Edukacja 2014, 2(127), 67–84
ISSN 0239-6858

Konsekwencje błędnego określenia rodzaju zadania testowego

Paulina Skórska, Karolina Świst, Henryk Szaleniec
Instytut Badań Edukacyjnych*

Przedmiotem artykułu jest wpływ niepoprawnego określenia rodzaju zadania na trafność wyników egzaminu. W wyniku przeglądu zadań części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2002–11 zidentyfikowano 9 zadań uznanych przez CKE za otwarte, mimo że treściowo i psychometrycznie funkcjonowały one jak zamknięte. Dla jednego z tych zadań przeprowadzono studium przypadku z wykorzystaniem modelowania IRT. Omówiony przypadek dowodzi zgadywania poprawnej odpowiedzi w zadaniach błędnie zakwalifikowanych jako zadania otwarte.

Słowa kluczowe: rodzaj zadania, trafność teoretyczna, trafność treściowa, zadania otwarte, zadania zamknięte.

Większość badaczy zajmujących się konstrukcją testów podkreśla znaczenie trafności treściowej (*content validity*) zadań egzaminacyjnych. Zadanie powinno sprawdzać określoną umiejętność ucznia, co znaczy, że zakres treściowy zadania powinien w jak największym stopniu pokrywać się z zakresem mierzonej umiejętności. W praktyce ważna jest nie tylko treść zadania, ale także jego rodzaj, forma i skala odpowiedzi. Rodzaj i forma zadania wpływają na psychometryczne właściwości skali, co z kolei decyduje o trafności teoretycznej (*construct validity*) oraz kryterialnej (*criterion validity*; Rauthmann, 2011). Poniżej omawiamy zagadnienie błędnego określenia rodzaju zadania w części matematyczno-przyrodniczej testowego egzaminu gimnazjalnego w latach 2002–2011 wraz z potencjalnymi konsekwencjami takiego określenia.

Od struktury umiejętności do zadania egzaminacyjnego

Procedura przygotowania arkusza egzaminacyjnego, niezależnie od tego, czy stosujemy strategię analityczną, czy holistyczną, rozpoczyna się od analizy treści kształcenia, która prowadzi do stworzenia planu i kartoteki tego arkusza, a w dalszej kolejności do konstruowania zadań. Treść kształcenia określa w Polsce podstawa programowa. Dla trzeciego etapu edukacyjnego (gimnazjum), podstawa programowa z lat 2002–2011 (MEN, 1999,

Artykuł powstał w ramach projektu systemowego „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" finansowanego ze środków Europejskiego Funduszu Społecznego w ramach Programu Operacyjnego Kapitał Ludzki (Priorytet III: Wysoka jakość systemu oświaty, Poddziałanie 3.1.1. Tworzenie warunków i narzędzi do monitorowania, ewaluacji i badań systemu oświaty).

* Adres do korespondencji: ul. Górczewska 8, 01-180 Warszawa. E-mail: p.skorska@ibe.edu.pl

s. 601) określiła następujące ogólne cele nauczania:

- Wprowadzenie ucznia w świat nauki przez poznanie języka, pojęć, twierdzeń i metod właściwych dla wybranych dyscyplin naukowych, w tym w językach obcych, na poziomie umożliwiającym dalsze kształcenie.
- Rozbudzenie i rozwinięcie indywidualnych zainteresowań ucznia.
- Wprowadzenie ucznia w świat kultury i sztuki.
- Rozwinięcie umiejętności społecznych ucznia przez zdobywanie prawidłowych doświadczeń we współżyciu i współdziałaniu w grupie rówieśniczej.

Dla każdego przedmiotu zdefiniowano cele edukacyjne, zadania szkoły oraz treść nauczania z wyróżnieniem oczekiwanych osiągnięć. Zgodnie z podstawą programową sformułowano także standardy wymagań będące podstawą egzaminu w ostatnim roku nauki w gimnazjum (MEN, 2007, s. 18–20). Dla części matematyczno-przyrodniczej zdefiniowano cztery obszary standardów:

- I: umiejętne stosowanie terminów, pojęć i procedur z zakresu przedmiotów matematyczno-przyrodniczych niezbędnych w praktyce życiowej i dalszym kształceniu;
- II: wyszukiwanie i stosowanie informacji;
- III: wskazywanie i opisywanie faktów, związków i zależności, w szczególności przyczynowo-skutkowych, funkcjonalnych, przestrzennych i czasowych;
- IV: stosowanie zintegrowanej wiedzy i umiejętności do rozwiązywania problemów.

Dla każdego z obszarów określono umiejętności, które definiują zakres tego, co podlega egzaminowaniu. Na przykład obszar III został opisany następująco.

Uczeń wskazuje prawidłowości w procesach, w funkcjonowaniu układów i systemów:

- wyodrębnia z kontekstu dane zjawisko,
- określa warunki jego występowania,
- opisuje przebieg zjawiska w czasie i przestrzeni,
- wykorzystuje zasady i prawa do objaśniania zjawisk (MEN, 2007, s. 19).

W podobny sposób zostały opisane pozostałe czynności, które uczeń powinien umieć samodzielnie wykonać. Odpowiadają one zbiorowi umiejętności, których próba mogła być w danym roku przedmiotem pomiaru na egzaminie.

Do 2011 r. planowanie i przygotowywanie arkuszy egzaminacyjnych dla gimnazjum odbywało się według zasad techniki autorskiego konstruowania narzędzi pomiarowych. W wypadku arkuszy standardowych (dla uczniów bez dysfunkcji i z dysleksją rozwojową) każda okręgowa komisja przygotowywała arkusz egzaminacyjny na dany rok, a Centralna Komisja Egzaminacyjna (CKE) decydowała o wyborze i ostatecznym kształcie finalnej wersji przeznaczonej do egzaminu w pierwszym i drugim terminie.

Specyfikacja arkusza egzaminacyjnego, mająca charakter ogólnego zarysu planu testu, określała:

- maksymalną liczbę punktów do uzyskania, z podziałem na standardy wymagań egzaminacyjnych,
- przedział, w którym powinna się mieścić średnia łatwość testu,
- przedziały średniej łatwości podtestów dla poszczególnych standardów wymagań egzaminacyjnych,
- proporcje zadań otwartych i zamkniętych,
- ogólny opis reprezentacji treści z poszczególnych przedmiotów (CKE, 2005a).

Zadaniem egzaminacyjnym nazywamy najmniejszą, względnie niezależną i osobno punktowaną część testu egzaminacyjnego obejmującą opis sytuacji, pytanie lub polecenie i ewentualnie gotowe odpowiedzi do wyboru (odpowiedź poprawną i dystraktory) lub wskazówki ukierunkowujące pracę ucznia. Zadania w części matematyczno-przyrodniczej powinny sprawdzać umiejętności z biologii, chemii, fizyki z astronomią, geografii i matematyki, z zachowaniem międzyprzedmiotowego charakteru egzaminu. Zgodnie ze specyfikacją liczba punktów

możliwych do uzyskania za rozwiązanie wszystkich zadań w arkuszu matematyczno-przyrodniczym wynosiła 50. Za zadania zamknięte wielokrotnego wyboru punktowane zero-jedynkowo uczeń mógł uzyskać maksymalnie 25 punktów, tyle samo, ile za zadania otwarte o rozszerzonej i krótkiej odpowiedzi. W specyfikacji części humanistycznej i matematyczno-przyrodniczej nie założono minimalnej ani średniej mocy różnicującej, jak również wartości wskaźnika rzetelności dla całego testu i dla podskal odpowiadających poszczególnym standardom wymagań egzaminacyjnych.

