Bartosz Kondratek
Centralna Komisja Egzaminacyjna
Polskie Towarzystwo Diagnostyki Edukacyjnej

Wykorzystanie modelowania IRT do statystycznej kontroli niereprezentatywności próby badawczej przy wyznaczaniu właściwości psychometrycznych zadań egzaminacyjnych

Wprowadzenie

W 2015 roku wprowadzono w Centralnej Komisji Egzaminacyjnej (CKE) modyfikację dotychczasowej procedury próbnego zastosowania zadań (pilotażu) z egzaminu gimnazjalnego. Jej celem była próba statystycznej kontroli następujących czynników zaburzających ocenę właściwości psychometrycznych zadań:

1. brak reprezentatywności prób standaryzacyjnych (zregionalizowany i celowy dobór szkół i uczniów);
2. potencjalnie obniżona motywacja uczniów związana z badaniem w sytuacji nieegzaminacyjnej (test niskiej stawki).

Wspomniana modyfikacja polegała na włączeniu do arkuszy testowych zawierających nowe zadania, próbnie stosowane przez okręgowe komisje egzaminacyjne (OKE), zestawu zadań kotwiczących wybranych spośród zadań wykorzystanych na egzaminach we wcześniejszych latach. Dzięki temu, że uczniowie rozwiązujący arkusze pilotażowe udzielali jednocześnie odpowiedzi na kilka zadań pochodzących z wcześniejszych egzaminów, możliwe było oszacowanie sposobu, w jaki uczniowie podchodzący do egzaminu we wcześniejszych latach odpowiadaliby na zadania poddane próbnemu zastosowaniu. Dzięki takiemu schematowi badania oraz zastosowaniu technik modelowania IRT (*Item Response Theory*) właściwości psychometryczne zadań pilotowanych w odmiennych arkuszach i w odmiennych grupach uczniów mogły być przedstawione na wspólnej skali uczniów podchodzących do egzaminu. Wyniki próbnego zastosowania były raportowane ekspertom przedmiotowym, układającym egzamin, w postaci arkusza kalkulacyjnego, który oprócz przedstawienia populacyjnych oszacowań właściwości zadań, szacował także właściwości całego zestawu zadań, ułożonego z wykorzystaniem dowolnego podzbioru zadań poddanych próbnemu zastosowaniu.

Taka procedura stanowi adaptację „wariantu B" przeprowadzania próbnego zastosowania zadań, który został zarekomendowany w raporcie *Porównywalne wyniki egzaminacyjne*, podsumowującym prace projektu zajmującego się zrównywaniem wyników egzaminów zewnętrznych, prowadzonego w Instytucie Badań Edukacyjnych w latach 2010–2015 (Szaleniec i in., 2015, s. 197–199). Istotną zmianą w porównaniu z proponowanym w raporcie schematem badawczym jest wprowadzenie zadań kotwiczących do wcześniejszych egzaminów.

Artykuł stawia sobie dwa główne cele:

1. przybliżenie procedury zbierania danych oraz przewidywania egzaminacyjnych charakterystyk zadań oraz testu złożonego z dowolnego podzbioru zadań z wykorzystaniem modelowania IRT;
2. ocenę jakości przewidywanych charakterystyk zadań poprzez zestawienie ich z wartościami faktycznie uzyskanymi na egzaminie.

Jako przykład wykorzystane zostaną zadania matematyczne z egzaminu gimnazjalnego z roku 2017 oraz 2018. Wszystkie analizy IRT zostały przeprowadzone z wykorzystaniem oprogramowania UIRT działającego w środowisku Stata (Kondratek, 2016).

Dobór i właściwości zadań kotwiczących

Próbne zastosowanie zadań z matematyki na egzamin gimnazjalny w 2017 oraz 2018 roku zostało przeprowadzone według tożsamych procedur, określonych w kolejnych zarządzeniach dyrektora CKE.

Tabela 1. Schemat rozmieszczenia zadań kotwiczących w arkuszach na próbnym zastosowaniu

	Zadanie kotwiczące	Pozycja zadania kotwiczącego w arkuszu pilotażowym					
Próbne zastosowanie na rok 2017		Arkusz M_17_1	Arkusz M_17_2	Arkusz M_17_3	Arkusz M_17_4	Arkusz M_17_5	Arkusz M_17_6
	gm_m_2014_z04	8	2	4	6	2	5
	gm_m_2014_z07	10	11	8	2	4	3
	gm_m_2015_z09	11	5	11	13	7	9
	gm_m_2014_z14	13	15	15	11	12	13
	gm_m_2013_z15	19	18	17	17	17	16
	gm_m_2014_z20	21	20	20	19	19	20
Próbne zastosowanie na rok 2018	Zadanie kotwiczące	Pozycja zadania kotwiczącego w arkuszu pilotażowym					
		Arkusz M_18_1	Arkusz M_18_2	Arkusz M_18_3	Arkusz M_18_4	Arkusz M_18_5	Arkusz M_18_6
	gm_m_2014_z04	8	2	4	6	2	5
	gm_m_2014_z07	10	11	8	2	4	3
	gm_m_2015_z09	11	5	11	13	7	9
	gm_m_2014_z14	13	15	15	11	12	13
	gm_m_2013_z15	19	18	17	17	17	16
	gm_m_2014_z20	21	20	20	19	19	20

W sześciu różnych OKE powstały odmienne arkusze matematyczne zawierające zadania – kandydatów do włączenia do przyszłego egzaminu gimnazjalnego. W każdym z arkuszy umieszczonych zostało sześć takich samych[1]

[1] Zastosowanie tych samych zadań w różnych arkuszach nie było podyktowane wymogami użytej procedury statystycznej. W różnych arkuszach można było wykorzystać różne zadania kotwiczące do poprzednich egzaminów, i tak też postąpiono np. w przypadku j. angielskiego, gdzie treść pozostałych zadań arkusza w większym stopniu ograniczała dowolność doboru zadań kotwiczących.

kotwic, które pochodziły z egzaminu z roku 2013 (z15), 2014 (z04, z07, z14, z20) oraz 2015 (z09). Zadania kotwiczące dobrano, kierując się ich treścią oraz właściwościami psychometrycznymi. Przesłanka treściowa zakładała, że zbiór zadań kotwiczących będzie pokrywał możliwie zróżnicowany zakres badanych umiejętności. Analizując właściwości psychometryczne, starano się wybrać zadania o możliwie umiarkowanej trudności oraz wysoko dyskryminujące, czyli zadania, które potencjalnie dostarczały jak najwięcej informacji o poziomie umiejętności uczniów. Było to istotne ze względu na niewielką liczbę zadań kotwiczących oraz przyjęte założenie, że będą to zadania zamknięte. Jak widać w tabeli 1, umiejscawiając zadania kotwiczące w pilotażowych arkuszach, starano się zachować ich relatywną pozycję z oryginalnego arkusza egzaminacyjnego, aby zminimalizować potencjalnie występujące efekty kolejności.

W tabeli 2 zestawiono łatwości sześciu zadań kotwiczących uzyskane na prawdziwym egzaminie z łatwościami, jakie zaobserwowano dla tych samych zadań w każdym z 12 pilotażowych arkuszy.

Tabela 2. Łatwość zadań kotwiczących zaobserwowana na egzaminie oraz w 12 grupach uczniów rozwiązujących arkusze na próbnym zastosowaniu; przypadki, gdy zadanie było trudniejsze niż na egzaminie, zostały wyróżnione

	Zadanie kotwiczące	Łatwości zadań zaobserwowane w różnych grupach uczniów						
		egzamin	M_17_1 (N=283)	M_17_2 (N=294)	M_17_3 (N=282)	M_17_4 (N=288)	M_17_5 (N=300)	M_17_6 (N=275)
Próbne zastosowanie na rok 2017	gm_m_2014_z04	50,7%	66,4%	57,8%	68,4%	60,1%	66,0%	73,5%
	gm_m_2014_z07	47,4%	66,4%	58,8%	59,6%	60,4%	60,3%	71,3%
	gm_m_2015_z09	50,9%	68,2%	57,1%	61,7%	54,9%	57,0%	70,5%
	gm_m_2014_z14	55,7%	67,8%	57,5%	59,6%	67,0%	61,7%	73,8%
	gm_m_2013_z15	50,8%	66,4%	57,8%	59,2%	59,4%	59,3%	74,9%
	gm_m_2014_z20	54,3%	62,2%	**53,1%**	**51,4%**	60,4%	57,0%	69,8%
	Zadanie kotwiczące	Łatwości zadań zaobserwowane w różnych grupach uczniów						
		egzamin	M_18_1 (N=319)	M_18_2 (N=309)	M_18_3 (N=334)	M_18_4 (N=320)	M_18_5 (N=321)	M_18_6 (N=310)
Próbne zastosowanie na rok 2018	gm_m_2014_z04	50,7%	73,7%	74,1%	58,4%	58,8%	66,7%	63,5%
	gm_m_2014_z07	47,4%	69,6%	61,8%	59,3%	53,8%	60,4%	61,9%
	gm_m_2015_z09	50,9%	68,0%	65,0%	53,9%	54,1%	66,7%	61,6%
	gm_m_2014_z14	55,7%	69,3%	74,4%	61,1%	**53,8%**	66,4%	61,6%
	gm_m_2013_z15	50,8%	77,1%	75,7%	60,8%	57,2%	62,9%	57,7%
	gm_m_2014_z20	54,3%	58,9%	60,5%	**53,9%**	**51,6%**	60,7%	57,7%

Analizując łatwości zebrane w tabeli 2, można zauważyć dwie rzeczy. Po pierwsze, uczniowie w każdej z 12 grup radzili sobie zasadniczo lepiej od uczniów na prawdziwym egzaminie (oprócz pięciu komórek tabeli wszystkie odsetki poprawnych rozwiązań są wyższe niż na egzaminie). Po drugie, grupy uczniów rozwiązujące odmienne arkusze na próbnym zastosowaniu ewidentnie różnią się poziomem

umiejętności matematycznych – w arkuszu M_17_6 uzyskano na sześciu kotwicach średni wynik 72,3%, natomiast uczniowie rozwiązujący te same sześć zadań w arkuszu M_18_4 mieli średnią rozwiązywalność na poziomie 54,9%.

Zestawienie łatwości w tabeli 2 jest doskonałą ilustracją problemu, z jakim należało się zmierzyć przy analizie danych pochodzących z próbnego zastosowania oraz doskonałą ilustracją korzyści z wykorzystania zadań kotwiczących. Zauważenie różnic w poziomie umiejętności uczniów między grupami nie byłoby możliwe bez włączenia do arkuszy wspólnych zadań. Oceniając łatwość zadań, narażeni bylibyśmy na systematyczne błędy wynikające z różnic między grupami rozwiązującymi różne arkusze, nawet rzędu kilkunastu punktów procentowych. Natomiast dzięki temu, że zadania kotwiczące pochodziły z faktycznych egzaminów, możliwe stało się odniesienie poziomu umiejętności uczniów uczestniczących w próbnym zastosowaniu do poziomu umiejętności uczniów na egzaminie.

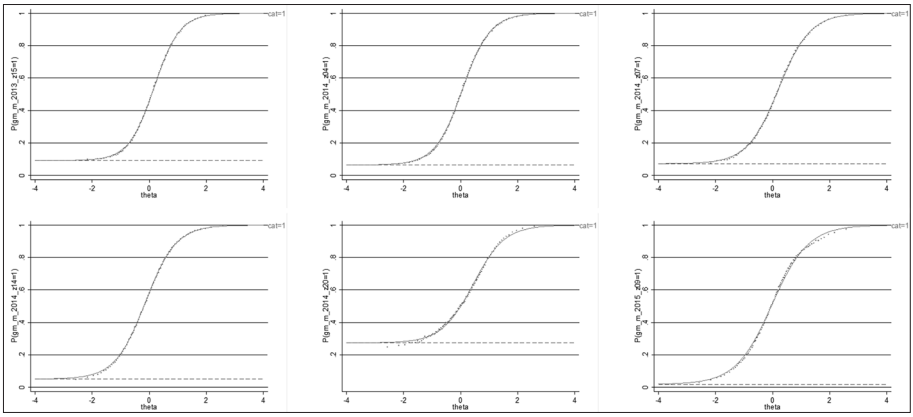
Opis zastosowanych technik modelowania IRT

Tabela 2 jest również dobrym punktem wyjścia do przedstawienia modelu IRT, który został wykorzystany do przeprowadzenia opisywanych w artykule analiz. Celem modelu IRT jest opisanie rozkładu prawdopodobieństwa wektora odpowiedzi, $U=(U_1,U_2,\ldots,U_n)$, udzielanych przez ucznia, którego wylosowano z pewnej populacji K(por. Kondratek i Pokropek, 2015):

$$P(\boldsymbol{U}=\boldsymbol{u}|K=k)=\int\prod_{i=1}^{n}f(u_i,\theta,\boldsymbol{\beta_i})\,\psi_k(\theta)\,d\theta \qquad (1)$$

Podstawowym elementem tego modelu dla zadań dychotomicznych jest tzw. funkcja charakterystyczna zadania, $f(u_i=1,\theta,\beta_i)$, która określa prawdopodobieństwo odpowiedzi poprawnej w zależności poziomu umiejętności ucznia θ oraz parametrów β_i opisujących właściwości zadania (prawdopodobieństwo odpowiedzi błędnej jest równe: $1-f(u_i=1,\theta,\beta_i)$).

