Karolina Świst
Instytut Badań Edukacyjnych
k.swist@ibe.edu.pl

Paulina Skórska
Instytut Badań Edukacyjnych
p.skorska@ibe.edu.pl

Artur Pokropek
Instytut Badań Edukacyjnych
a.pokropek@ibe.edu.pl

Wykorzystanie mieszanych modeli IRT do detekcji egzaminowanych o niskiej motywacji

1. Wstęp

Szacowanie umiejętności uczniów bez wzięcia pod uwagę ich motywacji (zwłaszcza podczas rozwiązywania testów niskiej stawki, nie wiążących się z poważnymi konsekwencjami dla zdającego) może prowadzić do problemów z trafnością (np. występowaniem wariancji niezwiązanej z mierzonym konstruktem – *construct-irrelevant variance*) (Messick, 1989; Messick 1995). Jednym z rozwiązań tego problemu może być odfiltrowanie danych niezmotywowanych uczniów rozwiązujących test niskiej stawki (Setzer, Wise, van den Heuvel i Ling, 2013). Odfiltrowanie oznacza jednak w praktyce utratę danych w analizie. Ponadto wymaga silnego założenia, iż prawdopodobieństwo poprawnej odpowiedzi nie jest związane z poziomem umiejętności uczniów. Założenie to jest kluczowe, gdyż w przypadku jego złamania podczas odfiltrowywania respondentów słabo zmotywowanych, usuwani z próby są również badani o niskim poziomie cechy ukrytej, co w oczywisty sposób zaburzać będzie wyniki (Sundre i Wise, 2003). Innym sposobem na radzenie sobie z potencjalnym problemem obniżonej trafności wyników w testach niskiej stawki jest uwzględnienie informacji o obniżonej motywacji uczniów w modelu. Wykorzystane w niniejszym rozdziale mieszane modele IRT pozwalają na detekcję klas ukrytych uczniów odpowiadających na pytania testowe w sposób jakościowo różny oraz sprawdzenie do jakiego stopnia uwzględnienie informacji o motywacji uczniów wpływa na przynależność do klas.

2. Motywacja testowa

2.1 Motywacja do podejmowania testów niskiej stawki

Motywacja testowa określa, do jakiego stopnia egzaminowani są w stanie poświęcić maksimum wysiłku rozwiązywanemu testowi oraz w sposób rzetelny odzwierciedlić swój poziom umiejętności w

dziedzinie mierzonej przez dany test (Wise i DeMars, 2005, s. 2). Poziom motywacji testowej badanego może być zależny od konsekwencji, które niesie za sobą dany test. Jeśli test nie niesie ze sobą poważnych konsekwencji dla badanego (w postaci np. wpływu na ocenę lub przyjęcie do wybranej szkoły), uczeń może nie traktować testu poważnie, zgadywać odpowiedzi w zadaniach lub ich nie rozwiązywać (omijać zadania). W związku z tym trafność wyników i decyzji podjętych na podstawie wyników testu niskiej stawki (*low-stakes assessment*) może się obniżać (Swerdzewski, Finney i Harmes, 2007).

Należy zauważyć, że znane badania umiejętności uczniów, takie jak PISA (*Programme for International Student Assessment*) czy TIMSS (*Trends in International Mathematics and Science Study*) są testami niskiej stawki. Z drugiej strony na podstawie wyników tych badań, mogą być podejmowane decyzje dotyczące całego systemu edukacji. Dlatego też uwzględnienie poziomu motywacji testowej jest kwestią kluczową dla zapewnienia trafności wyników jakiegokolwiek badania umiejętności (Skórska, Świst i Pokropek, 2014).

2.2 Sposoby mierzenia motywacji testowej

Istnieją dwa główne podejścia do pomiaru motywacji testowej. Pierwsze podejście opiera się na miarach samoopisowych. Kwestionariusz Motywacji (*The Motivation Questionnaire*) stworzony przez Wolf i Smitha (1993) składa się z dwóch podskal: ważności *(importance*) mierzonej przez pięć pytań i wysiłku *(effort*) mierzonego przez trzy pytania. Na podstawie tego kwestionariusza, Sundre (1999) stworzyła kwestionariusz *Student Opinion Survey* (SOS), uzupełniając skalę wysiłku o dwa pytania i modyfikując konceptualizację motywacji. Ważność określa, w jakim stopniu wynik osiągnięty na teście ma dla ucznia znaczenie, natomiast wysiłek określa postrzegany stopień pracy lub wysiłku mentalnego niezbędny dla ukończenia danego testu (Sundre, 2007). Krótkie kwestionariusze motywacji zostały stworzone także przez Eklöf (Eklöf, 2006; Eklöf, 2007; Eklöf 2008) i są wykorzystywane do badania motywacji testowej w TIMSS. W Polsce na potrzeby badania zrównującego, prowadzonego przez Zespół Analiz Osiągnięć Uczniów, skonstruowano Skalę Motywacji Testowej (Węziak-Białowolska, 2011; Skórska, Świst i Pokropek, 2014). Jednakże trafność metod samoopisowych zależy w dużej mierze od tego, czy badani mają świadomość własnego poziomu motywacji oraz czy posiadają umiejętność odwzorowania tego poziomu przy pomocy skali (Swerdzewski, Harmes i Finney, 2011).

Bezpośrednim sposobem pomiaru motywacji testowej jest analiza obiektywnych miar zachowania uczniów podczas rozwiązywania testów. Miara *Response Time Effort* (RTE), przedstawiona przez Wise'a i Konga (2005) opiera się na założeniu, że niezmotywowani badani nie poświęcają wystarczającego wysiłku oraz czasu na dokładne przeczytanie zadań i udzielenie odpowiedzi. Z kolei Dolata i Pokropek (2010) analizowali wskaźnik motywacji testowej w postaci liczby braków danych w teście. Należy jednak pamiętać, że miara ta może być także związana z poziomem umiejętności

ucznia (uczniowie o niskim poziomie umiejętności będą przejawiać podobny wzorzec odpowiedzi co uczniowie niezmotywowani), więc powinien on być zawsze kontrolowany podczas analizy motywacji testowej.

W literaturze (np. Wolf i Smith, 1995; Sundre i Kitsantas, 2004; Wise i DeMars, 2005; Thelk i in., 2009) wskazuje się na pozytywny związek pomiędzy wynikami ucznia i motywacją do podejmowania testu niskiej stawki. Metaanaliza przeprowadzona przez Wise'a i DeMars (2005) również wskazuje na pozytywny związek pomiędzy poziomem motywacji testowej a wynikami testu. Średnia wielkość efektu (standaryzowana różnic pomiędzy średnimi) wynosi 0,59. Oznacza to, że zmotywowani uczniowie osiągają wyniki wyższe o ponad 0,5 odchylenia standardowego niż niezmotywowani uczniowie.

Podczas analizy zjawiska motywacji do podejmowania testów niskiej stawki, ważne jest więc rozgraniczenie pomiędzy poziomem umiejętności uczniów, a ich poziomem motywacji oraz wykorzystanie miar w mniejszym stopniu podatnych na zniekształcenia trafności. Rozwiązaniem może być wykorzystanie opisanych w poniższym podrozdziale mieszanych modeli IRT.

3. Mieszany model IRT (*Mixture IRT model*)

Mieszany model IRT łączy ze sobą dwa rodzaje modeli wykorzystywanych w pomiarach cechy ukrytej: modelowanie IRT (*Item Response Theory*) oraz analizę klas ukrytych (*Latent Class Analysis*), będącą szczególnym przypadkiem skończonych modeli mieszanek (*Finite Mixture Models*) (zob. rozdział o LCA).

W uproszczonej postaci jednowymiarowy model IRT może być zapisany jako (bardziej szczegółowa specyfikacja oraz opis modelu znajduje się w rozdziale 1):

$$f(\mathbf{u}) = \int f(\mathbf{u} \mid \theta) d\theta \tag{1}$$

gdzie $f(x)$ to funkcja gęstości prawdopodobieństwa losowego wektora $\mathbf{u}=[u_1, u_2,..., u_I]$ uzyskanego na podstawie całkowania warunkowych (ze względu na poziom ciągłej cechy ukrytej) funkcji gęstości $f(\mathbf{u} \mid \theta)$ po ciągłej cesze ukrytej. Prawdopodobieństwo prawidłowej odpowiedzi na pytanie jest zatem zależne od parametrów funkcji $f()$ oraz ciągłej, latentnej cechy respondentów, pochodzących z homogenicznej populacji. Można powiedzieć, iż różnice między prawdopodobieństwami poprawnej odpowiedzi miedzy poszczególnymi respondentami maja charakter *ilościowy*, gdyż zależą od różnicy w wartości ciągłej cechy ukrytej.

Inaczej jest w przypadku modelu klas ukrytych (Lazarsfeld i Henry, 1968, patrz też rozdział o LCA):

$$p(\mathbf{u}) = \sum_{c=1}^{C} \pi_c \prod_{i=1}^{I} p_{ic}^{u} (1 - p_{ic})^{1-u} \tag{2}$$

gdzie $p(\mathbf{u})$ to funkcja gęstości prawdopodobieństwa losowego wektora u=[u1, u2,..., uI]. Natomiast πc to prawdopodobieństwo przynależenia danego respondenta do klasy *c*, a p_{ic}^{u} i $(1-p_{ic})^{1-u}$ to kolejno prawdopodobieństwa sukcesu i porażki. Zarówno π_c i p_{ic}^{u} są parametrami modelu, które w normalnych warunkach badawczych są estymowane.

W modelu cech ukrytych zmienna latentna ma charakter nominalny. Zakłada się, że respondenci w tej samej klasie odpowiadają w na pytania w zbliżony do siebie sposób, zaś członkowie innych klas różnią się znacząco od siebie. Można przyjąć interpretację, że model klas ukrytych stosowany jest do wnioskowania na temat grup, które różnią się między sobą w *jakościowym* sensie.

Mieszany model IRT łączy te dwa podejścia, funkcja gęstości prawdopodobieństwa losowego wektora odpowiedzi na zadania w tym modelu opisana jest za pomocą rozkładów warunkowych zarówno ze względu na ciągłą i nominalną cechę ukrytą:

$$p(\mathbf{u}) = \sum_{c=1}^{C} \pi_c \int \prod_{i=1}^{I} p_i(\mathrm{u}_i \mid \theta, c)\, f_c(\theta) d\theta \qquad (3)$$

gdzie $f_c(\theta)$ to specyficzna, dla danej klasy *c*, funkcja gęstości prawdopodobieństwa określająca rozkład zmiennej θ w danej klasie oraz specyficzne parametry zadań. Taka specyfikacja modelu rozluźnia założenie, iż respondenci wylosowani są z homogenicznej populacji. Mieszany model IRT pozwala, aby respondenci pochodzili z różnych populacji i byli charakteryzowani przez różne wartości parametrów między klasami. W procesie estymacji każdemu respondentowi przypisana jest wartość ciągłej cechy ukrytej, jak i prawdopodobieństwo przynależności do każdej z klas. Model mieszany łączy tym samym możliwość wnioskowania *ilościowego* na temat poziomu danej cechy ukrytej i wnioskowania *jakościowego* pozwalającego na interpretację różnic miedzy jednostkami w heterogenicznej populacji.

4. Pytania badawcze

Celem niniejszej analizy jest przedstawienie zastosowania mieszanego modelu IRT do detekcji niezmotywowanych uczniów oraz ocena adekwatności tej metody. Do osiągnięcia tego celu, ważna jest odpowiedź na trzy, przedstawione poniżej pytania badawcze:

1. Czy mieszany model IRT jest lepiej dopasowany do danych niż klasyczny model IRT, a jeśli tak, to ile klas posiada?
2. Jeżeli model mieszany jest lepiej dopasowany to czy klasy ukryte tego modelu mogą być interpretowane w kategoriach motywacji testowej?

3. Czy użycie dodatkowych zmiennych polepsza dopasowanie modelu i ułatwia detekcję uczniów charakteryzujących się obniżoną motywacją?

5. Metodologia

W tym rozdziale przyjmujemy założenie, że badana populacja uczniów jest heterogeniczna ze względu na motywację testową. Załóżmy także, iż motywacja jest charakteryzowana poprzez nominalną cechę ukrytą, w najprostszym układzie dwukategorialną: uczniowie zmotywowani i uczniowie niezmotywowani. Założenia te możemy testować za pomocą mieszanego modelu IRT, gdzie przyjmiemy, że badana populacja składa się przynajmniej z dwóch klas, charakteryzujących dwie populacje uczniów. W każdej z klas prawdopodobieństwo udzielenia poprawnej odpowiedzi będzie zależało od ciągłej latentnej cechy ukrytej, lecz w każdej klasie relacja ta będzie wyglądała inaczej. Zróżnicowanie relacji osiągamy przez to, iż w każdej z klas parametry zadań (trudność i dyskryminacja) będą różne. Odnosząc się do pierwszego pytania badawczego, porównaliśmy, czy mieszany model IRT jest lepiej dopasowany do danych niż klasyczny model IRT.

Interpretując uzyskane klasy ukryte w kontekście motywacji testowej porównaliśmy parametry trudności i dyskryminacji dla każdej z klasy oraz średni wynik uzyskany w kwestionariuszu SOS mierzący motywację testową. Sprawdziliśmy ponadto do jakiego stopnia wykorzystanie informacji o motywacji testowej poprawia dopasowanie modelu oraz jakość klasyfikacji.

W celu odpowiedzi na trzecie pytanie badawcze, zastosowano mieszany model IRT z wykorzystaniem zmiennej współwystępującej w postaci sumy punktów uzyskanych w kwestionariuszu motywacji SOS. W przypadku modelu z informacją o motywacji zastosowano podejście single-step regression (regresji bazującej wyłącznie na jednym etapie) polegające na jednoczesnym włączeniu zmiennej współwystępującej do analizy podczas modelowania klas ukrytych. Choć w literaturze (Clark i Muthén, 2009) opisuje się także inne podejścia, takie jak: a) regresja na podstawie najbardziej prawdopodobnej przynależności do klasy ukrytej (*most likely class regression*) b) regresja na podstawie prawdopodobieństwa (*probablity regression*) przynależności do danej klasy z rozkładu *a posteriori* wyników i c) ważona regresja na podstawie prawdopodobieństwa (ważenie prawdopodobieństwem przynależności do danej klasy z rozkładu *a posteriori)*, ich wyniki mogą być obciążone. Regresje bazujące na najbardziej prawdopodobnej przynależności do klasy ukrytej mogą być obarczone błędem, ze względu na to, że przynależność do klasy jest traktowana jako obserwowalna zmienna, co oznacza, że jednostki niejako „na siłę" są przypisane do klasy ukrytej. Grozi to zakłóceniem wyników oraz niepoprawnym oszacowaniem błędów standardowych. Dlatego też przyjęto podejście opierające się na jednoczesnym włączeniu zmiennej współwystępującej do oszacowania klas ukrytych, które nie grozi opisanymi powyżej problemami. Formalnie model ze

zmienną współwystępującą można rozpisać zmieniając jeden element równania (3) π_c na $\pi_{c|Z}$ czyli prawdopodobieństwo przynależności do danej klasy określić jako prawdopodobieństwo warunkowe ze względu na zmienną współwystępującą (a ogólnie zmienne współwystępujące). Takie warunkowe prawdopodobieństwo może być dalej rozwinięte do odpowiedniej postaci funkcyjnej. