też jej źródłem jest błąd pomiaru. Zmienność wyników w całej populacji piszących test jest sumą wariancji wyników prawdziwych poszczególnych uczniów i wariancji błędów pomiarowych.

$$war(X) = war(T) + war(E)$$

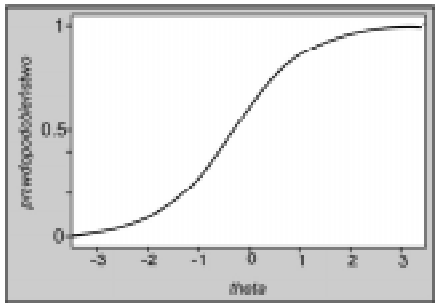
gdzie:
X – zmienna oznaczająca wynik zaobserwowany w wyniku testowania,
T – zmienna oznaczająca wynik prawdziwy,
E – zmienna oznaczająca błąd pomiaru.

Stosunek wariancji wyniku prawdziwego war(T) do wariancji wyników zaobserwowanych (liczby punktów) war(X) definiuje rzetelność wyników testowania. Oczywiście, nie da się oszacować rzetelności z formuły definicyjnej. Istnieje szereg metod szacowania rzetelności wyników testowania, ale w tym artykule nie będziemy się nimi zajmować. Wynik zaobserwowany (liczba punktów), wynik prawdziwy i rzetelność – to podstawowe pojęcia klasycznej teorii pomiaru dydaktycznego.

W probabilistycznej teorii wyniku zadania centralnym pojęciem jest nieobserwowalna cecha, która jest przedmiotem pomiaru (ang. *latent variable*). W pomiarze dydaktycznym może to być pojedyncza umiejętność lub zespół umiejętności. Cecha, która zmienia się w sposób ciągły od minus do plus nieskończoności i może być przedstawiona na osi liczb rzeczywistych. W praktyce najczęściej jest to przedział od -3 do +3. Punkt „0” na tej przedziałowej skali wybierany jest najczęściej w sposób arbitralny tak, aby odpowiadał średniej trudności testu. Celem pomiaru jest uszeregowanie osób najbardziej precyzyjnie jak to jest możliwe, ze względu na wyróżnioną cechę. Narzędziem do tego uszeregowania są zadania testowe. Poziom tej cechy to konkretny punkt na osi liczb rzeczywistych. Wzdłuż tej samej osi można uszeregować zadania ze względu na ich poziom trudności.



Krzywa charakterystyczna zadania w modelu deterministycznym



Krzywa charakterystyczna zadania w modelu probabilistycznym

Rys. 1. Zależność prawdopodobieństwa odpowiedzi na zadanie o trudności odpowiadającej punktowi 0 na skali w zależności od poziomu umiejętności ucznia (w deterministycznym i probabilistycznym modelu odpowiedzi na zadanie)

W klasycznej teorii pomiaru prawdopodobieństwo poprawnej odpowiedzi na zadanie można opisać za pomocą deterministycznego modelu. Jeżeli poziom umiejętności Romka jest niższy od poziomu trudności zadania (por. rys. 1. i 2.), to prawdopodobieństwo rozwiązania zadania przez testowanego wynosi 0. Natomiast, jeżeli poziom umiejętności Beaty jest wyższy od trudności zadania zastosowanego do pomiaru, to prawdopodobieństwo poprawnej odpowiedzi wynosi 1.

W modelu probabilistycznym prawdopodobieństwo poprawnej odpowiedzi na zadanie zmienia się w sposób ciągły od 0 dla bardzo niskiego poziomu umiejętności do 1 dla bardzo wysokiego, przyjmując wartość 0,5 dla poziomu umiejętności równej trudności zadania. Oznacza to m.in., że istnieje pewne prawdopodobieństwo, iż uczeń o bardzo niskim poziomie umiejętności badanej przez zadanie rozwiąże je poprawnie. Tak samo nie jest nieprawdopodobne, że uczeń o wysokim poziomie badanej umiejętności nie będzie potrafił go rozwiązać.

Podwaliny obecnie stosowanej teorii IRT stworzyli niezależnie Georg Rasch oraz Fred Lord. Georg Rasch rozwinął matematyczny model pomiaru, bazując na probabilistycznej relacji pomiędzy trudnością zadania i poziomem cechy badanej tym zadaniem. Model Rascha w najprostszym przypadku odnosi się do określenia warunkowego prawdopodobieństwa pozytywnej odpowiedzi na zadanie punktowane 0;1 jako funkcji różnicy pomiędzy trudnością zadania a poziomem umiejętności danego ucznia. Prawdopodobieństwo udzielenia poprawnej odpowiedzi i uzyskania 1 punktu przez ucznia *n* na zadanie *i* możemy wyrazić matematycznie za pomocą równania (1).

$$P_{ni}\,(x=1) = f\,(\theta_n - b_i) \tag{1}$$

Równanie (1) może być dokładniej przedstawione za pomocą funkcji logistycznej w następującej postaci:

$$P_{ni}(x_{ni}=1/\theta_n, b_i) = \frac{e^{(\theta_n - b_i)}}{1+e^{(\theta_n - b_i)}}, \tag{2}$$

gdzie:

P_{ni} – prawdopodobieństwo, że dany uczeń uzyska 1 punkt za zadanie „i" punktowane 0;1,

x_{ni} – wynik punktowy ucznia *n* za zadanie *i*,

b_i – parametr trudności zadania *i*,

θ_n – poziom umiejętności ucznia *n*,

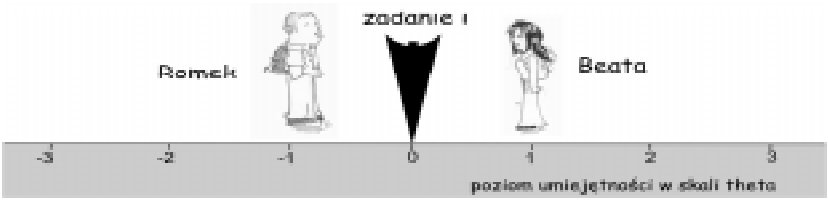
e = 2,7183 – odstawa logarytmów naturalnych.