Funkcje charakterystyczne zadań kotwiczących wyznaczone na danych egzaminacyjnych zebrano na rysunku 1. Modelem IRT dopasowanym do dychotomicznych zadań egzaminacyjnych był trójparametryczny model logistyczny (3plm). Jak widać na wykresach, dopasowanie modelu IRT dla wybranych zadań jest praktycznie idealne (empiryczne proporcje odpowiedzi układają się na linii wyznaczonej przez krzywą charakterystyczną). Kształt krzywych charakterystycznych zadań kotwiczących ilustruje spełnienie wspomnianego wcześniej psychometrycznego kryterium ich doboru. Mianowicie każde z kotwiczących zadań jest wysoce informatywne, punkt przegięcia krzywej charakterystycznej przypada w okolicach średniej populacyjnej (w każdym egzaminie rozkład umiejętności był zakotwiczony na rozkładzie standardowym) oraz krzywe odznaczają się znaczną „stromością". Innymi słowy, zadania kotwiczące dobrano tak, aby cechowały się znaczną zdolnością do oddzielania uczniów o wysokim poziomie umiejętności od uczniów o niskim poziomie umiejętności. Dzięki takiemu doborowi zadań kotwiczących, za pomocą niewielkiej ich liczby, możliwe było zebranie znacznej ilości informacji o poziomie umiejętności badanych uczniów.



Rysunek 1. Krzywe charakterystyczne zadań kotwiczących uzyskane z danych egzaminacyjnych wraz z empirycznymi punktami dopasowania na centylach poziomu umiejętności

Znając parametry modelu IRT (parametry zadań oraz rozkładu umiejętności), jesteśmy w stanie wyznaczyć różne klasyczne właściwości zadań. Przykładowo, przedstawione w tabeli 2 łatwości zadań kotwiczących zaobserwowane na egzaminie, $p_{i,egz}$, można uzyskać, całkując warunkowe prawdopodobieństwa poprawnej odpowiedzi w zależności od poziomu umiejętności, $f_i\,(u_i = 1, \theta, \beta_i)$, które przedstawiono na rysunku 2, przez rozkład umiejętności w populacji egzaminacyjnej, $\Psi_{egz}(\theta)$:

$$p_{i,egz} = \int f_i\,(u_i = 1, \theta, \beta_i)\;\Psi_{egz}(\theta)\;d\theta \qquad (2)$$

Praktyczna implementacja wzoru (2) sprowadza się do przeprowadzenia prostego eksperymentu Monte Carlo, w którym:

1. losujemy dużo (użyto: n = 10^7) obserwacji θ_j z rozkładu $\Psi_{egz}(\theta)$ (czyli z rozkładu standardowego normalnego),
2. dla każdego θ_j losujemy u_{ij} odpowiedź na zadanie zgodnie z funkcją $f_i\,(u_i = 1, \theta, \beta_i)$.

Przy takim podejściu wyznaczenie łatwości sprowadza się do prostego policzenia średniej z u_{ij} i podzielenie jej przez maksymalną możliwą liczbą punktów do zdobycia za zadanie (m_i, w przypadku zadań dychotomicznych równe 1):

$$p_{i,egz} = \frac{\sum_{j=1}^{n} u_{ij}}{10^7}\frac{1}{m_i} \qquad (3)$$

Dysponując parami (u_i, θ_j), można również policzyć korelację między poziomem umiejętności θ a zadaniem *i:*

$$r_{i,\theta,egz} = \mathrm{corr}\,(u_i, \theta_j) \qquad (4)$$

Współczynnik $r_{i,\theta,egz}$ stanowi analogiczny wskaźnik do klasycznego współczynnika dyskryminacji (korelacja zadania z testem lub z resztą testu). W odróżnieniu od klasycznych wskaźników jest on jednak niezależny od innych zadań

wchodzących w skład testu i dlatego został wykorzystany do raportowania mocy dyskryminacyjnej pilotowanych zadań ekspertom przedmiotowym konstruującym arkusz egzaminacyjny na podstawie wyników próbnego zastosowania.

Gdyby parametry zadań poddanych próbnemu zastosowaniu były wyznaczone na skali egzaminacyjnej, to możliwe byłoby zastosowanie wzoru (3) w celu wyznaczenia łatwości, jaką dane zadanie miałoby w populacji uczniów podchodzących do egzaminu, oraz wzoru (4) do wyznaczenia współczynnika dyskryminacji zadania na skali egzaminacyjnej. Otrzymanie tych dwóch wskaźników dla zadań poddanych próbnemu zastosowaniu było głównym celem przeprowadzanych analiz.

W celu wyznaczenia parametrów IRT pilotowanych zadań na skali umiejętności uczniów podchodzących do egzaminu gimnazjalnego, dla każdego poddanego próbnemu zastosowaniu nowego zadania, *s*, niezależnie dopasowano następujący model IRT[2]:

$$P(\boldsymbol{U}=\boldsymbol{u}) = \int f(u_s, \theta, \boldsymbol{\beta_s}) \prod_{i=1}^{6} f_i(u_i, \theta, \boldsymbol{\beta_i})\, \psi_{egz}(\theta)\, d\theta \qquad (5)$$

Zatem do każdego pilotowanego zadania dopasowano model, który opisywał statystycznie test składający się z 7 zadań: 6 kotwic, *i*, oraz danego zadania *s* o nieznanych parametrach β_s. Parametry sześciu zadań kotwiczących, β_i, oraz parametry rozkładu umiejętności, Ψ_{egz}, były ustalone na wartościach otrzymanych z danych egzaminacyjnych. Jedynymi szacowanymi z danych z próbnego zastosowania parametrami we wzorze (3) były parametry β_s. Dzięki ustaleniu pozostałych parametrów modelu (3) na wartościach populacyjnych skala, na której wyznaczone zostały parametry β_s była skalą populacji uczniów podchodzących do egzaminu.

[2] Zastosowana strategia estymacyjna może być nieintuicyjna i wymaga komentarza. Można sobie wyobrazić uzyskanie parametrów pilotowanych zadań na skali populacyjnej poprzez dopasowanie innych, alternatywnych, modeli IRT:

1) modelu uwzględniającego jednocześnie wszystkie nowe, pilotowane zadania:

$$P(\boldsymbol{U}=\boldsymbol{u}) = \int \prod_{s=1}^{n_s} f_s(u_s, \theta, \boldsymbol{\beta_i}) \prod_{i=1}^{6} f_i(u_i, \theta, \boldsymbol{\beta_i})\, \psi_{egz}(\theta)\, d\theta$$

2) modelu wielogrupowego, który uwzględniałby widoczne w tabeli 2 zróżnicowanie poziomu umiejętności matematycznych między grupami uczniów, *K*, którzy rozwiązywali różne pilotażowe arkusze:

$$P(\boldsymbol{U}=\boldsymbol{u}|K=k) = \int \prod_{s=1}^{n_s} f_s(u_s, \theta, \boldsymbol{\beta_i}) \prod_{i=1}^{6} f_i(u_i, \theta, \boldsymbol{\beta_i})\, \psi_K(\theta)\, d\theta$$

3) modelu dla 7-zadaniowego testu, który uwzględniałby wielogrupową naturę danych:

$$P(\boldsymbol{U}=\boldsymbol{u}|K=k) = \int f(u_s, \theta, \boldsymbol{\beta_s}) \prod_{i=1}^{6} f_i(u_i, \theta, \boldsymbol{\beta_i})\, \psi_K(\theta)\, d\theta.$$

Przeprowadzone zostały badania symulacyjne, które jednak jednoznacznie wskazują, że wszystkie wymienione trzy strategie prowadzą do uzyskania oszacowań rozwiązywalności zadań na skali egzaminacyjnej, które są albo obciążone (opcja wielogrupowa), albo charakteryzują się większą wariancją (opcja łącznej kalibracji wszystkich zadań). Szczegółowe omówienie wyników wspomnianych badań symulacyjnych będzie przedmiotem innej publikacji, która jest w przygotowaniu.