Proces opracowania arkuszy egzaminacyjnych przez autorskie zespoły w okręgowych komisjach egzaminacyjnych regulowały procedury tworzone i zatwierdzane przez dyrektorów okręgowych i Centralnej Komisji Egzaminacyjnej. Począwszy od 2001 r. procedury te ulegały zmianom polegającym na ich doprecyzowaniu. W ogólnym zarysie obejmowały one:

- założenia ogólne, dotyczące powoływania zespołów autorskich,
- zasady przygotowania kartoteki testu oraz tworzenia zadań,
- zasady przeprowadzenia próbnego zastosowania zestawów egzaminacyjnych na próbie celowej, warstwowanej ze względu na lokalizację szkoły,
- zasady opracowania wyników próbnego zastosowania z wyszczególnieniem koniecznych parametrów statystycznych zadań i testu,
- ustalenia dotyczące recenzji nauczycielskiej i akademickiej,
- zasady przekazywania przez OKE finalnej wersji arkuszy egzaminacyjnych do CKE,
- zasady analizy zestawów egzaminacyjnych w CKE i kryteria wyboru do zastosowania w kraju.

Słabym punktem autorskiej techniki przygotowywania narzędzi pomiarowych do egzaminu jest po pierwsze, ograniczony zakres próbnego zastosowania testu (ze względu na tajemnicę egzaminacyjną) i po drugie, nieznany (przed egzaminem) wpływ korekt dokonanych w CKE już po wybraniu arkusza do zastosowania w całym kraju.

Klasyfikacje zadań egzaminacyjnych i ich znaczenie

Przedstawiony proces tworzenia zadań egzaminacyjnych odnosi się do jeszcze jednej kwestii – doboru rodzaju oraz formy zadań. Najbardziej ogólna klasyfikacja zadań egzaminacyjnych (Downing, 2009) dzieli je na zamknięte (*selected-response item format*, SR) i otwarte (*constructed-response item format*, CR). Zadania otwarte wymagają od ucznia samodzielnego sformułowania i zapisania odpowiedzi w reakcji na bodziec, którym zazwyczaj jest pytanie lub stwierdzenie. Udzielone odpowiedzi są następnie analizowane przez egzaminatorów, którzy przydzielają im określoną liczbę punktów, zgodnie z wcześniej przygotowanym schematem punktowania obejmującym kryteria i skalę. Zarówno klasyczne definicje (Cronbach, 1984; Niemierko, 1999), jak i te bardziej współczesne (Hohensinn i Kubinger, 2011) kładą nacisk na samodzielne wytworzenie odpowiedzi przez ucznia, a nie na sam fakt wpisywania czegokolwiek w arkuszu testowym. Natomiast zadania zamknięte wymagają od egzaminowanego wybrania, a następnie zaznaczenia poprawnej odpowiedzi lub (częściej) najlepszej odpowiedzi z podanej listy możliwych wariantów (Hohensinn i Kubinger, 2011).

Zarówno zadania zamknięte, jak i otwarte mają swoje zalety i wady (Downing, 2009, Niemierko, 1975). Zaletą stosowania zadań otwartych jest to, że odpowiedzi pozwalają obserwować logikę myślenia, kroki w dochodzeniu do rozwiązania problemu, dając pogłębiony obraz umiejętności ucznia. Główne ich wady są związane z koniecznością wydłużenia

czasu testowania i z gorszą reprezentacją treści kształcenia z uwagi na konieczność ograniczenia liczby zadań. Proces oceny odpowiedzi jest znacznie dłuższy i wymaga zaangażowania egzaminatorów, większy jest również koszt oceniania. Dodatkowo zadania te są podatne na efekt egzaminatora (problemy z subiektywizmem i powtarzalnością punktowania). Zadania zamknięte z kolei mają obiektywny system oceny, który jest szybki, powtarzalny i łatwy do uzasadnienia, są jednak trudne w konstrukcji. Dlatego na etapie pilotażu tekstu zaleca się przygotowanie zadania zamkniętego w wersji otwartej oraz analizę odpowiedzi udzielanych przez uczniów, w celu konstrukcji dystraktorów spójnych z prawidłową odpowiedzią (Tyralska-Wojtycza, 2010). Co więcej, zadania zamknięte charakteryzują się podatnością na zgadywanie i używanie cząstkowej wiedzy do eliminowania mniej prawdopodobnych odpowiedzi (Ebel i Frisbie, 1991; Niemierko, 1999). Uczniowie mogą także wykazywać tendencję do zapamiętywania prawidłowych odpowiedzi oraz do innych nieetycznych zachowań w sytuacji egzaminacyjnej, takich jak ściąganie.

Należy jednak pamiętać, że zadania otwarte i zadania zamknięte mają inne przeznaczenie, ponieważ rozwiązywanie różnych rodzajów zadań aktywizuje różne rodzaje pamięci ucznia. Podczas udzielania odpowiedzi otwartej częściej aktywizowana jest wiedza proceduralna, a podczas odpowiadania na pytanie zamknięte – wiedza deklaratywna (Ackerman i Smith, 1988). Na przykład w zadaniach otwartych mających formę eseju bada się trzy procesy kognitywne: (a) planowanie i strukturyzację eseju, (b) przełożenie tych planów na zdania oraz (c) przejrzenie tekstu pod kątem jego ulepszenia i eliminacji pojawiających się błędów (Hayes i Flower, 1980). W wypadku zadań zamkniętych wielokrotnego wyboru dwa pierwsze procesy nie są konieczne, wymagana jest natomiast analiza informacji niezbędnych do wyboru rozwiązania zadania przedstawionych w trzonie, ocena zestawu możliwych odpowiedzi, dokonanie wyboru i zakreślenie najlepszego wariantu.

W literaturze naukowej często występują bardziej szczegółowe klasyfikacje zadań testowych, niż prosty podział na zadania zamknięte i otwarte. Jedna z nich pochodzi od Stevena Downinga (2009, s. 152–154). Zadania otwarte dzieli on pod względem formy na zadania krótkiej odpowiedzi (*short answer constructed-response*, KO) dla odpowiedzi nie dłuższych niż trzy zdania oraz zadania rozszerzonej odpowiedzi (*long answer constructed-response*, RO) z tekstem nie dłuższym niż pięć stron. Bardziej rozbudowany jest podział zadań zamkniętych na następujące formy.

- Tradycyjne zadania zamknięte wielokrotnego wyboru (*multiple-choice item*, MCQ). Są to zadania z jedną prawidłową odpowiedzią.
- Złożone zadania zamknięte wielokrotnego wyboru (*complex multiple-choice*, Type K). W zadaniach tych dostępne są alternatywy typu „wszystkie podane odpowiedzi są poprawne", „żadna z podanych odpowiedzi nie jest poprawna", „dwie lub więcej z podanych odpowiedzi są poprawne".
- Zadania zamknięte typu prawda/fałsz (*true-false*, TF).
- Wielokrotne zadania zamknięte typu prawda/fałsz (*multiple true-false*, MTF). Ocena na skali prawda/fałsz nie następuje dla pojedynczego stwierdzenia, ale dla zestawu stwierdzeń powiązanych tematycznie.
- Zadania zamknięte z alternatywnym wyborem (*alternate-choice*, AC). Są to zadania, w których dostępne odpowiedzi nie ograniczają się do stwierdzeń „prawda" lub „fałsz", ale mogą być dowolnymi dwoma, wzajemnie wykluczającymi się słowami/stwierdzeniami.

- Tradycyjne zadania zamknięte na dopasowywanie/przyporządkowanie (*traditional matching*, TM). W zadaniach tych prosi się ucznia o dopasowanie lub przyporządkowanie elementów jednej listy do elementów drugiej listy zgodnie z kryterium zamieszczonym w poleceniu.
- Zadania zamknięte na dopasowywanie/przyporządkowanie typu rozszerzonego (*extended matching*, EM). Różnią się od powyższych tym, że jedna z list wykorzystywanych w dopasowywaniu zawiera, zamiast krótkich stwierdzeń, poszerzony opis, zwykle 3–6-zdaniowy (Wood, 2003).
- Wiązka zadań/zestaw zadań powiązanych ze względu na kontekst (*testlets/ /context-dependent item sets*). Jest to grupa zadań, które posiadają wspólny trzon. Tym trzonem może być na przykład tekst, rysunek, mapa itd.[1].