Punkt wyjścia do stworzenia miary poziomu umiejętności w modelu Rascha jest taki sam jak w klasycznej teorii pomiaru. Jest to obliczenie proporcji poprawnych odpowiedzi na zadania w teście udzielonych przez danego ucznia (przy punktowaniu 0;1 – łatwość testu dla ucznia) i proporcji poprawnych odpowiedzi na dane zadanie udzielonych przez wszystkich uczniów, czyli łatwość danego zadania. Wyniki surowe w teście, tworzące tylko skalę porządkową, są zarówno

warunkiem koniecznym, jak wystarczającym do oszacowania poziomu badanej umiejętności θ_n i trudności zadania b_i. Różnica pomiędzy trudnością zadania a poziomem umiejętności osoby rozwiązującej zadanie jest podstawą do oszacowania prawdopodobieństwa rozwiązania tego zadania przez daną osobę. Logika modelu jest stosunkowo prosta. Dla wszystkich rozwiązujących zadania większe jest prawdopodobieństwo osiągnięcia sukcesu przy rozwiązywaniu łatwych zadań niż w przypadku rozwiązywania zadań trudniejszych. Pierwszym krokiem oszacowania poziomu umiejętności osoby *n* jest przekształcenie proporcji poprawnych odpowiedzi w prawdopodobieństwo sukcesu, które jest szacowane poprzez obliczenie ilorazu proporcji poprawnych odpowiedzi *p* do proporcji błędnych odpowiedzi (1-*p*). Logarytm naturalny z tego ilorazu stanowi w najprostszym przypadku oszacowanie umiejętności danej osoby. Dokładnie taka sama procedura powtarzana jest dla oszacowania trudności zadań. Dla danego zadania *i* obliczamy logarytm naturalny ilorazu proporcji poprawnych odpowiedzi na dane zadanie przez wszystkie osoby rozwiązujące oraz proporcji błędnych odpowiedzi osób rozwiązujących to zadanie. Następnie trudność zadania b_i i poziom umiejętności danej osoby θ_n lokowane są na wspólnej skali i wyrażane w jednostkach logarytmu naturalnego z proporcji nazywanych logit. Średnia wartość logit przyjmowana jest arbitralnie jako 0. Dodatnie wartości logit oznaczają poziom badanej cechy θ (np. umiejętności matematycznych badanych w sprawdzianie) wyższy od średniej, a ujemne wartości – niższy.

Przedstawiony powyżej opis dotyczy modelu Rascha, który zakłada jednakową dla wszystkich zadań moc różnicującą równą 1. Trudność zadania jest jedynym szacowanym parametrem dla zadań.

Przeanalizujmy trzy przykłady obliczania prawdopodobieństwa poprawnej odpowiedzi na zadanie o trudności $b_i = 0$ logit dla Romka i Beaty (rys. 2.), dla których znamy poziom umiejętności badanych danym testem.



Rys. 2. Przykład lokalizacji na osi umiejętności pozycji uczniów ze względu na poziom umiejętności i zadania ze względu na jego trudność

Przykład 1

Dla Beaty poziom umiejętności $\theta_n = 1$ logit. Jeżeli poziom umiejętności Beaty jest o 1 logit wyższy od trudności zadania, to różnica pomiędzy poziomem umiejętności i trudnością zadania wyrażona w jednostkach logit wyniesie 1 (rys. 2.). Podstawiając te wartości do równania (2) otrzymujemy:

$$P_{ni}(x_{ni}=1/\theta_n(1),b_i(0))=\frac{e^{(1-0)}}{1+e^{(1-0)}}=\frac{2{,}7183^1}{1+2{,}7183^1}=0{,}73 \qquad (3)$$

Otrzymany wynik informuje nas, że prawdopodobieństwo uzyskania 1pkt przez Beatę wynosi 0,73. Warto podkreślić, że, pomimo iż poziom umiejętności Beaty jest wyższy od trudności zadania, to istnieje prawdopodobieństwo równe 0,27, że uzyska ona za to zadanie 0 pkt. Jeżeli ten model idealnie przystaje do naszego pomiaru, to 27% uczniów o takim poziomie umiejętności jak Beata uzyska za to zadanie 0 pkt.

Przykład 2

Rozważmy ucznia, którego poziom umiejętności jest równy trudności zadania, czyli $\theta_n - b_i = 0$. Każda liczba podniesiona do potęgi 0 daje wynik 1.

$$P_{ni}(x_{ni}=1/\theta_n(0),b_i(0))=\frac{e^{(0-0)}}{1+e^{(0-0)}}=\frac{2{,}7183^0}{1+2{,}7183^0}=\frac{1}{1+1}=0{,}50 \qquad (4)$$

Obliczone prawdopodobieństwo poprawnej odpowiedzi dla takiego ucznia wyniesie 0,5. W idealnym przypadku tylko 50% takich uczniów uzyska za to zadanie 1 pkt.

Przykład 3

Rozważmy jeszcze jeden przypadek, odpowiadający na rys. 2. pozycji Romka, którego poziom umiejętności jest o 1 logit niższy od trudności zadania ($\theta_n - b_i = -1$).

$$P_{ni}(x_{ni}=1/\theta_n(-1),b_i(0))=\frac{e^{(-1-0)}}{1+e^{(-1-0)}}=\frac{2{,}7183^{-1}}{1+2{,}7183^{-1}}=\frac{0{,}3679}{1+0{,}3679}=0{,}27$$

Dla Romka, zgodnie z analizowanym modelem, prawdopodobieństwo poprawnej odpowiedzi na to zadanie wynosi tylko 0,27.

Obliczenia zastosowane w przykładach były stosunkowo łatwe, gdyż znaliśmy trudność zadania i poziom umiejętności wszystkich trzech uczniów. Oszacowanie trudności wszystkich zadań w teście i umiejętności wszystkich uczniów jest złożonym i zaawansowanym matematycznie zadaniem, wymagającym specjalistycznego oprogramowania komputerowego.

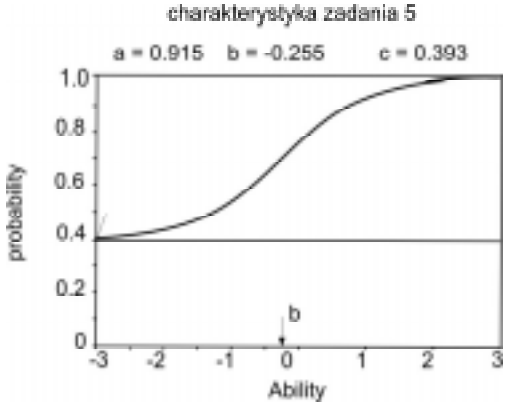
2. Modele wykorzystujące więcej niż jeden parametr

W praktyce pomiarowej stosowanych jest wiele modeli w ramach teorii odpowiedzi na zadanie testowe (IRT) różniących się matematycznym przedstawieniem funkcji charakterystycznej zadania czy liczbą parametrów tworzących

model. Każdy z modeli stosuje jeden lub więcej parametrów opisujących zadanie oraz jeden lub więcej parametrów opisujących egzaminowanego. Pierwsze próby analiz prowadzone były w Polskim Systemie Egzaminacyjnym z zastosowaniem takich programów komputerowych, jak RUMM (Andrich-Australian Council for Education), MULTILOG (*Multiple, Categorical Item Analysis and Test Scoring* –USA) i OPLM *(One Parameter Logistic Model*, opracowany w CITO przez N.D. Verhelsta, C.A.W. Glasa i H.H.F.M. Vertralen). Oprogramowanie to wykorzystuje jedno-, dwu- lub trójparametryczny model logistyczny sformułowany po raz pierwszy przez A. Birnbauma. W opisie modelu matematycznego parametry są oznaczane małymi literami: a, b, c. Zgodnie z modelem, prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie egzaminacyjne przez ucznia o poziomie umiejętności θ może być przedstawione funkcją logistyczną:

$$P_{ni}(\theta_n, b_i) = c_i + \frac{(1-c_i)}{1+e^{-Da_i\ (\theta - b_i)}} \tag{5}$$

gdzie:
b_i – parametr trudności zadania,
a_i – moc różnicująca,
c_i – współczynnik zgadywania odpowiedzi,
D – stała skalowania równa 1,7,
e – podstawa logarytmów naturalnych,
θ_n – poziom badanej cechy (umiejętności).