Większość zadań poddanych próbnemu zastosowaniu była zamkniętymi zadaniami wyboru. Właściwym modelem IRT do opisu funkcjonowania takich zadań jest model trójparametryczny, uwzględniający parametr pseudozgadywania (*c*), który opisuje niezerowe prawdopodobieństwo odpowiedzi poprawnej przy niskim poziomie umiejętności (por. rys. 1, gdzie wartość parametru *c* zaznaczono poziomą przerywaną linią). Taki model jest domyślnie stosowany przy analizie danych egzaminacyjnych, gdzie dysponujemy setkami tysięcy uczniowskich wyników. Jednak przy niewielkiej liczebności próby oszacowania parametru *c* są obciążone bardzo dużym błędem, gdyż funkcja wiarygodności dla parametru *c* ma kształt płaskiego siodła (Baker i Kim, 2004, s. 53–54). W celu wiarygodnego wyznaczenia wartości parametru *c* potrzeba licznej próby uczniów o niskim poziomie umiejętności, a taką nie dysponowano w próbnym zastosowaniu (cała próba była rzędu około 300 uczniów na jedno pilotowane zadanie). Znaczny błąd w szacowaniu parametru *c* miałby bezpośrednie przełożenie na oszacowanie klasycznej łatwości zadania przy zmianie poziomu umiejętności uczniów. W związku z tym zdecydowano, że do modelowania wszystkich pilotowanych dychotomicznych zadań, *s*, zastosowany zostanie dwuparametryczny model logistyczny (2plm), w którym parametr pseudozgadywania jest równy 0. Dla nielicznych zadań ocenianych na szerszej skali niż 0–1 zastosowano model oceny stopniowanej (*graded response model, GRM,* Samejima, 1969).

Parametry β_s uzyskane w wyniku dopasowania do każdego zadania *s* modelu (5) wykorzystano do obliczenia populacyjnej łatwości zadań (3) oraz populacyjnego współczynnika dyskryminacji zadań (4). Dodatkowo, prezentując wyniki analizy zespołom ekspertów przedmiotowych układającym arkusze egzaminacyjne, wykorzystano formuły umożliwiające dynamiczne szacowanie łatwości oraz rzetelności całego testu ułożonego z dowolnego podzbioru standaryzowanych zadań. Jeżeli *X* przez oznaczymy zbiór indeksów wybranych zadań, to średnią w całym teście wyliczymy jako:

$$M_{X,egz} = \sum_{s \in X} m_s p_{s,egz} \tag{6}$$

gdzie m_s jest maksymalną możliwą liczbą punktów do zdobycia za zadanie *s*, a $p_{s,egz}$ jest wcześniej opisaną łatwością zadania na skali egzaminacyjnej (3).

W przypadku szacowania rzetelności testu na podstawie parametrów IRT pojedynczych zadań problem jest trochę bardziej skomplikowany. Za punkt wyjścia przyjęto formułę na współczynnik rzetelności, w którym rzetelność testu jest ujmowana jako miara redukcji wariancji błędu pomiaru w wariancji całkowitej $(1 - \frac{\sigma_E^2}{\sigma_X^2})$.

W celu wyznaczenia odpowiednika populacyjnej wariancji błędu pomiaru, σ_E^2, opartego na parametrach zadań IRT wykorzystano zależność między warunkową funkcją informacji testu a wariancją błędu pomiaru umiejętności:

$$var(\hat{\theta}|\theta) \approx \frac{1}{I(\hat{\theta}|\theta)} = \frac{1}{\sum_{s \in X} I_s(\theta, \boldsymbol{\beta_s})} \quad (7)$$

Występująca w mianowniku (7) warunkowa funkcja informacji testu, $I(\hat{\theta}|\theta)$, poprzez właściwość addytywności informacji (Rao, 1982, s. 342), została wyrażona jako suma niezależnych warunkowych funkcji informacji pojedynczych zadań, $I_s(\theta,\beta_s)$, które zależą jedynie od poziomu umiejętności ucznia, θ, oraz parametrów zadań, β_s. Bezwarunkową wariancję błędu, $var(\hat{\theta})$, oszacowano, całkując po rozkładzie umiejętności funkcje informacji pojedynczych zadań:

$$var(\hat{\theta}) \approx \frac{1}{\int \sum_{s \in X} I_s(\theta, \boldsymbol{\beta_s})\, \psi_{egz}(\theta)\, d\theta} \quad (8)$$

Całkowanie we wzorze (8) przeprowadzono numerycznie. Ostatecznie, podstawiając (8) za uśrednioną wariancję błędu oraz przyjmując $\sigma_X^2 = 1$ (wyniki populacji egzaminacyjnej były zakotwiczone na rozkładzie standardowym normalnym), otrzymano następującą formułę przybliżającą rzetelność testu w rozumieniu klasycznym:

$$rzet_{X,egz} \approx 1\text{-}var(\hat{\theta}) \quad (9)$$

Przyjęty do oszacowania klasycznej rzetelności testu wzór (9) ma istotne wady. W szczególności dla niewielkiej liczby zadań może nawet przyjąć wartości ujemne, co jest sprzeczne z definicją współczynnika rzetelności. Jego zaletą jest jednak postać matematyczna formuły, która wymaga zsumowania niezależnego wkładu informacji pojedynczych zadań (por. (8)), co znacznie ułatwia zaaplikowanie go w arkuszu kalkulacyjnym jako podstawę do dynamicznego szacowania rzetelności testu złożonego z dowolnego podzbioru zadań z próbnego zastosowania.

W dalszej części opracowania, przy korzystaniu z wyników zebranych na prawdziwym egzaminie gimnazjalnym w latach 2017 oraz 2018, zostanie zweryfikowana skuteczność przewidywania właściwości pilotowanych zadań (wzory (3) i (4)) oraz całego testu (wzory (6) i (9)) na skali populacji egzaminacyjnej.

Na koniec omówienia metod IRT wykorzystanych w analizie przybliżymy oszacowany poziom umiejętności matematycznych w każdej z dwunastu grup rozwiązujących poszczególne arkusze wykorzystane w próbnym zastosowaniu. Jest to wątek poboczny do głównych celów analizy, jednak w sposób interesujący uzupełnia opis danych, który przedstawiono w tabeli 2. Do całości danych standaryzacyjnych na egzamin w roku 2017 oraz do całości danych standaryzacyjnych na egzamin w roku 2018 dopasowano wielogrupowy model IRT, w którym jedynymi szacowanymi parametrami były wartości rozkładu umiejętności w każdej z grup, K, uczniów rozwiązujących poszczególne arkusze[3]:

$$P(\boldsymbol{U} = \boldsymbol{u}|K = k) = \int \prod_{s=1}^{n_S} f_s(u_s, \theta, \boldsymbol{\beta_s}) \prod_{i=1}^{6} f_i(u_i, \theta, \boldsymbol{\beta_i})\, \psi_K(\theta)\, d\theta \quad (10)$$

[3] Parametry nowych zadań, β_s, były w tej analizie ustalone na wartościach otrzymanych wcześniej, po dopasowaniu dla każdego zadania niezależnie modelu (5).