W badaniu PISA 2012 (OECD, 2013) w zakresie przedmiotów przyrodniczych korzystano z zadań otwartych (krótkiej i rozszerzonej odpowiedzi) i zamkniętych. W ramach pytań zamkniętych korzystano z dwóch form zadań: klasycznego zadania zamkniętego wielokrotnego wyboru (*simple multiple choice*) z jedną poprawną odpowiedzią spośród kilku możliwości oraz złożonego zadania zamkniętego wielokrotnego wyboru (*complex multiple choice*). Te drugie obejmowały serię zadań typu prawda/ /fałsz, zadania z więcej niż jedną poprawną odpowiedzią, zadania z uzupełnieniem luki wybranym z listy elementem lub wymagające porządkowania, lub dobierania elementów. Badania TIMSS i NAEP (Neidorf, Binkley i Stephens, 2006) stosują najprostszą klasyfikację rodzajów zadań (zamknięte i otwarte), z wyróżnieniem zadań otwartych wymagających krótkiej lub rozszerzonej odpowiedzi.

W Polsce najczęstszym punktem odniesienia do konstrukcji zadań egzaminacyjnych jest klasyfikacja Bolesława Niemierki (1975). Dzieli on zadania na dwa rodzaje: otwarte i zamknięte. Wśród zadań otwartych wyróżnia trzy formy: rozszerzonej odpowiedzi (RO), krótkiej odpowiedzi (KO) oraz z luką (L). Dla zadań zamkniętych wyróżnione są także trzy formy: zadania wielokrotnego wyboru (WW), prawda/fałsz (PF) oraz zadania na dobieranie (D).

W egzaminie gimnazjalnym do 2012 r. CKE stosowała prosty podział zadań na dwa rodzaje – zamknięte i otwarte. W kartotekach arkuszy egzaminacyjnych rzadko pojawia się specyfikacja ich formy (np. CKE, 2005b). Specyfikacja arkusza egzaminacyjnego w latach 2002–2011 ustalała liczbę zadań zamkniętych, ich rodzaj i podział punktów możliwych do uzyskania za każdy z dwóch rodzajów zadań.

Problem i pytania badawcze

W literaturze przedmiotu (np. Haladyna, Downing i Rodriguez, 2002) znajdujemy szereg wytycznych dotyczących konstrukcji zadań testowych, brakuje jednak informacji, jakie konsekwencje dla analiz i interpretacji wyników egzaminu może mieć zaklasyfikowanie jako otwartych zadań funkcjonujących psychometrycznie i treściowo jak zadania zamknięte. Podejmując próbę zmierzenia się z tym problemem, poszukujemy odpowiedzi na trzy pytania:

- Jak często w arkuszach egzaminu gimnazjalnego z lat 2002–2011 kwalifikowano zadania zamknięte do grupy zadań otwartych?
- W jaki sposób zadania zidentyfikowane jako problematyczne pod względem przy-

[1] W kontekście polskiego systemu egzaminów zewnętrznych, w których często występują wiązki zadań, można traktować tego typu rodzaj zadania przekrojowo. Pytania w ramach wiązki mogą przyjmować różne formy (spośród pozostałych form zadań wymienionych przez Downinga). Można się zastanawiać, czy wiązka zadań powinna tworzyć odrębną grupę w klasyfikacji. W artykule zachowano jednak oryginalny charakter klasyfikacji Downinga.

pisanego rodzaju lub formy zostałyby zakwalifikowane w znanych systemach klasyfikacyjnych?
- Jakie mogą być psychometryczne konsekwencje niepoprawnego zakwalifikowania zadań egzaminacyjnych?

Dane i metody analizy

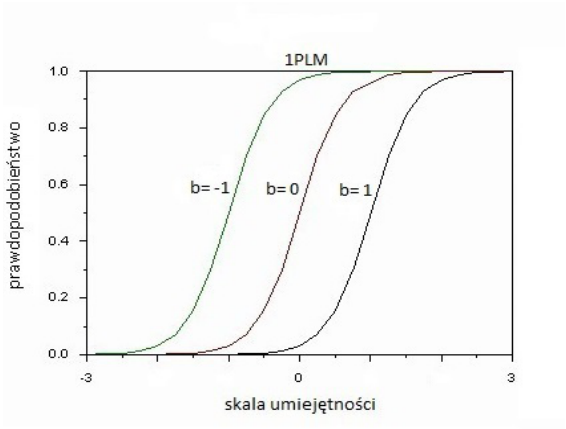
Analizie poddano wszystkie standardowe arkusze egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej, łącznie ze schematami oceniania z lat 2002–2011. W części humanistycznej nie wystąpiły żadne zadania otwarte z błędnie przypisanym rodzajem lub formą. W części matematyczno-przyrodniczej wśród zadań umieszczonych w testach na pozycjach od 26. wzwyż, czyli zadeklarowanych w kartotekach testów jako otwarte (CKE, 2011a), zidentyfikowano 9 zadań, które naszym zdaniem mają cechy zadań zamkniętych.

Zastosowano dwie metody analizy problematycznych zadań: jakościową i ilościową. Metoda jakościowa opierała się na przeglądzie arkuszy egzaminacyjnych pod kątem wybranych klasyfikacji zadań. W części ilościowej przeprowadzono analizę funkcjonowania zadań testowych za pomocą modeli odpowiedzi na pozycje testowe (*item response theory*, IRT) (por. np. Kondratek i Pokropek, 2013). Teoria IRT dostarcza narzędzi statystycznych pozwalających analizować zachowania ucznia w stosunku do pojedynczego zadania testowego, a nie całego testu (van der Linden i Hambleton, 1997). Funkcjonowanie zadania jest graficznie zilustrowane przez tzw. krzywą charakterystyczną zadania (*item characteristic curve*, ICC). Funkcja logistyczna pozwalająca otrzymać ICC łączy prawdopodobieństwo sukcesu ucznia (najczęściej równoznacznego z udzieleniem poprawnej odpowiedzi) z poziomem mierzonej umiejętności oraz z cechami zadania, takimi jak trudność, dyskryminacja i podatność na zgadywanie (Hambleton, Swaminathan i Rogers, 1991).

Trzy najbardziej popularne modele IRT przyjmują swoje nazwy od liczby parametrów (charakterystyk) zadania. Jednoparametryczny model logistyczny (*one-parameter logistic model*, 1PLM) jest często nazywany modelem Rascha (Rasch, 1960) z uwagi na to, że jest jego matematycznym ekwiwalentem. W modelu 1PLM zadanie testowe jest charakteryzowane przez jeden parametr – trudność (*difficulty parameter*), oznaczany najczęściej jako *b*. Wskazuje on punkt na skali umiejętności ucznia (θ), w którym prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie wynosi 0,5. Im wyższa trudność zadania, tym większy wymagany poziom umiejętności ucznia, by szansa udzielenia poprawnej odpowiedzi na zadanie wyniosła 50%. Parametr trudności zadania jest czasem nazywany parametrem pozycji (*location parameter*), gdyż wskazuje na położenie ICC względem skali umiejętności. Niższe wartości *b* wiążą się z przesunięciem krzywej charakterystycznej zadania w lewo względem skali umiejętności (ku niższym poziomom umiejętności ucznia), co świadczy o tym, że zadanie jest łatwe. ICC dla trudnego zadania będzie przesunięta w prawo względem skali umiejętności uczniów (ku wyższym poziomom umiejętności). Krzywe charakterystyczne zadań analizowanych za pomocą modelu 1PLM różnią się wyłącznie położeniem (przesunięcie ICC na lewo lub prawo skali umiejętności – patrz Rysunek 1[2]).

Dwuparametryczny model logistyczny (*two-parameter logistic model*, 2PLM; Birnbaum, 1968) szacuje także parametr dyskryminacji zadania (*discrimination parameter*, *a*). W modelu 1PLM przyjmuje on stałą wartość dla wszystkich zadań, wynoszącą 1.

[2] Rysunki 1, 2 i 3 – opracowanie własne na podstawie http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm.

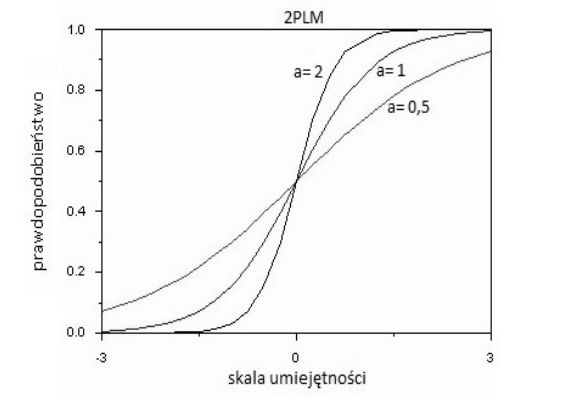


Rysunek 1. Przykładowe krzywe charakterystyczne zadań w modelu 1PLM.