Rys. 3 Krzywa charakterystyczna zadania wielokrotnego wyboru z uwzględnieniem wszystkich trzech parametrów – a, b, c. Charakterystyka uzyskana z wykorzystaniem programu MultiLog dla jednego z zadań sprawdzianu

Jeszcze raz podkreślmy, że w modelach bazujących na probabilistycznej teorii zadania, inaczej niż w klasycznej teorii pomiaru, trudność zadań wyrażona jest w tej samej skali co poziom badanej umiejętności i może przyjmować wartości zarówno ujemne, jak dodatnie. Oczywiście, ponieważ skala jest przedziałowa, zawsze można ją przekształcić bez uszczerbku dla pomiaru tak, aby poziom umiejętności był wyrażany tylko wartościami dodatnimi.

Pierwszym krokiem analizy, niezależnie od wybranego modelu, jest oszacowanie parametrów zadania (a, b, c).

Wybrany do analizy model może być stosowny lub nie do zastosowania względem zbioru danych empirycznych. Oznacza to, że model może niewłaściwie przewidywać i wyjaśniać wyniki egzaminu. Dlatego jednym z najważniejszych kroków podczas stosowania teorii analizy zadania testowego do oceny wyników egzaminów jest oszacowanie, czy wybraliśmy właściwy model i czy w ogóle możemy zastosować IRT do analizy naszych danych. Procedury szacowania, czy dane empiryczne spełniają wymagania danego modelu, zwykle są integralną częścią programów komputerowych umożliwiających praktycznie stosowanie IRT.

Kiedy model właściwie opisuje dane empiryczne, uzyskujemy opis szeregu istotnych cech pojedynczych zadań, jak i egzaminowanych uczniów, które są pożądane dla pomiaru dydaktycznego. Jeżeli arkusz egzaminacyjny został trafnie przygotowany do danego egzaminu oraz gdy wyniki egzaminu spełniają założenia wybranego modelu, to, po pierwsze, oszacowany poziom umiejętności egzaminowanego jest niezależny od zastosowanego arkusza egzaminacyjnego oraz oszacowane parametry zadania są niezależne od grupy egzaminowanych danym arkuszem egzaminacyjnym. Innymi słowy, poziom umiejętności oszacowany na podstawie różnych zbiorów zadań mierzących tę samą umiejętność jest w granicach błędu pomiarowego taki sam. Po drugie, parametry zadania oszacowane na postawie różnych grup egzaminowanych są takie same w granicach błędu związanego z wyborem próby. Możemy więc powiedzieć, że dla IRT parametry opisujące zadanie, jak i parametry opisujące poziom osiągnięć ucznia są inwariantem (niezmiennikiem). Niezależność parametrów osiągnięć egzaminowanego od parametrów zastosowanych zadań oraz niezależność parametrów zadań od wyboru próby egzaminowanych uczniów jest koronnym atutem teorii odpowiedzi na zadanie testowe. Ta niezależność osiągana jest poprzez wykorzystanie informacji o zadaniach do oszacowania poziomu osiągnięć uczniów i poprzez symultaniczne wykorzystanie informacji o osiągnięciach uczniów do oszacowania parametrów zadań.

3. Metody szacowania parametrów zadań

Procedury szacowania parametrów zadań są stosunkowo złożone i wymagają zaawansowanego oprogramowania komputerowego. Norman D. Verhelst wyróżnia cztery metody leżące u podstaw analiz z wykorzystaniem modeli IRT[3]. Pierwsza z metod, nazwana w skrócie ML – *maksimum prawdopodobieństwa* (ang. *Maximum Likelihood*), polega na takim wyborze parametrów zadań (trudności i mocy różnicującej), że dane empiryczne uzyskane z testu są wartościami

[3] Verhelst N.D., *Manual for relating Language Examinations to CEF*, 2005.

najbardziej prawdopodobnymi jak to jest tylko możliwe. Jakość oszacowania tą metodą parametrów zadań istotnie zależy od wielkości próby uczniów.

Kolejną metodą jest *maksimum łącznego prawdopodobieństwa* – JML (ang. *Joint Maximum Likelihood).* W tej metodzie szacowane są łącznie parametry zadań i poziom umiejętności uczniów. Jeżeli przyjmiemy model Rascha, w którym wszystkie zadania mają taki sam parametr mocy różnicującej, to np. dla 10 uczniów rozwiązujących 10 zadań szacowane będzie łącznie 10 parametrów trudności zadań i 10 parametrów poziomów umiejętności. Wraz ze wzrostem próby szybko rośnie liczba parametrów koniecznych do oszacowania.

Trzecia metoda to *maksimum prawdopodobieństwa brzegowego* – MML (ang. *Marginal Maximum Likelihood*). W metodzie tej zakłada się, że badana cecha ma rozkład normalny w populacji, z której my dysponujemy losową próbą. W takim przypadku szacowane są parametry zadań i dwa parametry rozkładu normalnego w populacji – średnia i odchylenie standardowe. W praktyce to drugie założenie nie jest łatwe do spełnienia.

Czwarta metoda to *warunkowe maksimum prawdopodobieństwa* – CML (ang. *Conditional Maximum Likelihood*). W tej metodzie parametry zadań są szacowane na podstawie założenia, że znane są liczby punktów za test dla każdej osoby i procent poprawnych odpowiedzi za każde zadanie. N. Verhelst ilustruje podstawy tej metody w najprostszym przypadku testu złożonego z dwóch zadań punktowanych 0 i 1. Przeanalizujmy przykładowy rozkład odpowiedzi na te zadania.

Tab. 1. Częstość odpowiedzi na zadania[4]

		zadanie 1.		
		1	0	razem w zadaniu 2.
zadanie 2.	1	70 (2 pkt za test)	30 (1 pkt za test)	100
	0	80 (1 pkt za test)	120 (0 pkt za test)	200
	Razem w zadaniu 1.	150	150	300

Jak można odczytać z tabeli, 150 osób, czyli 50% próby odpowiedziało poprawnie na zadanie 1. Na zadanie 2. poprawnie odpowiedział co trzeci uczeń. Czyli zadanie 2. jest trudniejsze od zadania 1. Do takiej samej konkluzji można dojść w inny sposób, analizując częstość odpowiedzi na obydwa zadania uczniów, którzy za dwuzadaniowy test uzyskali tyle samo punktów, czyli 1. Ponieważ wśród uczniów z wynikiem 1 pkt jest mniej takich, którzy uzyskali ten wynik za zadanie 2. niż za zadanie 1, wnioskujemy, że zadanie 2. jest względnie trudniejsze od zadania 1. Wraz ze wzrostem liczby zadań tego typu względne porównania stają się bardziej złożone.