Uzyskane po dopasowaniu modelu (10) oszacowania średniej oraz odchylenia standardowego rozkładu umiejętności, $\Psi_K(\theta)$, dla każdej z grup uczniów przedstawiono w tabeli 3.

Zgodnie z obserwacjami poczynionymi przy okazji omówienia obserwowanych łatwości zadań kotwiczących przedstawionych w tabeli 2, widzimy, że wszystkie 12 grup uczniów charakteryzował wyższy poziom umiejętności matematycznych od populacji uczniów podchodzących do egzaminu gimnazjalnego w edycjach, z których zaczerpnięto zadania kotwiczące (średnia w populacji egzaminacyjnej w każdym roku miała ustaloną wartość 0). W szczególności wyróżniona wcześniej grupa uczniów rozwiązująca zadania zebrane w arkuszu M_17_6 miała najwyższy średni poziom umiejętności matematycznych (0,63 odchylenia standardowego powyżej średniej), natomiast grupę uczniów rozwiązującą arkusz M_18_4 charakteryzował relatywnie najniższy średni poziom umiejętności (0,11 odchylenia standardowego powyżej średniej). Kolejną obserwacją, która nie była bezpośrednio uchwytna przy analizie łatwości odpowiedzi, jest niższy od populacyjnego poziom zróżnicowania umiejętności matematycznych w grupach rozwiązujących odmienne arkusze podczas próbnego zastosowania. Populacyjne odchylenie standardowe wyników było ustalone na wartości 1, natomiast w każdej z dwunastu grup odchylenie standardowe jest poniżej jedności. Takie obniżenie zróżnicowania jest spodziewanym rezultatem przy niereprezentatywnej selekcji uczniów o ponadprzeciętnym poziomie umiejętności.

Tabela 3. Parametry rozkładów umiejętności matematycznych uczniów rozwiązujących arkusze na próbnym zastosowaniu wyrażone na skali populacji egzaminacyjnej

Próbne zast. na rok 2017	Parametry $\Psi_K(\theta)$	Arkusz M_17_1	Arkusz M_17_2	Arkusz M_17_3	Arkusz M_17_4	Arkusz M_17_5	Arkusz M_17_6
	średnia θ	0,43	0,16	0,26	0,28	0,27	0,63
	SD θ	0,83	0,89	0,89	0,92	0,92	0,90
Próbne zast. na rok 2018	Parametry $\Psi_K(\theta)$	Arkusz M_18_1	Arkusz M_18_2	Arkusz M_18_3	Arkusz M_18_4	Arkusz M_18_5	Arkusz M_18_6
	średnia θ	0,53	0,50	0,20	0,11	0,38	0,27
	SD θ	0,87	0,84	0,99	0,97	0,91	0,87

Walidacja jakości przewidywania egzaminacyjnych właściwości psychometrycznych zadań poddanych próbnemu zastosowaniu

W tabeli 4 zebrano właściwości psychometryczne zadań uzyskane podczas próbnego zastosowania wraz z właściwościami zadań uzyskanymi na egzaminie gimnazjalnym w 2017 roku[4]. Dla każdego zadania podana jest surowa łatwość w badaniu pilotażowym, łatwość przewidywana na podstawie modelowania IRT (wzór (3)) oraz faktycznie zaobserwowana łatwość zadania w populacji egzaminacyjnej. Podobne zestawienie przedstawiono dla korelacji za-

[4] Analiza wyników egzaminacyjnych w latach 2017 oraz 2018 została przeprowadzona dla populacji uczniów rozwiązujących arkusz dla uczniów bez dysfunkcji (tzw. arkusz standardowy).

dania z poziomem umiejętności matematycznych (wzór (4)) oraz dla średniej z całego testu (wzór (6)). W przypadku współczynnika rzetelności możliwe było jedynie porównanie wartości przewidywanej z wykorzystaniem wzoru (9) z faktycznie uzyskaną wartością oszacowaną za pomocą współczynnika α-Cronbacha, gdyż nie sposób wyliczyć surowego, klasycznego oszacowania współczynnika rzetelności dla zadań pochodzących z różnych testów. W tabeli 5 znajdują się analogiczne informacje dla zadań wykorzystanych w egzaminie gimnazjalnym w 2018 roku. Przy numerze zadania w arkuszu standaryzacyjnym w tabelach 4 oraz 5 zaznaczono gwiazdką zadania, które ulegały drobnym modyfikacjom po przeprowadzeniu próbnego zastosowania (znacznie zmodyfikowane zadania zostały wyłączone z zestawienia). Dla najważniejszego, z praktycznego punktu widzenia, parametru łatwości zadania, różnice między przewidywaniami IRT oraz surowymi oszacowaniami łatwości przedstawiono dodatkowo w sposób graficzny na rysunku 2.

W przypadku obu egzaminów obserwujemy wyraźnie większą precyzję w szacowaniu poziomu łatwości zadań z wykorzystaniem modelowania IRT kotwiczącego rozkład umiejętności uczniów w danych egzaminacyjnych. Dla egzaminu w 2017 roku różnica w średnim wymodelowanym przewidywanym wyniku dla wszystkich analizowanych zadań, a uzyskanym na egzaminie średnim wyniku to 0,3 punktu (na 28 możliwych). Średnia różnica w łatwości pojedynczego zadania to jedynie 0,9 punktu procentowego. Dla egzaminu w 2018 średni wynik przewidziany z wykorzystaniem modelowania IRT prawie idealnie trafił w faktyczną średnią uzyskaną na egzaminie; w skali 24-punktowego zestawu zadań różnica między średnimi wynosiła setną część punktu. Przeciętna różnica między przewidzianymi łatwościami IRT a łatwościami zaobserwowanymi na egzaminie wynosiła dla 2018 roku 0,6 punktu procentowego.

Surowe łatwości zadań, na które bylibyśmy zdani bez zastosowania omawianej w artykule kontroli statystycznej, istotnie zawyżałyby przewidywany średni wynik. Średnia surowa łatwość z próbnego zastosowania była wyższa od faktycznej egzaminacyjnej wartości o 6,7 punktu procentowego dla egzaminu w roku 2017 oraz o 6,0 punktu procentowego dla egzaminu w roku 2018, co przekładało się na wyższe średnie w całym teście, odpowiednio o 1,9 (na 28) oraz 1,5 (na 24) punktu. Jest to spodziewany rezultat z racji wyższego poziomu umiejętności matematycznych uczniów biorących udział w próbnym zastosowaniu (tab. 3). Porównanie jakości przewidywanych na podstawie IRT łatwości z surowymi łatwościami z danych standaryzacyjnych podsumowują wykresy na rysunku 2. Widzimy, że w roku 2017 w przypadku 19 zadań, z 22 rozpatrywanych, oszacowanie IRT było bliższe prawdziwej, uzyskanej na egzaminie wartości, natomiast w roku 2018 przewidywania IRT były lepsze w 15 z 17 przypadków. Zaprezentowane wyniki w jednoznaczny sposób ukazują korzyści z wprowadzenia nowych procedur zbierania i analizy danych z próbnego zastosowania dla prognozowania łatwości egzaminu.

Różnice między oszacowaną surową mocą dyskryminacyjną zadań a mocą dyskryminacyjną zadań przewidywaną na podstawie IRT były niewielkie. Nie powinno to być zaskoczeniem, gdyż zastosowany współczynnik dyskryminacji jest miarą korelacyjną i przez to jest o wiele mniej podatny na niereprezentatywność

próby, w porównaniu ze średnią, będącą podstawą wyliczenia łatwości. Średnia różnica między surowymi oszacowaniami mocy dyskryminacyjnej a wartościami uzyskanymi na egzaminie wynosiła +0,01 dla 2017 roku oraz -0,05 dla 2018 roku. Współczynniki oparte na IRT wypadły podobnie, różnica dla 2017 roku wynosiła +0,02, a dla roku 2018 -0,04. Obie wartości (surowa i wykorzystująca populacyjną informację poprzez zadania kotwiczące) dają dobre przybliżenie tego, jak dane zadanie będzie korelowało z umiejętnością na prawdziwym egzaminie.

Tabela 4. Porównanie właściwości psychometrycznych zadań egzaminu z egzaminu gimnazjalnego z matematyki w 2017 roku z wartościami przewidzianymi na danych z próbnego zastosowania; gwiazdką zaznaczono zadania, w których dokonano drobnych zmian już po przeprowadzeniu badań pilotażowych

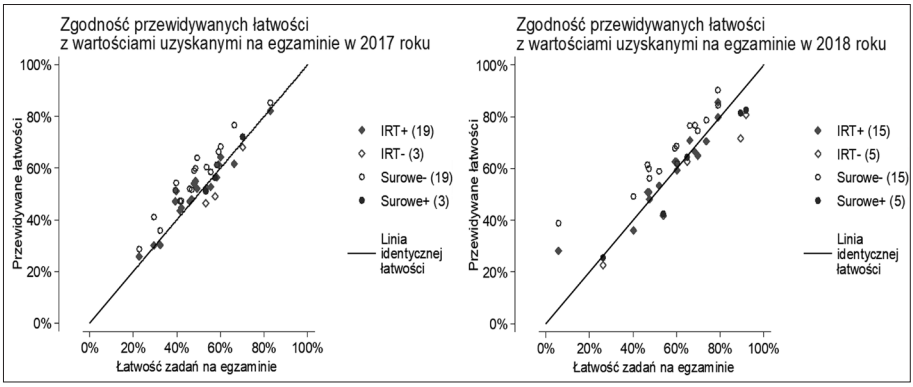
Pozycja zadania – próbne zast.	Pozycja zadania – egzamin	Łatwość					Dyskryminacja				
		próbne zastosowanie		egzamin	różnica z egzaminem		próbne zastosowanie		egzamin	różnica z egzaminem	
		surowe	IRT		surowe	IRT	surowe	IRT		surowe	IRT
M1_z12	z01	66,4%	61,3%	59,1%	7,3	2,2	0,25	0,28	0,29	-0,05	-0,01
M4_z03	z02	60,4%	52,3%	53,6%	6,8	-1,4	0,61	0,66	0,49	0,12	0,17
M2_z07	z03	51,7%	47,9%	46,8%	4,9	1,2	0,52	0,52	0,55	-0,03	-0,03
M6_z01	z04	64,0%	52,1%	49,2%	14,8	2,8	0,44	0,43	0,47	-0,04	-0,03
M3_z02	z05*	54,3%	51,0%	39,6%	14,6	11,4	0,26	0,29	0,09	0,17	0,19
M2_z04	z06	72,1%	68,1%	70,2%	1,9	-2,1	0,50	0,52	0,45	0,04	0,07
M3_z01	z08	51,4%	47,2%	39,3%	12,1	7,9	0,42	0,37	0,40	0,01	-0,03
M5_z10	z09	59,0%	54,0%	47,8%	11,2	6,2	0,41	0,43	0,47	-0,06	-0,04
M4_z10	z10*	51,0%	46,3%	53,2%	-2,2	-6,9	0,36	0,41	0,39	-0,03	0,02
M2_z06	z11*	47,3%	43,6%	41,4%	5,8	2,1	0,54	0,54	0,58	-0,05	-0,04
M3_z13	z12*	58,5%	52,7%	55,5%	3,0	-2,8	0,47	0,49	0,35	0,12	0,14
M5_z03	z13	85,3%	82,1%	82,9%	2,5	-0,8	0,38	0,36	0,42	-0,06	-0,06
M6_z15	z14	41,1%	30,1%	29,5%	11,6	0,7	0,40	0,42	0,46	-0,05	-0,04
M4_z14	z15*	56,3%	49,1%	57,5%	-1,2	-8,4	0,60	0,61	0,52	0,09	0,09
M5_z20	z16	61,3%	56,4%	58,3%	3,0	-1,9	0,40	0,43	0,32	0,09	0,10
M6_z19	z17	76,7%	61,7%	66,3%	10,4	-4,7	0,46	0,57	0,48	-0,01	0,09
M4_z16	z18	52,1%	47,0%	45,9%	6,2	1,2	0,47	0,45	0,42	0,05	0,03
M4_z20	z19	68,4%	64,2%	60,1%	8,3	4,1	0,38	0,39	0,38	0,02	0,01
M2_z21	z20*	47,3%	44,6%	42,0%	5,3	2,6	0,38	0,39	0,40	-0,03	-0,01
M2_z22	z21	28,6%	25,8%	22,7%	5,8	3,1	0,60	0,59	0,59	0,01	0,00
M2_z23	z22	59,9%	54,9%	48,6%	11,3	6,3	0,63	0,64	0,67	-0,06	-0,03
M3_z24	z23*	35,9%	30,2%	32,3%	3,6	-2,1	0,62	0,64	0,72	-0,10	-0,08
		średni wynik w teście			średnia różnic łatwości		rzetelność testu			średnia różnic dyskryminacji	
		15,05	13,49	13,19	6,7	0,9	-	0,90	0,85	0,01	0,02

W przypadku współczynnika rzetelności obserwujemy natomiast nieznaczne zawyżenie przewidywanej za pomocą wzoru (9) wartości, w porównaniu z tą, którą uzyskano na egzaminie. Oszacowanie rzetelności rozpatrywanego zbioru zadań

dla egzaminu w 2017 roku wynosiło 0,90 natomiast α-Cronbacha na danych egzaminacyjnych miało wartość 0,85. W roku 2018 przewidziano rzetelność na poziomie 0,86, natomiast α-Cronbacha też była niższa, wyniosła 0,84, mimo tego, że przewidywane współczynniki dyskryminacji zadań były nieznacznie niższe od faktycznie uzyskanych. Niekoniecznie jednak wskazuje to na obciążenie przewidywanej na podstawie IRT rzetelności testu, gdyż współczynnik α-Cronbacha wyznacza dolną granicę rzetelności testu (Novick i Levis, 1967). Aby ostatecznie rozstrzygnąć kwestię jakości przewidywania rzetelności testu, należałoby użyć w obu przypadkach tej samej miary. Jest to pole do dalszych badań, gdyż jak już wspomniano przy komentarzu do wzoru (9), zastosowany do szacowania rzetelności współczynnik oparty na modelowaniu IRT nie jest pozbawiony wad.