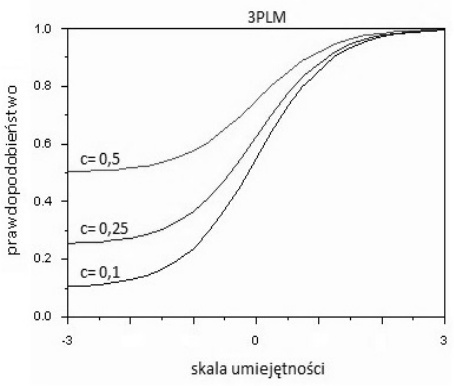
Parametr dyskryminacji jest proporcjonalny do nachylenia ICC. W modelu 2PLM zadania różnią się więc nie tylko położeniem, ale i nachyleniem (Rysunek 2). Im wyższy parametr dyskryminacji, tym większa zdolność danego zadania do różnicowania uczniów pod względem różnych poziomów ich umiejętności sprawdzanej tym zadaniem.

Trzyparametryczny model logistyczny (*three-parameter logistic model*, 3PLM; Samejima, 1969) dopuszcza niezerową dolną asymptotę krzywej charakterystycznej (Rysunek 3). Umożliwia on oszacowanie trzeciego parametru, zwanego parametrem zgadywania/pseudozgadywania[3] (*guessing parameter/pseudo-chance-level parameter*, *c*), który reprezentuje prawdopodobieństwo udzielenia poprawnej odpowiedzi w zadaniu przez uczniów o bardzo niskim poziomie umiejętności. Ten model zakłada, że nawet uczniowie o najniższym poziomie umiejętności mogą mieć niezerowe prawdopodobieństwo udzielenia poprawnej odpowiedzi (np. zgadując, która z odpowiedzi jest poprawna).

[3] Interpretacja parametru *c* pozostaje do dziś przedmiotem dyskusji. Zainteresowany czytelnik znajdzie więcej na ten temat w publikacjach: Hambleton i in. (1991), Lord (1974) i Han (2012).



Rysunek 2. Przykładowe krzywe charakterystyczne zadań w modelu 2PLM.



Rysunek 3. Przykładowe krzywe charakterystyczne zadań w modelu 3PLM.

Podsumowując, bardziej złożone modele IRT (spośród opisywanych) są uogólnieniami mniej złożonych. Model 1PLM jest specjalnym przypadkiem modelu 2PLM, w którym $a = 1$, a model 2PLM jest specjalnym przypadkiem modelu 3PLM, w którym $c = 0$.

Zadania egzaminacyjne zaklasyfikowane jako problematyczne poddano ocenie funkcjonowania w modelu 2PLM oraz 3PLM. Spośród 9 zidentyfikowanych zadań (Tabela 1) wybrano jedno w celu zilustrowania postawionego problemu. Przedstawiono je w studium przypadku[4]. W dalszych częściach artykułu porównamy ICC w modelach 2PLM i 3PLM dla wybranego zadania. Gdyby się okazało, że model z poprawką na zgadywanie (3PLM) będzie lepiej dopasowany do danych niż model 2PLM, zasadne będzie twierdzenie, że błędne zakwalifikowanie zadania zamkniętego jako otwartego może doprowadzić do wyboru nietrafnej techniki analizy danych. Założenie, że dane zadanie ma charakter otwarty, wyklucza bowiem zastosowanie modelu 3PLM, który umożliwia wprowadzenie poprawki na zgadywanie.

[4] Siedem zadań ma charakter wielopunktowy i dla nich skorzystano z modelu GRM (*graded response model*). W takich zadaniach o potrzebie uwzględnienia parametru zgadywania świadczy niedoszacowanie przez model prawdopodobieństwa uzyskania określonej punktacji przez uczniów wykazujących niski poziom umiejętności.

Wyniki

W Tabeli 1 przedstawiono obszary standardów i umiejętności sprawdzane przez 9 problematycznych zadań, a także określenia ich rodzajów w trzech systemach klasyfikacyjnych: Niemierki (1975), PISA 2012 (OECD, 2013) oraz Downinga (2009). Według klasyfikacji CKE wszystkie te zadania są otwarte. Zgodnie z klasyfikacją Niemierki wśród analizowanych zadań występują następujące typy: (a) zamknięte na dobieranie/ /przyporządkowanie (D1); (b) zamknięte wielokrotnego wyboru z jedną odpowiedzią prawdziwą (WW1) oraz (c) zamknięte typu prawda/fałsz (PF1). Według klasyfikacji PISA 2012 (OECD, 2013) wszystkie zadania można przypisać do typu złożonych zadań zamkniętych wielokrotnego wyboru (*complex multiple-choice*). Według Downinga (2009) wśród analizowanych zadań znajdują się tradycyjne zadania zamknięte na dopasowanie/przyporządkowanie (*traditional matching*, TM), zadanie zamknięte wielokrotnego wyboru (*multiple-choice item*, MCQ) oraz wielokrotne zadanie zamknięte typu prawda/fałsz (*multiple true-false*, MTF). Konkludując, wszystkie zadania ujęte w Tabeli 2 są zadaniami zamkniętymi. Klasyfikowanie zadania przez CKE jako otwarte prawdopodobnie wynika z uznawania za otwarte każdego

zadania, w którym uczeń wpisuje cokolwiek w arkuszu. Tymczasem wszystkie znane z literatury klasyfikacje za cechę definicyjną zadań otwartych uznają samodzielne wytworzenie odpowiedzi przez ucznia (wpisywanie tej odpowiedzi w arkuszu uważa się za drugorzędne).

Analiza ilościowa i studium przypadku[5]

Wybrane do studium przypadku zadanie pochodzi z części matematyczno-przyrodniczej arkusza standardowego egzaminu gimnazjalnego z 2011 r. Analizowane zadanie 33 (CKE, 2011a) składa się z trzech części (Tabela 1).

Poprawne odpowiedzi na poszczególne części tego zadania są następujące: (1) Natężenie prądu elektrycznego wzrosło. (2) Opór elektryczny opornika nie zmienił się. (3) Moc opornika wzrosła 4 razy. W zadaniu zastosowano następujące kryteria oceniania: za poprawne uzupełnienie 3 zdań uczeń uzyskiwał 3 punkty, za dwa poprawnie uzupełnione zdania – 2 punkty i jedno poprawnie uzupełnione zdanie – 1 punkt (CKE, 2011b). Egzaminatorzy na karcie odpowiedzi odnotowywali oddzielnie punktację odpowiedzi poprawnych dla kolejnych zdań.

Druga i trzecia część tego zadania zostały zanalizowane przez porównanie dopasowania modelu dwuparametrycznego i trzyparametrycznego IRT[6]. W literaturze wskazuje się na wskaźnik G^2 (McKinley i Mills, 1985), oparty na ilorazie wiarygodności (*likelihood ratio*), oraz χ^2 Pearsona jako miary dobroci dopasowania dla modeli 2PLM i 3PLM. W niniejszym artykule decydujemy się jednak z nich nie korzystać ze względu na trudności z poprawnym oszacowaniem błędu pierwszego rodzaju (Orlando i Thissen, 2000). Co więcej, te miary dopasowania mówią o ogólnym dopasowaniu modelu do danych (Maydeu-Olivares, 2013), natomiast w przeprowadzonej analizie ważne jest przede wszystkim lokalne niedopasowanie do danych, które wskazuje na niedoszacowanie lub przeszacowanie prawdopodobieństwa udzielenia odpowiedzi przez uczniów o konkretnym poziomie umiejętności.

Z uwagi na bardzo niski poziom korelacji pierwszej części zadania 33 z sumą punktów z całego testu po wykluczeniu danego zadania (0,0414) nie został dla niej oszacowany model IRT. Rysunek 4 wskazuje na proporcję poprawnych oraz niepoprawnych

[5] Do analizy funkcjonowania wybranego zadania wykorzystano program irt.ado, będący dodatkiem do programu STATA.

[6] Korelacja drugiej części z sumą punktów z całego testu po wykluczeniu danego zadania wyniosła 0,0449. Korelacja trzeciej części z sumą punktów po wykluczeniu danego zadania wyniosła 0,2214.

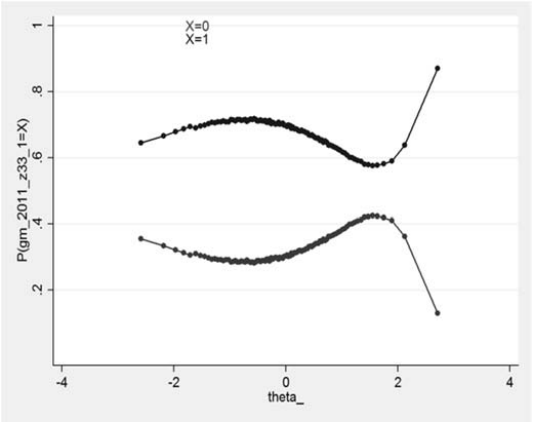
Tabela 1
Treść zadania 33 z części matematyczno-przyrodniczej arkusza egzaminu gimnazjalnego z 2011 r.

Zadanie 33. (0–3) Jacek zastąpił baterię w obwodzie dwiema takimi samymi bateriami połączonymi szeregowo. Zauważył wówczas, że napięcie na oporniku wzrosło dwukrotnie. Uzupełnij zdania.
Natężenie prądu elektrycznego..... . wzrosło/ nie zmieniło się/zmalało
Opór elektryczny opornika ..... . wzrósł/nie zmienił się/zmalał
Moc opornika..... ..... razy. wzrosła/zmalała 2/4

Tabela 2
Zidentyfikowane zadania problematyczne (wraz z obszarami standardów) i ich rodzaje w wybranych klasyfikacjach zadań

Rok	Nr zadania	Obszar standardów	Nazwa sprawdzanej umiejętności	Nazwa sprawdzanej czynności	Rodzaj i forma zadania wg Niemierki (1975)	Rodzaj i forma zadania wg PISA 2012 (OECD, 2013)	Rodzaj i forma zadania wg Downinga (2009)
2003	31	I	stosowanie terminów i pojęć matematyczno-przyrodniczych	wybieranie właściwych terminów do opisu obiektów przyrodniczych	zadanie zamknięte na dobieranie/ /przyporządkowanie (D1)	złożone pytanie zamknięte wielokrotnego wyboru (*complex multiple-choice*)	tradycyjne zadanie zamknięte na dopasowywanie/ /przyporządkowanie *(traditional matching, TM)*
2005	27	II	selekcjonowanie informacji	lokalizowanie na mapie państw sąsiadujących z Polską	zadanie zamknięte na dobieranie/ /przyporządkowanie (D1)	złożone pytanie zamknięte wielokrotnego wyboru	tradycyjne zadanie zamknięte na dopasowywanie/ /przyporządkowanie
2007	27	II	operowanie informacją	Interpretowanie informacji przedstawionych na schemacie	zadanie zamknięte typu prawda/fałsz (PF1)	złożone pytanie zamknięte wielokrotnego wyboru	wielokrotne zadanie zamknięte typu prawda/ fałsz (*multiple true-false*, MTF)
2007	28	IV	tworzenie modelów sytuacji problemowej	dobieranie wykresów ilustrujących charakter zależności wysokości poziomu wlewanej do naczyń wody od czasu	zadanie zamknięte na dobieranie/ /przyporządkowanie (D1)	złożone pytanie zamknięte wielokrotnego wyboru	tradycyjne zadanie zamknięte na dopasowywanie/ /przyporządkowanie
2008	29	II	operowanie informacją	• analizowanie schematu obwodu elektrycznego; • opisywanie stanu wyłączników, przy którym prąd elektryczny płynie przez część obwodu; 3. określanie, czy urządzenie będzie pracować przy zadanym stanie wyłączników	zadanie zamknięte na dobieranie/ /przyporządkowanie (D1)	złożone pytanie zamknięte wielokrotnego wyboru	tradycyjne zadanie zamknięte na dopasowywanie/ /przyporządkowanie

2009	26	II	operowanie informacją	selekcjonowanie informacji	zadanie zamknięte na dobieranie/ /przyporządkowanie (D1)	złożone pytanie zamknięte wielokrotnego wyboru	tradycyjne zadanie zamknięte na dopasowywanie/ /przyporządkowanie
2010	35	III	wskazywanie prawidłowości w procesach, w funkcjonowaniu układów i systemów	nazywanie procesów warunkujących obieg węgla w biosferze	zadanie zamknięte na dobieranie/ /przyporządkowanie (D1)	złożone pytanie zamknięte wielokrotnego wyboru	tradycyjne zadanie zamknięte na dopasowywanie/ /przyporządkowanie
2011	31	III	stosowanie zintegrowanej wiedzy do objaśnienia zjawisk przyrodniczych	wskazywanie zależności między działalnością człowieka a jej przyrodniczymi uwarunkowaniami	zadanie zamknięte na dobieranie/ /przyporządkowanie (D1)	złożone pytanie zamknięte wielokrotnego wyboru	tradycyjne zadanie zamknięte na dopasowywanie/ /przyporządkowanie
2011	33	III	wskazywanie prawidłowości w procesach, w funkcjonowaniu układów i systemów	wykorzystywanie zasad i praw do objaśniania zjawisk	zadanie zamknięte wielokrotnego wyboru z jedną odpowiedzią prawdziwą (WW1)	złożone pytanie zamknięte wielokrotnego wyboru	zadania zamknięte wielokrotnego wyboru (*multiple-choice item, MCQ*)



Rysunek 4. Rozkład prawdopodobieństwa poprawnej odpowiedzi dla pierwszej części zadania 33.

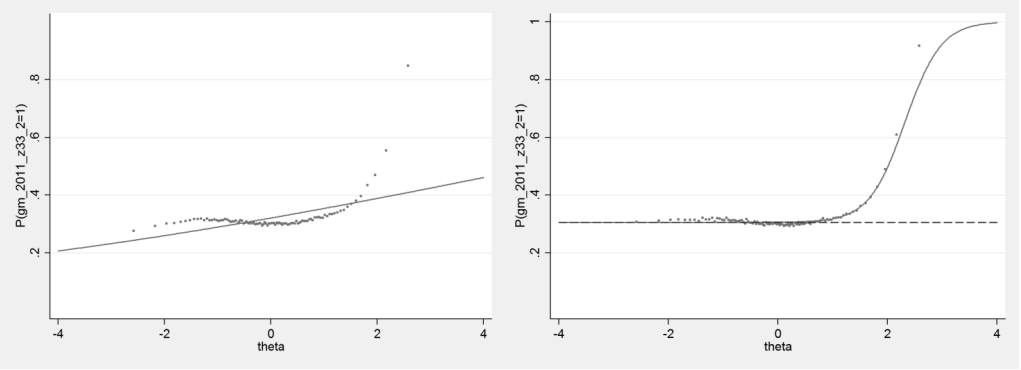
odpowiedzi (odpowiednio: $X = 0$ oraz $X = 1$) dla pierwszej części zadania 33 w kolejnych centylach poziomu umiejętności na skali zmiennej ukrytej θ. Poziom umiejętności egzaminowanych uczniów na skali θ wykorzystany do sporządzenia wykresu został oszacowany na podstawie modelu IRT dopasowanego do pozostałych zadań egzaminu.

Dla pierwszej części zadania 33 funkcja opisująca rozkład prawdopodobieństwa udzielenia poprawnej odpowiedzi nie rośnie monotonicznie, co wyraźnie narusza założenia modelu IRT (tłumaczy to, dlaczego nie dopasowano do tej części zadania krzywej IRT oraz dlaczego zadanie tak nisko korelowało z wynikiem sumarycznym w reszcie testu). Dla najniższego poziomu umiejętności prawdopodobieństwo udzielenia odpowiedzi poprawnej wzrasta, po czym maleje dla średniego poziomu umiejętności uczniów, by znowu wzrosnąć dla uczniów o najwyższym poziomie zdolności.

Proporcje poprawnych odpowiedzi ($X = 1$) w określonych przedziałach umiejętności uczniów dla pierwszej części wskazują na niezerowe prawdopodobieństwo udzielenia odpowiedzi poprawnych na to zadanie niezależnie od poziomu umiejętności, wynikające najpewniej z mechanizmu odgadywania odpowiedzi poprawnej. Poziom zgadywania dla pierwszej części wynosi około 0,6[7].

Porównanie dopasowania modelu dwu- i trzyparametrycznego IRT dla drugiej części zadania wskazuje na skrajne niedopasowanie modelu dwuparametrycznego. Jak pokazuje Rysunek 5, model 2PLM nie doszacowuje prawdopodobieństwa udzielenia poprawnej odpowiedzi przez uczniów o niskim i wysokim poziomie umiejętności (θ). Dla uczniów o średnim poziomie umiejętności prawdopodobieństwo udzielenia poprawnej odpowiedzi jest przeszacowane. Po oszacowaniu parametrów zadania za pomocą modelu 3PLM okazało się, że moc dyskryminacyjna tej części zadania jest wyższa, niż by to wynikało z oszacowań uzyskanych dzięki modelowi 2PLM.

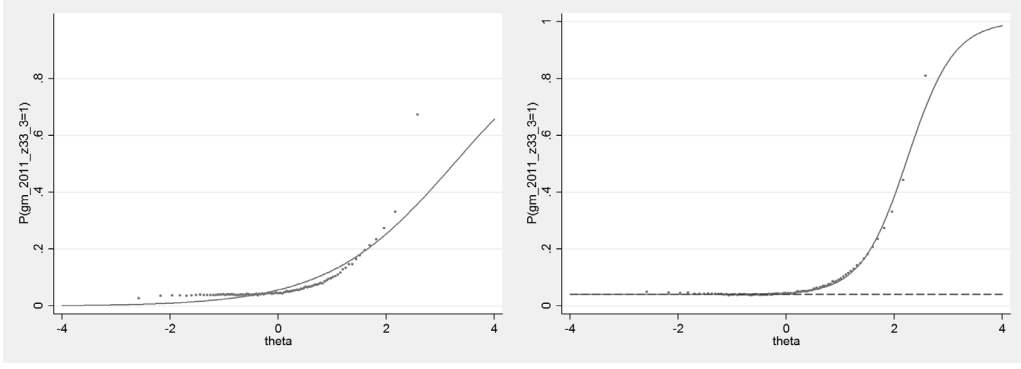
[7] Kształt krzywej wskazujący na spadek prawdopodobieństwo udzielenia poprawnej odpowiedzi dla uczniów o wysokim (ale nie najwyższym) poziomie umiejętności (w porównaniu do uczniów o niskim i średnim poziomie) może być konsekwencją wysokiego poziomu zgadywania i tego, że jest to skuteczniejsza strategia, niż próba rozwiązania tego zadania przez uczniów o wysokim poziomie umiejętności. Hipoteza ta, a także próba odpowiedzi na pytanie o to, jakie strategie odpowiadania na zadanie przyjmują uczniowie o wysokim poziomie umiejętności, wymagają osobnego sprawdzenia.



Rysunek 5. Krzywa charakterystyczna dla zadania 33_2 w modelu dwuparametrycznym (2PLM) i trzyparametrycznym (3PLM) (w modelu 2PLM *a* = 0,1481, *b* = 5,0461; w modelu 3PLM *a* = 3,0506, *b* = 2,3192, *c* = 0,3068).

Porównanie dopasowania modelu dwu- i trzyparametrycznego IRT dla trzeciej części zadania ujawnia duże niedopasowanie modelu 2PLM. Jak można zauważyć na Rysunku 6, dla uczniów o niskim poziomie umiejętności (θ) model dwuparametryczny nie doszacowuje prawdopodobieństwa udzielenia poprawnej odpowiedzi przez ucznia na to zadanie. Dla uczniów o średnim poziomie umiejętności (θ) prawdopodobieństwo udzielenia poprawnej odpowiedzi jest z kolei przeszacowane. Dla uczniów o najwyższym poziomie umiejętności (θ) mamy do czynienia z rosnącym niedoszacowaniem prawdopodobieństwa poprawnej odpowiedzi. Dla kontrastu wykorzystanie modelu 3PLM prowadzi do trafnego oszacowania prawdopodobieństwa udzielenia poprawnej odpowiedzi na zadanie, w zależności od poziomu umiejętności uczniów (lekkie niedopasowanie można zaobserwować tylko dla najwyższego poziomu umiejętności (θ)).

Gdyby to zadanie zostało zaklasyfikowane poprawnie, wszystkie części powinny mieć formę zadań zamkniętych wielokrotnego wyboru. Według ekspertów zajmujących się konstrukcją zadań gimnazjalnych



Rysunek 6. Krzywa charakterystyczna zadania 33_3 w modelu dwuparametrycznym (2PLM) i trzyparametrycznym (3PLM) (w modelu 2PLM *a* = 0,8635 i *b* = 3,2484, w modelu 3PLM *a* = 2,3467, *b* = 2,2455 i *c* = 0,0411).

z części matematyczno-przyrodniczej (Tyralska-Wojtycza, 2010), zmiana ta nie powinna zmniejszać możliwości zadania w zakresie pomiaru złożonych umiejętności. Przykładowo, trzecia część zadania powinna wyglądać następująco:

Moc opornika:
A. wzrosła 2 razy
B. wzrosła 4 razy
C. zmalała 2 razy
D. zmalała 4 razy

Potraktowanie tego zadania jako otwartego wyklucza uwzględnienie w analizie jego funkcjonowania parametru zgadywania. Biorąc pod uwagę wyłącznie poziom wykonania pierwszej części zadania 33, należałoby stwierdzić, że jest łatwe – prawie 70% uczniów rozwiązuje je poprawnie (dla drugiej części zadania poziom poprawnego wykonania zadania wyniósł 32%, a dla trzeciej zaledwie 8%). Na tej podstawie nauczyciel mógłby stwierdzić, że badana umiejętność jest opanowana. Uwzględnienie informacji o tym, że nawet uczeń o minimalnym poziomie umiejętności ma około 60% szans na odgadnięcie poprawnej odpowiedzi, zmienia interpretację wyników. Nie można więc z dostateczną dozą pewności przypisać sukcesu uczniów ich wysokim umiejętnościom, gdyż wynik odzwierciedla również komponent zgadywania.

Poza psychometrycznymi konsekwencjami niepoprawnego przypisania rodzaju i formy zadania dla analizy możliwe jest także obniżenie jego trafności. W trzeciej części zadania 33. uczeń otrzymuje 1 punkt, gdy poprawnie wskaże jednocześnie dwie (z czterech możliwych) odpowiedzi (CKE, 2011b). Przyglądając się wynikom uczniów, nie mamy pewności, czy kod 0 odnosi się do sytuacji, w której:

- uczeń opuścił jedną lub obie luki w zadaniu,
- uczeń wpisał złą odpowiedź w jednej lub obu lukach.

Szczegółowe przyjrzenie się sposobowi udzielania przez uczniów odpowiedzi na zadanie 33 wykazało, że pojawiła się tu znikoma liczba opuszczeń. Rozkład wyborów poszczególnych wariantów odpowiedzi możemy oszacować na podstawie losowo wybranych arkuszy egzaminacyjnych z krakowskiej (100 prac) i wrocławskiej (112 prac) okręgowej komisji egzaminacyjnej[8]. Uczniowie najczęściej uzupełniali luki w następujący sposób: „moc opornika wzrosła 2 razy" (62% w OKE Kraków; 66% w OKE Wrocław). Sugeruje to, że egzaminowani rozumieją relacje pomiędzy napięciem, natężeniem i mocą, jednak nie mają liczbowej intuicji na temat wielkości tej relacji. Może o tym także świadczyć niższy odsetek wariantów uwzględniających odpowiedź „moc zmalała…". Wśród odpowiedzi „moc zmalała…" więcej uczniów wybiera wariant „moc zmalała 2 razy" (14% dla OKE Kraków i 13% dla OKE Wrocław) niż „moc zmalała 4 razy" (jedynie 2% w obydwu OKE), co tym bardziej sugeruje, że uczniowie nie rozumieją liczbowych relacji pomiędzy tymi wielkościami.

Z perspektywy przedmiotu nauczania zadanie sprawdzało znajomość prawa Ohma (I i II część zadania) oraz zrozumienie, od czego zależy moc opornika w opisanym w zadaniu obwodzie. Zdaniem praktyków uczących fizyki w gimnazjum z pierwszą częścią zadania (pytanie o zmianę natężenia prądu – wzrost/spadek po dołączeniu szeregowo drugiego ogniwa) uczniowie nie powinni mieć problemu. Powinni rozumieć i pamiętać tę zależność dzięki przeprowadzonemu doświadczeniu. Rzeczywiście, w tej części zadania poprawną odpowiedź zarejestrowano u prawie 70% uczniów osiągających zarówno bardzo słaby, jak i dobry wynik w całym teście. W drugiej części

[8] Informacja o sposobie wykonania poszczególnych elementów zadania przez ucznia przy potraktowaniu go jako zadanie otwarte nie jest dostępna w analizie po zakończeniu oceniania.

zadania 32% uczniów udzieliło poprawnej odpowiedzi. Trzecia część zadania okazała się dla gimnazjalistów bardzo trudna (8% poprawnych odpowiedzi). Dla nauczycieli nie było to zaskoczeniem. Z 11 godzin zwykle przeznaczanych na dział „Prąd elektryczny" zwykle tylko jedna jest poświęcana na pracę i moc prądu elektrycznego.

Potraktowanie tego zadania jako otwartego (rezygnacja z rejestrowania wyboru odpowiedzi i szacowania zgadywania) ograniczyło informację pozyskiwaną zarówno z próbnego zastosowania zadań podczas prac nad konstruowaniem arkusza egzaminacyjnego, jak i z samego egzaminu, redukując jego funkcję dostarczania informacji o efektach uczenia się (por. Szaleniec i Dolata, 2012).

Podsumowanie i wnioski

Dobór modelu analizy danych egzaminacyjnych powinien odpowiadać rodzajom zadań w strukturze arkusza egzaminacyjnego. Jeśli zadanie zostało zaklasyfikowane w kartotece testu jako otwarte, to logicznym wyborem w analizie danych jest model jedno- lub dwuparametryczny. Rzeczywiste zadania otwarte w niewielkim stopniu są podatne na zgadywanie. Nie ma powodu ani możliwości technicznej oszacowania na podstawie zapisu uczniowskiej odpowiedzi parametru zgadywania. W sytuacji opisanej w tym artykule, błędne określenie rodzaju zadania (zadanie otwarte z luką), które w rzeczywistości jest wiązką zadań zamkniętych wielokrotnego wyboru (WW), prowadzi do zastosowania modelu innego niż 3PLM i nieuwzględnienia parametru zgadywania. To z kolei wpływa na niewłaściwe oszacowanie prawdopodobieństwa udzielenia przez ucznia poprawnej odpowiedzi na zadanie w zależności od jego poziomu umiejętności. W konsekwencji umiejętności ucznia zostają oszacowane niewłaściwie. Nie bierze się pod uwagę tego, że nawet uczeń zupełnie nieznający odpowiedzi na zadanie, zgadując, ma pewną szansę wybrania poprawnej odpowiedzi.

Właściwe zakwalifikowanie zadania ze względu na formę ma też znaczenie dla organizacji egzaminu. Na podstawie znajomości formy zadania można szacować czas potrzebny uczniowi na jego rozwiązanie, a w konsekwencji na rozwiązanie całego testu. Czas rozwiązania poszczególnych zadań przekłada się na długość testu, a liczba zadań w teście ma bezpośrednie konsekwencje dla jego rzetelności. Właściwy dobór rodzaju zadania ma też konsekwencje techniczne, gdyż proces sprawdzania zadań zamkniętych jest zautomatyzowany. Oznaczenie zadania zamkniętego jako otwarte i ocenianie go przez egzaminatora generuje dodatkową wariancję wyniku pochodzącą z efektu egzaminatora. Dodatkowo koszt sprawdzania zadań otwartych w takim wypadku jest zawyżony.

Przy błędnie dobranym lub zaklasyfikowanym rodzaju zadania egzaminacyjnego pojawia się też problem z reprezentatywnością zadań testowych dla całego zestawu mierzonych umiejętności. Test idealny powinien zawierać różne rodzaje zadań, dobrze sprawdzające się w mierzeniu konkretnych procesów umysłowych. Zbyt niska reprezentacja zadań otwartych w teście (co ma miejsce w przypadku oznaczenia zadania de facto zamkniętego jako otwartego) może oznaczać, że część umiejętności, które można dobrze mierzyć zadaniami otwartymi, nie jest w teście sprawdzana. Potencjał zadań otwartych w zakresie mierzenia specyficznych umiejętności uczniów zostaje zmarnowany. Jak wskazują badania, rodzaj zadania wpływa na jego trudność, a zadania otwarte są zwykle trudniejsze niż zadania zamknięte (Hohensinn i Kubinger, 2011; In'nami i Koizumi, 2009). Nadreprezentacja w teście zadań zamkniętych (w stosunku do zaplanowanych i uwidocznionych w kartotece testu) może prowadzić do tego, że test

w rzeczywistości sprawdza poziom wiedzy uczniów o niższym poziomie umiejętności. Inne badanie (DeMars, 2000) wskazuje, że chłopcy uzyskują wyższe wyniki w zadaniach zamkniętych, a dziewczynki – w otwartych. Zbyt duża liczba zadań zamkniętych może prowadzić do pogłębienia różnic w wynikach chłopców i dziewczynek z zakresu przedmiotów przyrodniczych (chłopcy zwykle uzyskują w nich nieznacznie wyższe wyniki). Wszystko to sprawia, że ograniczona może być możliwość uogólniania wyników testu na całe spectrum umiejętności uczniów.

Podsumowując, błędny wybór, a także niepoprawne oznaczenie rodzaju zadania egzaminacyjnego może obniżać trafność wyników egzaminu gimnazjalnego oraz mieć psychometryczne konsekwencje dla ich analizy. To z kolei wiąże się z faktem, że wnioski wyciągane na podstawie wyników testu mogą być nieadekwatne, a w związku z tym nie powinny być podstawą podejmowania decyzji. Ma to znaczenie nie tylko dla omawianego egzaminu gimnazjalnego, którego wyniki decydują o dalszych losach edukacyjnych uczniów, co zalicza go do egzaminów wysokiej stawki (*high-stakes test*), ale także dla procesu nauczania. Jeśli przyjmiemy, że pojęcie trafności testu odnosi się do stopnia, w jakim dane empiryczne oraz teoria uzasadniają interpretację wyników, to procesowi walidacji powinien być poddany kierunek i zakres tej interpretacji (AERA, APA i NCME, 2007). Błędne zakwalifikowanie zadań wielokrotnego wyboru jako zadań otwartych prowadzi do nieuprawnionych wniosków na temat poziomu opanowania umiejętności, do których pomiaru zaplanowano dane zadania, i typów popełnianych przez uczniów błędów. Poza tym zaklasyfikowanie wskazanych w artykule zadań zamkniętych jako zadania otwarte narusza wytyczne zawarte w dokumencie *Przygotowanie propozycji pytań, zadań i testów do przeprowadzenia sprawdzianu i egzaminu gimnazjalnego* (CKE, 2005a). Wytyczne wymagały równej liczby punktów za zadania zamknięte i otwarte w egzaminie gimnazjalnym w części matematyczno-przyrodniczej (po 25 punktów za zadania obydwu typów). Wbrew temu za zadania zamknięte przyznawano więcej punktów, niż należało.

Niniejsze badanie ma charakter eksploracyjny. Potrzebne są dalsze badania pogłębiające wiedzę o wpływie rodzaju zadania na oszacowanie umiejętności uczniów. Jedno z nich mogłoby polegać na wykonaniu opisanego w tym artykule testu przez dwie losowe grupy uczniów. W jednej grupie problematyczne zadanie 33 miałoby oryginalną formę, w drugiej – formę zamkniętą wielokrotnego wyboru. Pozwoliłoby to porównać obie grupy pod względem uzyskanych wyników i określić obciążenie obu wersji testu zróżnicowanym funkcjonowaniem zadań (*differential item functioning*, DIF).

Literatura

Ackerman, T. A. i Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement*, *12*(2), 117–128.

AERA, APA i NCME. (2007). *Standardy dla testów stosowanych w psychologii i pedagogice*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Birnbaum, A. (1968). Some latent trait models. W: F. M. Lord i M. R. Novick (red.), *Statistical theories of mental test scores* (s. 397–472). Reading, MA: Addison–Wesley.

CKE (2005a). *Przygotowanie propozycji pytań, zadań i testów do przeprowadzenia sprawdzianu i egzaminu gimnazjalnego* [Niepublikowane procedury ustalone na zebraniu dyrektorów CKE i OKE w dniu 24 listopada 2005 r.].

CKE (2005b). *Egzamin gimnazjalny 2005. Sprawozdanie*. Pobrano z http://cke.edu.pl/images/stories/Sprawozdania2005/egzamin%20gimnazjalny%202005%20sprawozdanie.pdf

CKE (2007). *Osiągnięcia uczniów kończących gimnazjum w roku 2007. Sprawozdanie z egzaminu gimnazjalnego 2007*. Pobrano z http://cke.edu.pl/images/stories/EGZ_GIMN_07_SPRAW.zip

CKE (2008).*Osiągnięcia uczniów kończących gimnazjum w roku 2008. Sprawozdanie z egzaminu gimnazjalnego 2008*. Pobrano z http://cke.edu.pl/images/stories/Egz_gimn_2008_sprawozdanie.rar

CKE (2009).*Osiągnięcia uczniów kończących gimnazjum w roku 2009. Sprawozdanie z egzaminu gimnazjalnego 2009*. Pobrano z http://cke.edu.pl/images/stories/Wyniki_09/raport_gimnazjum_2009.pdf

CKE (2010).*Osiągnięcia uczniów kończących gimnazjum w roku 2010. Sprawozdanie z egzaminu gimnazjalnego 2010*. Pobrano z http://cke.edu.pl/images/stories/001_Gimnazjum/spr_gimn_2010.pdf

CKE (2011a). *Egzamin w klasie trzeciej gimnazjum z zakresu przedmiotów matematyczno-przyrodniczych (GM-1-112).* Pobrano z http://cke.edu.pl/images/stories/0001_Gimnazja_2011/mat/gm-1-112.pdf

CKE (2011b). *Egzamin gimnazjalny 2011 część matematyczno-przyrodnicza. Klucz punktowania zadań*. Pobrano z http://www.cke.edu.pl/images/stories/0001_Gimnazja_2011/mat/klucz%20punktowania_1.pdf

CKE (2011c).*Osiągnięcia uczniów kończących gimnazjum w roku 2011. Sprawozdanie z egzaminu gimnazjalnego 2010*. Pobrano z http://cke.edu.pl/images/stories/0001_Gimnazja_2011/spr_gim.pdf

Cronbach, L. J. (1984). *Essential of psychological testing* (wyd. 4). New York: Harper&Row.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55–77.

Downing, S. M. (2009). Written tests: constructed-response and selected-response formats. W: S. M. Downing i R. Yudkowsky (red.), *Assessment in Health Professions Education* (s. 149–185). New York, NY: Routledge.

Ebel, R. L. i Frisbie, D. A. (1991). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice Hall.

Haladyna, T. M., Downing, S.M. i Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement In Education*, *15*(3), 309–334.

Hambleton, R. K., Swaminathan, H., i Rogers, H. J. (1991*). Fundamentals of item response theory*. Newbury Park, CA: Sage.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation*, *17*(1). Pobrano z http://pareonline.net/pdf/v17n1.pdf

Hayes, J. R., i Flower, L. S. (1980). Identifying the organization of writing processes. W: E. W. Gregg i E. R. Steinberg (red.), *Cognitive processes in writing* (3–30). Hillsdale, NJ: Lawrence, Erlbaum.

Hohensinn, C. i Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response format. *Educational and Psychological Measurement*, *71*(4), 732–746.

In'nami, Y., i Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244.

Kondratek, B. i Pokropek, A. (2013). IRT i pomiar edukacyjny. *Edukacja*, *124*(4), 42–66.

Linden, W. J. van der i Hambleton, R. K. (red.). (1997). *Handbook of modern item response theory*. New York: Springer.

Lord, F. M. (1974) Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*(2), 247–264.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspective*, *11*(3), 71–101.

McKinley, R., i Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*(1), 49–57.

MEN (1999). *Rozporządzenie Ministra Edukacji Narodowej i Sportu z dnia 15 lutego 1999 r. w sprawie podstawy programowej kształcenia ogólnego* (Dz.U. 1999 Nr 14, poz. 129). Pobrano z http://isap.sejm.gov.pl/DetailsServlet?id=WDU19990140129

MEN (2007). *Standardy wymagań będące podstawą przeprowadzania egzaminu w ostatnim roku nauki w gimnazjum*. Załącznik do rozporządzenia Ministra Edukacji Narodowej z dnia 28 sierpnia 2007 r. (Dz.U. z 2007 r. Nr 157, poz. 1102). Pobrano z http://www.gim-nt.com/download/rozporzadzenie_28082007.pdf

Neidorf, T. S., Binkley, M., i Stephens, M. (2006). *Comparing science content in the National Assessment of Educational Progress (NAEP) 2000 and Trends in International Mathematics and Science Study (TIMSS) 2003 Assessments (NCES 2006–026)*. Pobrano z http://nces.ed.gov/pubsearch

Niemierko, B. (1975). *ABC Testów osiągnięć szkolnych*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.

Niemierko, B. (1999). *Pomiar wyników kształcenia*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.

OECD (2013). *PISA 2012 Assessment and analytical framework: mathematics, reading, science,*

problem solving and financial literacy. Pobrano z http://dx.doi.org/10.1787/9789264190511-en

Okręgowa Komisja Egzaminacyjna w Krakowie (bdw.). *Zadania egzaminacyjne i ich ocenianie*. Pobrano z http://www.oke.krakow.pl/szkolenia/materialy_III.pdf

Orlando, M. i Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.

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

Rauthmann, J. F. (2011). Not only item content but also item format is important: taxonomizing item format approaches. *Social Behavior and Personality*, *39*(1), 119–128.

Samejima, F. (1969). *Estimation of latent ability using a pattern of graded scores* [Psychometric Monograph nr 17]. Richmond, VA: Psychometric Society.

Szaleniec, H. i Dolata, R. (2012). Funkcje krajowych egzaminów w systemie edukacji. *Polityka Społeczna. Polityka edukacyjna: szanse i wyzwania. Nr 1 tematyczny*, 37–41.

Tyralska-Wojtycza, E. (2010). Nowa formuła egzaminu gimnazjalnego – strata czy zysk dla przedmiotów przyrodniczych? W: B. Niemierko i K. Szmigel (red.), *Teraźniejszość i przyszłość oceniania szkolnego.* Materiały z XVI Krajowej Konferencji Diagnostyki Edukacyjnej, Toruń, 22–24. 10. 2010 r. Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Wood, E. (2003). *What are extended matching sets questions? BEE-j (1,1).* Pobrano z http://bio.ltsn.ac.uk/journal/vol1/beej-1-2.htm

Podziękowania

Autorzy artykułu składają podziękowania Bartoszowi Kondratkowi za udostępnienie autorskiego oprogramowania do analiz z wykorzystaniem modeli IRT oraz Okręgowym Komisjom Egzaminacyjnym w Krakowie i we Wrocławiu za udostępnienie informacji o sposobach odpowiadania na zadanie 33 z 212 losowo wybranych arkuszy egzaminacyjnych (GM-1-112, 2011 r.).