[4] Ibidem.

Zaletą tej metody jest niezależność oszacowania od tego, jak skomponowana jest próba uczniów do kalibracji. Np. rozdzielne próby uczniów mogą rozwiązywać zamiast całego testu tylko jego części (które nazwiemy podtestami) zawierające wspólne zadania. Również nie jest istotne, czy próba jest losowa, czy też nie. Cecha ta jest często nazywana niezależnością od próby uczniów, na której zastosowany jest test do kalibracji zadań. Ważne jest jednak, aby w próbie znaleźli się zarówno uczniowie, dla których każde zadanie jest bardzo łatwe, jak i tacy, dla których każde zadanie jest bardzo trudne. Ograniczeniem jest to, że nie wszystkie modele IRT mogą być przy tej metodzie zastosowane. Metoda ta może być stosowana tylko w jednoparametrycznym modelu Rascha, gdzie liczba punktów jest po prosu równa liczbie poprawnie udzielonych odpowiedzi. W modelu dwuparametrycznym liczba punktów dla ucznia jest ważoną sumą, a wagami są parametry mocy różnicującej zadań a_i.

Rozwiązaniem tego problemu jest zastosowanie jednoparametrycznego logistycznego modelu OPLM (ang. *One Parameter Logistic Model*). Formalnie jest to model do oszacowania dwóch parametrów, przy czym różne parametry mocy różnicującej dla poszczególnych zadań są znane (oszacowane uprzednio) i w modelu traktowane jako stałe, a nie jako parametry. W ten sposób dwuparametryczny model sprowadzany jest do jednego parametru, umożliwiając zastosowanie metody warunkowego maksimum prawdopodobieństwa CML. To rozwiązanie zostało wykorzystane podczas szacowania parametrów zadań połączonego testu zawierającego zadania ze sprawdzianów 2003, 2004 i 2005 r.

3. Zrównywanie wyników sprawdzianu 2003, 2004 i 2005

Wiele instytucji zajmujących się testowaniem na szeroką skalę wykorzystuje modele IRT do budowania banków zadań, tworzenia testów. W tych instytucjach wykorzystuje się IRT także do zrównywania testów. W obydwu wymienionych procesach pierwszy etap, czyli szacowanie i kalibrowanie parametrów zadań, jest wspólny. Dlatego też zrównywanie wyników stało się przyczynkiem do stworzenia zalążku banku zadań z wykorzystaniem programu komputerowego *FastTest.*

4. Procedura wykorzystana do zrównywania wyników sprawdzianów 2003–2005

Zastosowana procedura składa się z dwóch etapów, które mogą być przeprowadzone niezależnie. Pierwszy etap to oszacowanie parametrów zadań zastosowanych w sprawdzianach 2003, 2004 i 2005 r. z wykorzystaniem testek kotwiczących 2004 i 2005. Testka kotwicząca 2004 była tak dobrana, aby dobrze reprezentowała sprawdziany 2003 i 2004 r., natomiast testka 2005 – sprawdziany 2003 i 2005 r.

Rys. 4. Procedura zrównywania wyników sprawdzianu 2005 i 2004 r. do 2003 r.

Drugi etap to oszacowanie różnic średniej wartości poziomu osiągnięć dla populacji szóstoklasistów w kolejnych latach: 2003, 2004 i 2005. Do analizy wybrano jednakowe próby losowe uczniów, każda po 10 000 przypadków, z populacji szóstoklasistów piszących w każdym z analizowanych lat arkusz standardowy A1. Na tym etapie analizy sprawdziany z poszczególnych lat stanowiły podtesty jednego połączonego testu, którego parametry zadań zostały wykalibrowane w ten sposób, aby średnia połączonego testu była równa 0. W zastosowanej analizie regresji zmienną niezależną była wielkość mająca trzy kategorie określające kolejno populacje szóstoklasistów 2003, 2004 i 2005 r.

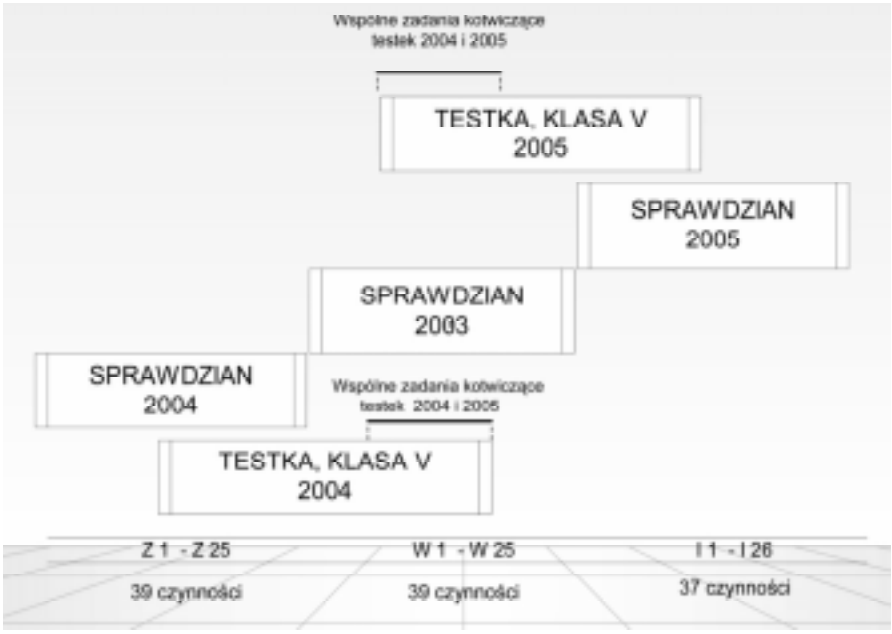
5. Pierwszy etap analizy – szacowanie parametrów zadań

Do oszacowania parametrów zadań tworzących sprawdziany na zakończenie klas VI w trzech kolejnych latach wykorzystano próby celowe uczniów klas V (2004 i 2005 r.) piszących testki kotwiczące oraz próby uczniów wylosowane z populacji piszących arkusz standardowy A1 w trzech kolejnych latach. Do powiązania parametrów zadań zastosowano dwie testki kotwiczące zastosowane wobec prób celowych uczniów klasy V w 2004 i w 2005 r. Obydwie testki zawierały również wspólne zadania kotwiczące. W części reprezentującej sprawdzian 2003 r. różniły się tylko 1 na 15 zadań.

W zadaniach rozwiniętej odpowiedzi, wymagających wykonania wielu czynności, jako zadanie testowe zdefiniujemy również czynność, która wymagała odrębnych działań i dla której określone były kryteria punktowania. W ten sposób np. dla zadania 22. w sprawdzianie z 2003 r. szacowane było 5 parametrów mocy różnicującej i 5 parametrów trudności, czyli dla każdej czynności, która była oceniana w uczniowskim rozwiązaniu. Przy takim rozumieniu zadań sprawdziany z 2003 i 2004 r. zawierały po 39 zadań, a sprawdzian z 2005 r. – 37 zadań. Testki kotwiczące zapewniające powiązanie ze sobą sprawdzianów zawierały odpowiednio po 54 i 55 zadań.

Sprawdziany z trzech kolejnych lat łącznie tworzą narzędzie do sprawdzania 115 czynności, dając możliwość uzyskania maksymalnie 120 pkt. Zgodnie z przyjętą na użytek tego artykułu definicją powiemy, że obejmują 115 zadań. Każdy z uczniów, którego wyniki brane były pod uwagę, rozwiązywał tylko część z tych zadań, zawartych w jednym z pięciu arkuszy sprawdzianów lub

testek. Zarówno sprawdziany z kolejnych lat, jak i testki 2004 i 2005 możemy wspólnie nazywać podtestami dużego testu obejmującego zadania z trzech sprawdzianów.



Rysunek 5. Struktura połączonego z trzech sprawdzianów testu z wykorzystaniem dwóch testek kotwiczących, zastosowanych do zrównywania

Każdy z podtestów był rozwiązywany przez różne grupy uczniów. Każdy uczeń miał wyniki, jakie otrzymał w swoim podteście oraz opuszczenia dla zadań z podtestów, w których nie uczestniczył.

W tab. 2. przedstawiona jest struktura odpowiedzi uczniowskich na test złożony z zadań zawartych w arkuszach sprawdzianów 2003, 2004 i 2005 r. Uczniowie, którzy pisali sprawdzian w 2003 r. mają wyniki dla 39 pierwszych zadań, podczas gdy dla pozostałych 76 mają opuszczenia (ang. *missing*) zaznaczone cyfrą 9. Szóstoklasiści, którzy pisali sprawdzian w 2004 r. mają opuszczenia w pierwszych 39 zadaniach, ponieważ zadania sprawdzianu 2004 r. zaczynają się od pozycji 40. do 78., dla których występują w zbiorze danych ich odpowiedzi: 0, 1 lub 2. Również dla zadań od 79. do 115. widzimy cyfry 9 symbolizujące opuszczenia. Piątoklasiści z 2004 r., którzy pisali test reprezentujący zadania z obydwu sprawdzianów, mają 0, 1 lub 2 dla zadań w ich teście, a dla pozostałych opuszczenia zaznaczone dziewiątkami. Podobnie w rekordach danych ze sprawdzianu 2005 r. dla pierwszych 78 zadań (zadania sprawdzianu 2003 i 2004 r.) obserwujemy opuszczenia oznaczone cyfrą 9 i wartości symbolizujące kody odpowiedzi dla ostatnich 39 zadań. W rekordach zawierających zapis odpowiedzi piątoklasistów z 2005 r. obserwujemy kody odpowiedzi dla zadań testki z 2005 r. (podtesty 2003 i 2005) i symbole opuszczeń dla pozostałych zadań.

Tab. 2. Struktura odpowiedzi uczniowskich na zadania połączonego testu

zad.	123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345
Sprawdzian 2003	01100110110101010100000000011000001010099999999999999999999999999999999999999999999999999999999999999999999999999999999999999999
Sprawdzian 2003	10001000011011110010000000110100000010999999999999999999999999999999999999999999999999999999999999999999999999999999999999999999
Testka kl. V 04	99999999991111001001000110111110000000110000000000999999999999999999999999999999999999999999999999999999999999999999999999999999
Testka kl. V 04	99999999991111001001000110111110000000110000000000999999999999999999999999999999999999999999999999999999999999999999999999999999
Sprawdzian 2004	99999999999999999999999999999999999991110101001010101011100010010000100000119999999999999999999999999999999999999
Sprawdzian 2004	99999999999999999999999999999999999991100100010011100010000000001101000101009999999999999999999999999999999999999
Sprawdzian 2005	99999999999999999999999999999999999999999999999999999999999999999999999999999999999999100100000111001101101000000100100000
Sprawdzian 2005	99999999999999999999999999999999999999999999999999999999999999999999999999999999999999100100000000110101112100000110100000 1
Testka kl. V 05	99999999999990111111111111111111001001999999999999999999999999999999999999999999999999999900011011110000000110000900009000
Testka kl. V 05	99999999999999111111111111111111001100019999999999999999999999999999999999999999999999999999999900000010100010000000000000000009

Z tak zestawionej struktury danych niekompletnych (zawierających opuszczenia dla każdego ucznia) postaramy się oszacować parametry (moc różnicującą i trudność) wszystkich zadań tworzących 115-zadaniowy test.

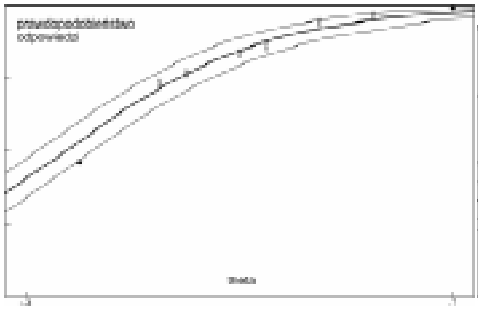
Tab. 3. Charakterystyka prób zastosowanych do szacowania parametrów zadań

populacja	rodzaj próby	wielkość próby	zastosowane narzędzia	numery zadań w połączonym teście
szóstoklasiści 2003	losowa	1836	sprawdzian 2003	1-39
szóstoklasiści 2004	losowa	1836	sprawdzian 2004	40-78
piątoklasiści 2004	losowa	1836	testka 2003-2004	1-5,8,10-12,14,15,17, 20, 22-26,28-37, 41-43, 45-47, 52-54,56-67,73-77
szóstoklasiści 2005	losowa	1836	sprawdzian 2005	79-1016
piątoklasiści 2005	losowa	1836	testka 2003-2005	1-5, 7, 10-12,14,15, 17,20,22-26,28-37, 80-84,87,88,90,91,93,94,96, 98-101,103-107,109-114

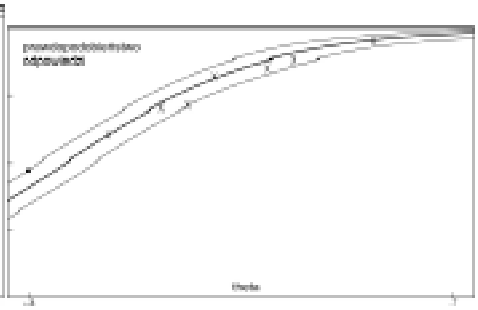
Do jednoczesnego oszacowania parametrów zadań i umiejętności uczniów zastosowano wspomniany już wcześniej program komputerowy OPLM. W pierwszym kroku oszacowano wstępne parametry zadań, zakładając stałą i równą 1 moc różnicującą zadań ($a = 1$). W drugim kroku oszacowano parametry mocy różnicującej zadań tak, aby wartości empiryczne, proporcje poprawnych odpowiedzi dla 8 wyróżnionych ze względu na poziom umiejętności grup uczniów były najbardziej zgodne z krzywymi charakterystycznymi zadań opisującymi prawdopodobieństwo poprawnej odpowiedzi. W kolejnym kroku dokonano analizy krzywych charakterystycznych dla poszczególnych zadań kotwiczących dla populacji szóstoklasistów i uczniów klas V. Rys. 6. przedstawia krzywe charakterystyczne zadań wraz z zaznaczeniem punktów empirycznych.

Przyjrzyjmy się teraz przykładom teoretycznych i empirycznych krzywych charakterystycznych dla zadań wspólnych zastosowanych w modelu. Na osi poziomej wykresów przedstawiona jest skala θ (theta) wspólna dla trudności zadań i poziomu umiejętności uczniów. Na osi pionowej mamy prawdopodobieństwo poprawnej odpowiedzi na zadanie – od 0 do 1. Pogrubiona krzywa przedstawia oszacowane na podstawie modelu CML prawdopodobieństwo udzielenia po-

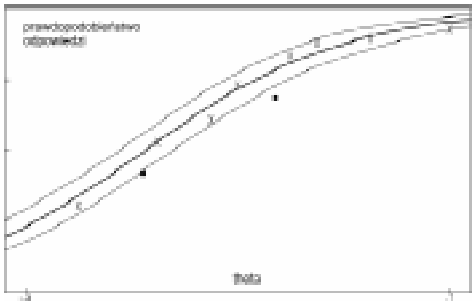
prawnej odpowiedzi przez uczniów o danym poziomie umiejętności badanych sprawdzianem. Krzywe zaznaczone jaśniejszą linią wyznaczają obszar 95-procentowego poziomu ufności. Punkty zaznaczone na wykresach przedstawiają dane empiryczne (zaobserwowana trudność zadań w 8 grupach, na które ze względu na poziom umiejętności została podzielona próba uczniów). Jak można zauważyć na podstawie rysunku, rozkład poziomu umiejętności piątoklasistów badanych tymi zadaniami przesunięty jest w kierunku niższych umiejętności w stosunku do uczniów klasy VI.



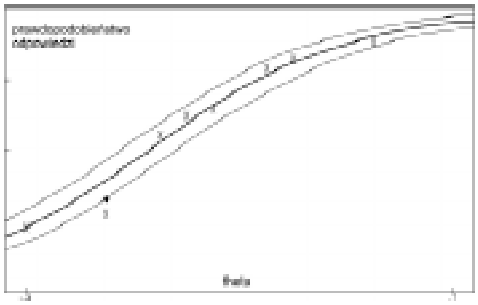
Populacja szóstoklasistów 2003 – zadanie 10. (10)



Populacja piątoklasistów 2004 – zadanie 10. (10)



Populacja szóstoklasistów 2004 – zadanie 17. (58)

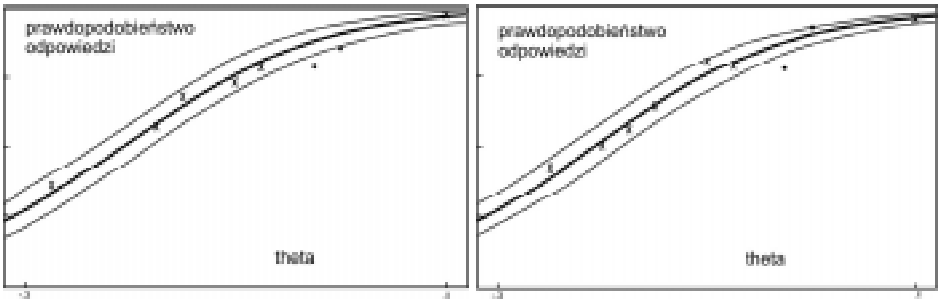


Populacja piątoklasistów 2004 – zadanie 17, (58)

Rys. 6. Krzywe charakterystyczne zadania 10. ze sprawdzianu 2003 r. i zadania 17. ze sprawdzianu 2004 r. Jaśniejszymi liniami zaznaczono obszar 95-procentowego prawdopodobieństwa zgodności wyników empirycznych z krzywą teoretyczną

Jednak podstawową informacją, którą możemy odczytać z wykresów jest dobre dopasowanie modelu do opisu zaobserwowanych danych empirycznych. Nie wszystkie zadania zastosowane do powiązania arkuszy z trzech sprawdzianów tak dobrze opisywane są przez model. Dla większości zadań dopasowanie modelu jest wystarczające, aby nie usuwać ich z analizy, szczególnie że model nie jest zbyt czuły na niewielkie odstępstwa od takiej samej mocy różnicującej zadań w populacji klas V i VI.

Zadanie 15. w sprawdzianie 2005 r. jest zadaniem wielokrotnego wyboru. Jest wielce prawdopodobne, że uczniowie o najniższym poziomie umiejętności uzyskali wyższy wynik od przewidywanego na skutek zgadywania.



Populacja szóstoklasistów 2005 – zadanie 15. (93)

Populacja piątoklasistów 2005 – zadanie 15. (93)

Rys. 7. Przykład zadania, w którym piątoklasiści o najniższym poziomie umiejętności osiągnęli znacznie wyższy wynik od przewidywanego na podstawie zastosowanego modelu

Jeżeli zgodzimy się, że zastosowany model analizy dobrze opisuje dane empiryczne dla prób z 5 populacji uczniów biorących udział w zrównywaniu, to możemy przyjrzeć się wynikom testu utworzonego z arkuszy 2003, 2004 i 2005. Rys. 7. przedstawia relacje pomiędzy skalą wyników surowych od 0 do 120 a rezultatem wyrażonym w przedziałowej skali theta ze średnią 0. Skali, która jest wspólna dla poziomu umiejętności uczniów i dla trudności zadań.

W ostatnim, czwartym kroku tej analizy, po ustaleniu parametrów mocy różnicującej zastosowano metodę CML (warunkowe maksimum prawdopodobieństwa) do oszacowania trudności zadań.

Tab. 4. Oszacowane parametry zadań – wskaźnik mocy różnicującej *a* i trudności zadań *b* (przy szacowaniu parametrów zadań wskaźnik mocy różnicującej został unormowany do średniej geometrycznej g = 4,007)

Identyfikator zadania		Parametry zadań – trudność i moc różnicująca		Przeskalowane parametry zadań	
łącznie	w sprawdzianie	a	b	a/g	b*g
1	w_1	4	-0.273	0.998	-1.093
2	w_2	3	0.003	0.749	0.013
3	w_3	5	-0.266	1.248	-1.066
.....					
112	i_25_4	5	0.411	1.248	1.645
113	i_25_5	6	0.166	1.497	0.667
114	i_25_6	6	0.143	1.497	0.573

Podsumowanie pierwszego etapu analizy

Analiza rezultatów zamieszczonych w poniższej tabeli pozwala zauważyć niższą wartość rzetelności sprawdzianu 2005 r. w stosunku do sprawdzianu w poprzednich latach. Można próbować to wyjaśnić efektem pułapu spowodowa-

nym faktem, że test był w tym roku łatwiejszy w stosunku do poprzednich lat, co zmniejszyło wariancję wyników.

Tab. 5. Charakterystyka wyników uzyskanych za pomocą testów i testek zastosowanych do zrównania

Opis statystyczny	Sprawdzian			Uczniowie klas V	
	2003	2004	2005	2004	2005
Liczba obserwacji	1836	1836	1836	1836	1836
Średnia dla theta	0,295	0,321	0,327	0,153	0,156
Wariancja dla oszacowanej wartości theta	0,106	0,103	0,090	0,084	0,072
Wariancja dla prawdziwej wartości theta	0,087	0,90	0,071	0,77	0,065
Rzetelność	0,825	0,878	0,792	0,918	0,904
Korelacja punktów surowych i oszacowanej wartości theta	0,953	0,985	0,942	0,992	0,978

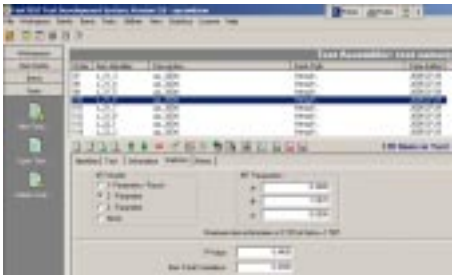
Zastosowany model pomiaru pozwolił oszacować relacje pomiędzy zadaniami zastosowanymi do pomiaru a nieobserwowalną zmienną $\underline{\theta}$, jaką jest poziom umiejętności badanych w sprawdzianie.

6. Budowanie banku zadań dla sprawdzianu

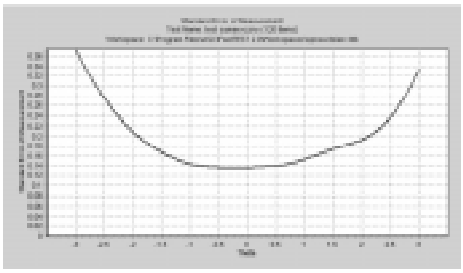
Wykalibrowane parametry zadań połączonego sprawdzianu z trzech lat zostały wprowadzone do banku zadań z wykorzystaniem oprogramowania komputerowego *FastTest* opracowanego przez Assessment System Corporation. Na strukturę banku składa się treść zadania oraz metryczka każdego zadania, która obejmuje: identyfikator zadania (przyjęto zgodnie z oznaczeniami w połączonym teście – por. rys. 4.), autora (przyjęto CKE), opis typu zadania, parametry zadania zgodne z IRT (trudność *b* i moc różnicującą *a*), parametry zadań zgodne z klasyczną teorią testu (łatwość *p*, moc różnicującą wyrażoną wskaźnikiem r_{pb}). Bank może być systematycznie rozbudowywany poprzez empiryczne zastosowanie testów, których zadania kotwiczące mogą stanowić już wykalibrowane zadania ze sprawdzianów 2003–2005.

Konstruując test z wykalibrowanych zadań z banku, można lepiej niż to ma miejsce w dotychczasowej praktyce przewidzieć parametry kolejnych sprawdzianów. Rozbudowanie banku zadań do rozmiarów, które pozwolą na praktyczne wykorzystanie go do tworzenia nowych arkuszy, wymaga systematycznego budowania i przeprowadzania testów w podobny sposób jak miało to miejsce w przypadku testek 2004 i 2005. Jest to duże i czasochłonne przedsięwzięcie, które jednak już w najbliższym czasie powinno stać się stałą praktyką w centralnej i okręgowych komisjach egzaminacyjnych. Znajomość kształtu krzywej informacyjnej dla arkusza egzaminacyjnego ma szczególne znaczenie na poziomie

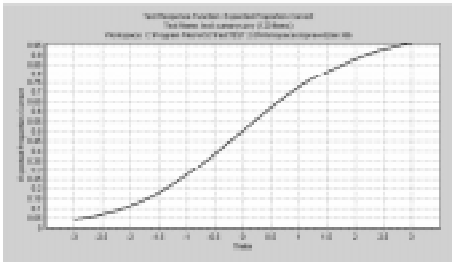
egzaminów maturalnych[5]. Powinna ona osiągać swoje maksimum w arkuszach na poziom podstawowy dla θ odpowiadającego progowi zaliczenia (30% punktów), ponieważ błąd standardowy pomiaru osiąga wtedy swoje minimum.



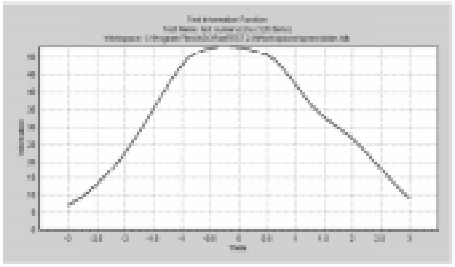
Metryczka zadania w banku



Przewidywana wielkość błędu standardowego dla wyników uczniów o różnym poziomie umiejętności piszących połączony test



Zależność pomiędzy poziomem umiejętności a spodziewanym procentem punktów za połączony test



Krzywa informacyjna wygenerowana dla połączonego testu

Rys. 8. Charakterystyki testu wygenerowanego z wykalibrowanych zadań znajdujących się w banku

7. Oszacowanie zmian poziomu osiągnięć szóstoklasistów na podstawie wyników sprawdzianu 2003–2005

Aby ustalić różnice pomiędzy wskaźnikami osiągnięć uczniów mierzonych sprawdzianem, w kolejnych latach zastosowano strukturalną analizę regresji zmiennej latenntnej (ang. *Structural Analysis of Univariate Latent Variable*) i programu komputerowego SAUL opracowanego w CITO przez N.D. Verhelsta i H.H.F.M. Verstralen.

Zastosowany poprzednio model pomiaru pozwolił oszacować relacje pomiędzy zadaniami zastosowanymi do pomiaru a nieobserwowalną zmienną θ,

[5] Szaleniec H., *Krzywa informacyjna zadań jako narzędzie w konstruowaniu arkusza egzaminacyjnego*, [w:] B. Niemierko, J. Brzdąk (red.), *Dwa rodzaje oceniania szkolnego. Ocenianie wewnątrzszkolne i zewnętrzne a jakość szkoły,* materiały VII ogólnopolskiej konferencji z cyklu „Diagnostyka Edukacyjna", Katowice 2002.

jaką jest poziom umiejętności badanych w sprawdzianie. Model strukturalny pozwala na oszacowanie relacji pomiędzy zmienną θ a innymi zmiennymi, takimi jak np.: populacje uczniów kończących szkołę podstawową, lokalizacja szkół, wielkość szkół, stosowane programy nauczania itp. Aby oszacować relację pomiędzy poziomem osiągnięć badanych sprawdzianem (zmienna zależna θ) a zmienną niezależną, jaką są populacje szóstoklasistów w kolejnych trzech latach, potrzebujemy na wejściu dwóch zbiorów danych:

1) parametrów zadań zastosowanych w kolejnych latach, wyrażonych jako wynik prawdziwy (wynik zaobserwowany – błąd pomiaru), przedstawionych w tej samej skali,
2) liczb punktów uzyskanych przez poszczególnych uczniów za te zadania w kolejnych populacjach.

Mocną stroną takiego podejścia jest możliwość zastosowania oszacowanych (łącznie dla arkuszy 2003, 2004, 2005) parametrów wszystkich zadań. Dzięki temu, że model pomiarowy i model strukturalny są rozseparowane, próba stosowana do oszacowania relacji pomiędzy poziomem umiejętności szóstoklasistów a trzema kategoriami zmiennej (populacja 2003, 2004, 2005) może się różnić od próby stosowanej do kalibracji zadań. Np. można zastosować wyniki nie tylko z próby, ale i z całej populacji. Można też wybrać do analiz wyniki jednego województwa czy też jednej komisji egzaminacyjnej.

Ujemną stroną takiego podejścia jest ignorowanie błędu oszacowania parametrów zadań, gdyż parametry te są przekazywane w postaci wyniku prawdziwego oszacowanego w programie OPLM.

Do analizy wybrano krajową próbę losową równą po 30 000 uczniów z połączonej populacji szóstoklasistów piszących arkusz standardowy A1 w latach 2003, 2004 i 2005.

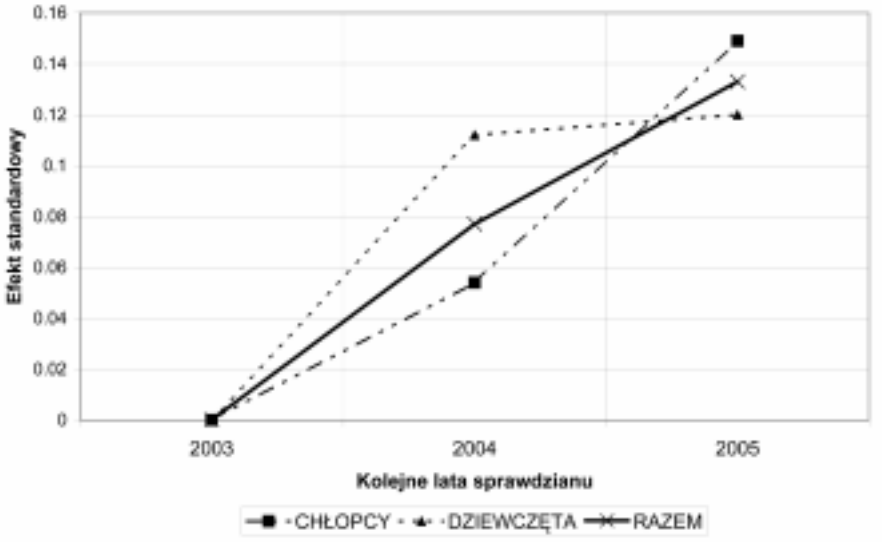
Wszystkie efekty są komunikowane w takiej skali, że średnia geometryczna mocy różnicującej jest równa 1.

stała (*additive parameter*) = 1,1556 (SE = 0,0127; z = 91,153)
wariancja sprawdzianu 2003 (*reference variance*) = 1,4320 (SE = 0,0141)
odchylenie standardowe 2003 (*ref. stand. dev*.) = 1,1966 (SE = 0,0059)

Tab. 6. Efekt główny dla zmiennej populacja sprawdzianu

Populacja sprawdzianu	Efekt	SE	n	z	Wielkość efektu standardowego
2003	ustalony na 0	-	10 331	-	0
2004	0,092	0,018	10 174	5,190	0,077
2005	0,160	0,018	9495	8,760	0,133
Kontrast parami					
2005–2004	0,067	0,018		3,718	0,056

Do oszacowania istotności różnic poziomu osiągnięć w kolejnych latach zastosowano test t – Studenta. Obydwie różnice w stosunku do 2003 r. są istotne na poziomie istotności $\theta = 0{,}01$. Różnice pomiędzy latami 2004 i 2005 jest istotna dla $\theta = 0{,}05$.



Rys. 9. Efekt standardowy przyrostu umiejętności mierzonych sprawdzianem w latach 2003–2005 z uwzględnieniem rozwarstwienia ze względu na płeć

Pozostaje nam teraz odpowiedzieć na pytanie, czy te różnice są znaczące. Do tego potrzebna jest nam jednostka pomiarowa mierzonego efektu. W tym przypadku za jednostkę pomiarową przyjęto odchylenie standardowe analizowanej zmiennej podstawowej, które wynosi 1,1966. Ostatnia kolumna tab. 6. podaje wielkość efektu standardowego dla populacji 2004 i 2005.

Podsumowanie

Oszacowana różnica osiągnięć badanych sprawdzianem pomiędzy 2003 i 2004 r. wynosi prawie 7,7% odchylenia standardowego sprawdzianu 2003 r., a pomiędzy 2004 i 2005 r. – 5,6% odchylenia standardowego. Oszacowany przyrost poziomu osiągnięć (efekt standardowy) pomiędzy populacjami 2003 i 2005 to 13,3% odchylenia standardowego**.** Czy jest to duży przyrost? Trudno powiedzieć. W kraju nie mamy żadnego układu odniesienia. Wynik ten jest wysoce zbieżny z rezultatem, jaki otrzymał prof. Niemierko, stosując ekwicentylową metodę zrównywania[6]. Z wnioskowaniem o stałym trendzie w skali całego kraju musimy być jednak bardzo ostrożni. Dysponujemy wynikami dopiero z czterech lat. Analizując zmiany poziomu osiągnięć szóstoklasistów w dwóch warstwach

[6] Niemierko B., *Zrównywanie wyników sprawdzianu 2005...*, op. cit.

(chłopcy – dziewczęta), zaobserwowano istotne różnice w dynamice oszacowanych zmian osiągnięć w poszczególnych warstwach. Dla chłopców zaobserwowano słabszy przyrost pomiędzy 2003 i 2004 r. – efekt standardowy wynosi 5,4% – i większy pomiędzy 2004 i 2005 r. (9,4%). Odpowiednio dla dziewcząt oszacowany przyrost wynosi 11,2% pomiędzy 2003 i 2004 r. Pomiędzy sprawdzianami 2004 i 2005 r. oszacowana różnica dla dziewcząt wynosi 0,8% i jest nieistotna statystycznie. W rozważanym okresie dziewczęta osiągały wyższe wyniki i oszacowany efekt standardowy porównania wyników w warstwach wynosi 28,7%. Problematyka ta wymaga dalszych pogłębionych badań, jak i rozszerzenia ich na poziom gimnazjalny. Zastosowanie metod bazujących na probabilistycznej teorii zadania daje duże możliwości prowadzenia pogłębionych analiz, które mogą istotnie uzupełniać analizy prowadzone z wykorzystaniem klasycznej teorii. Wymagają one jednak bardziej rygorystycznego przestrzegania wielu założeń, które leżą u podstaw konstruowania arkuszy egzaminacyjnych.