Tabela 5. Porównanie właściwości psychometrycznych zadań egzaminu z egzaminu gimnazjalnego z matematyki w 2018 roku z wartościami przewidzianymi na danych z próbnego zastosowania; gwiazdką zaznaczono zadania, w których dokonano drobnych zmian już po przeprowadzeniu badań pilotażowych

Pozycja zadania – próbne zast.	Pozycja zadania – egzamin	Łatwość					Dyskryminacja				
		próbne zastosowanie		egzamin	różnica z egzaminem		próbne zastosowanie		egzamin	różnica z egzaminem	
		surowe	IRT		surowe	IRT	surowe	IRT		surowe	IRT
M6_z01	z01*	82,6%	80,7%	91,9%	-9,3	-11,2	0,23	0,20	0,34	-0,11	-0,14
M1_z06	z03	61,4%	50,8%	46,7%	14,7	4,2	0,36	0,43	0,34	0,02	0,09
M1_z09	z04	74,6%	64,9%	69,7%	4,9	-4,8	0,38	0,42	0,44	-0,06	-0,02
M1_z04	z06	59,9%	50,8%	47,3%	12,6	3,5	0,34	0,38	0,34	0,00	0,04
M4_z05	z07	64,4%	62,7%	64,9%	-0,5	-2,2	0,38	0,38	0,49	-0,11	-0,11
M6_z10	z08*	67,7%	62,7%	59,2%	8,5	3,5	0,41	0,43	0,49	-0,08	-0,06
M2_z11	z09*	84,5%	79,7%	79,1%	5,4	0,7	0,22	0,26	0,41	-0,19	-0,15
M2_z06	z10	76,7%	66,4%	68,4%	8,3	-1,9	0,39	0,46	0,47	-0,08	-0,01
M2_z19	z11	49,2%	36,0%	40,2%	9,0	-4,2	0,57	0,58	0,54	0,03	0,04
M1_z12	z12*	81,5%	71,6%	89,4%	-7,9	-17,8	0,40	0,46	0,39	0,01	0,07
M1_z10	z13	78,7%	70,4%	73,7%	5,0	-3,3	0,34	0,38	0,42	-0,08	-0,04
M2_z13	z14	90,3%	85,6%	78,9%	11,4	6,7	0,27	0,30	0,42	-0,15	-0,12
M5_z17	z15	76,6%	70,8%	66,1%	10,5	4,7	0,41	0,39	0,54	-0,13	-0,15
M4_z16	z16	61,9%	59,3%	60,2%	1,7	-0,9	0,53	0,52	0,62	-0,09	-0,10
M4_z20	z17	42,5%	41,7%	53,9%	-11,4	-12,2	0,23	0,20	0,29	-0,06	-0,09
M2_z18	z18	68,6%	62,5%	60,0%	8,6	2,5	0,29	0,28	0,38	-0,09	-0,10
M5_z14	z19	58,9%	53,4%	52,0%	6,9	1,4	0,35	0,34	0,36	-0,01	-0,02
M1_z19	z20	56,1%	48,1%	47,5%	8,6	0,6	0,34	0,33	0,32	0,02	0,01
M1_z22	z21	38,9%	28,1%	5,8%	33,1	22,3	0,58	0,56	0,39	0,19	0,17
M3_z24	z22*	25,5%	22,5%	26,2%	-0,7	-3,7	0,67	0,60	0,73	-0,06	-0,13
		średni wynik w teście			średnia różnic łatwości		rzetelność testu			średnia różnic dyskryminacji	
		14,16	12,64	12,65	6,0	-0,6	-	0,86	0,82	-0,05	-0,04



Rysunek 2. Porównanie przewidywania łatwości zadań w podejściu wykorzystującym modelowanie IRT oraz zadania kotwiczące z surowymi oszacowaniami łatwości dla egzaminu z 2017 roku (z lewej) oraz z 2018 roku (z prawej); puste wskaźniki oznaczają przypadki, w których dane oszacowanie było gorsze

Podsumowanie

Porównywalność wyników egzaminacyjnych między poszczególnymi edycjami egzaminu stanowi bardzo istotny element oceny jakości systemu egzaminacyjnego. Jak wykazały badania przeprowadzone w IBE w latach 2010–2015 przez zespół Henryka Szaleńca (Szaleniec i in., 2015), wyniki egzaminacyjne w latach objętych analizą[5] odznaczały się znacznymi zmianami rozkładu, które tylko do pewnego stopnia można było wytłumaczyć populacyjnymi zmianami w poziomie umiejętności uczniów podchodzących do egzaminów w poszczególnych latach (op. cit., rysunki: 6.1, 6.3, 6.6, 6.8, 6.10 oraz 6.12).

Opisana w artykule nowa procedura przeprowadzania próbnego zastosowania zadań miała na celu umożliwienie spełnienia warunku porównywalności tworzonych w CKE egzaminów. Możliwość zestawienia właściwości psychometrycznych testu gimnazjalnego z matematyki w 2017 oraz 2018 roku z charakterystykami zadań oraz testu przewidzianymi na podstawie próbnego zastosowania zadań stworzyła unikalną szansę na ocenę jakości nowego podejścia do kontroli właściwości psychometrycznych egzaminu.

Uzyskane wyniki są obiecujące. Najistotniejszy parametr, jakim jest trudność testu, został dla egzaminu w 2017 oraz 2018 roku przewidziany bardzo trafnie. Różnica między przewidzianymi średnimi a faktycznie uzyskanymi dla analizowanych zbiorów zadań nie przekroczyła 1% zakresu możliwych do uzyskania punktów (w 2018 roku różnica średnich wynosiła setną część punktu na 24 punkty możliwe do uzyskania). Dla porównania, korzystając z surowych oszacowań łatwości zadań, bez zastosowania modelowania IRT z zadaniami kotwiczącymi, przewidywana średnia w teście byłaby zawyżona o ponad 6% w skali możliwych do uzyskania wyników. Przewidywanie mocy dyskryminacyjnej zadań również

[5] W badaniu kontrolowano populacyjne zmiany w poziomie umiejętności uczniów podchodzących do egzaminów w latach 2002–2013: sprawdzianu w klasie szóstej i egzaminu gimnazjalnego oraz w latach 2010–2013 do egzaminu maturalnego na poziomie podstawowym z matematyki, języka polskiego oraz języka angielskiego.

było satysfakcjonujące, przy czym surowe wskaźniki dyskryminacji miały podobną zgodność z wartościami uzyskanymi na egzaminie, jak te uzyskane z wykorzystaniem modelowania IRT oraz zadań kotwiczących.

Oszacowana rzetelność testu była natomiast nieznacznie zawyżana w zastosowanej procedurze. Może to być efekt porównania rzetelności wyliczonej za pomocą IRT z rzetelnością opartą na klasycznym wskaźniku α-Cronbacha. Uzyskany rezultat skłania jednak do refleksji nad wprowadzeniem innego sposobu szacowania rzetelności dowolnego zbioru zadań z wykorzystaniem IRT. Dobrym kandydatem byłby wskaźnik oparty na odmianie wzoru α-Cronbacha (znanej też jako K_{20}, *20 Kudera-Richardsona*, Machowski, 1993), w którym rzetelność testu jest określona poprzez właściwości pojedynczych zadań:

$$K_{20} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sum_{i=1}^{n}\sigma_i^2 + \sum_{j\neq i}^{n}\sigma_{ij}^2}\right) \qquad (11)$$

Występujące we wzorze (11) wariancje, σ_i^2, oraz kowariancje, σ_{ij}^2, pojedynczych zadań można wyznaczyć z wykorzystaniem parametrów IRT oszacowanych na skali egzaminacyjnej analogicznie jak poczyniono z klasyczną łatwością (wzory (2) oraz (3)). Ocena funkcjonowania takiego alternatywnego wskaźnika rzetelności opartego na IRT wydaje się interesującym tematem dla dalszych badań.

Potencjalnie istotnym ograniczeniem przeprowadzonych analiz jest przyjęcie założenia o braku zmian w poziomie umiejętności uczniów między latami. Z jednej strony wykorzystane w próbnym zastosowaniu zadania kotwiczące miały dla każdego z egzaminów, z którego zostały zaczerpnięte (lata 2013, 2014 oraz 2015), parametry ustalone na takim samym rozkładzie populacyjnym. Z drugiej strony przewidywane charakterystyki egzaminacyjne pilotowanych zadań na egzaminach w latach 2017 oraz 2018 też zakładały ten sam rozkład umiejętności uczniów. Jak wykazano w projekcie badającym porównywalność wyników egzaminacyjnych, poziom umiejętności w populacji uczniów może podlegać między latami drobnym fluktuacjom. Przeprowadzenie dodatkowych badań zrównujących wyniki egzaminu gimnazjalnego z lat 2013–2018 na pewno rzuciłoby dodatkowe światło na przedstawione wyniki. Przedstawiona w artykule bardzo duża zgodność między przewidywanymi właściwościami zadań i całego testu z wartościami uzyskanymi na egzaminach w roku 2017 oraz 2018 sugeruje jednak, że udział fluktuacji umiejętności między latami nie ma tu istotnego znaczenia.

Warto wspomnieć, że uzyskana w wyniku zastosowania opisanej metodologii kontrola właściwości psychometrycznych tworzonych w CKE egzaminów mogłaby, teoretycznie, zostać osiągnięta przez włączenie do egzaminów niejawnego bloku zadań kotwiczących, który byłby wspólny między latami. Alternatywnie można by też próbne zastosowanie nowych zadań, przewidzianych na przyszłe egzaminy, przeprowadzić, włączając takie zadania do arkuszy egzaminacyjnych w głównej sesji w bloku niejawnym. Zaletą takich rozwiązań byłoby zbieranie informacji o różnicy w poziomie umiejętności uczniów podchodzących egzaminu w różnych latach lub o właściwościach psychometrycznych pilotowanych zadań w faktycznej sytuacji egzaminacyjnej, tj. przy specyficznych warunkach motywacji oraz przygotowania uczniów, jakie występują

podczas egzaminu. Rozwiązanie ze stałymi kotwicami wymagałoby, po zebraniu wyników, dokonania na danych egzaminacyjnych dość skomplikowanych przekształceń zrównujących wyniki między latami, których naturę trudno byłoby przekazać bezpośrednim odbiorcom wyników egzaminacyjnych oraz szerszej opinii publicznej. Natomiast pilotaż nowych zadań w trakcie egzaminu wiązałby się z drukiem wielu wersji arkuszy oraz umieszczeniem w arkuszach egzaminacyjnych zadań, których rozwiązania nie byłyby uwzględniane w wyniku ucznia, co może budzić kontrowersje.

Największe ograniczenia przed wprowadzeniem niejawnych zadań w obu omówionych scenariuszach mają jednak naturę prawną – arkusze egzaminacyjne są jawne, można je fotografować, uczniowie mają możliwość wglądu do swoich prac. Należy także uwzględnić fakt, że utrzymanie niejawności powszechnie stosowanego w populacji zbioru zadań jest w praktyce bardzo trudne, jeżeli nie niemożliwe, do realizacji. Reasumując, rozwiązania metodologiczne oparte na włączeniu zadań niejawnych do arkuszy egzaminacyjnych, mimo pewnych oczywistych korzyści z takiego podejścia, nie stanowią realnej alternatywy dla zwiększenia porównywalności wyników egzaminacyjnych między latami do podejścia opisanego w artykule. Taka konkluzja jest w mocy zwłaszcza w świetle pozytywnej oceny jakości przewidywań opartych na relatywnie taniej i nieingerującej w sesję egzaminacyjną procedury opisanej w artykule.

Wysoka jakość przewidywania właściwości psychometrycznych egzaminu gimnazjalnego z matematyki poprzez kotwiczenie danych z próbnego zastosowania do wcześniejszych egzaminów i wykorzystanie modelowania IRT sugeruje, że warto taką procedurę stosować jako standardowe rozwiązanie przy tworzeniu egzaminów. Obecnie w ten sposób tworzony jest egzamin ósmoklasisty (z każdego przedmiotu, włączając egzamin z przedmiotów dodatkowych) oraz egzamin maturalny z wybranych przedmiotów (matematyka, biologia, język angielski oraz język niemiecki). Opisane rozwiązanie powinno być także brane pod uwagę we wszystkich innych kontekstach badawczych, gdy dysponujemy zbiorem pozycji testowych o znanych populacyjnych parametrach IRT i stoimy przed zadaniem wyznaczenia populacyjnych właściwości pozycji testowych (lub całych testów) na potencjalnie niereprezentatywnej próbie badawczej.

Bibliografia

Baker, F. B. & Kim, S. (2004). *Item Response Theory. Parameter estimation techniques*. New York: Marcel Dekker.

Kondratek, B. (2016). *UIRT: Stata module to fit unidimensional Item Response Theory models*. Statistical Software Components S458247, Boston College Department of Economics.

Kondratek, B. i Pokropek, A. (2015) Teoria odpowiedzi na pozycje testowe: jednowymiarowe modele dla cech ukrytych o charakterze ciągłym. W: A. Pokropek (red.) *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania* (s. 15–32). Warszawa: Instytut Badań Edukacyjnych.

Machowski, A. (1993). *Rzetelność testów psychologicznych. Dwa ujęcia modelowe*. Warszawa–Poznań: Wydawnictwo Naukowe PWN.
Novick, M. R. & Levis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32(1), 1–13.
Rao, C. R. (1982). Modele liniowe statystyki matematycznej. Warszawa: Państwowe Wydawnictwo Naukowe.
Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond: Psychometric Society.
Szaleniec, H., Kondratek, B., Kulon, F., Pokropek, A., Skórska, P., Świst, K., Wołodźko, T. i Żółtak, M. (2015). *Porównywalne wyniki egzaminacyjne*. Warszawa: Instytut Badań Edukacyjnych.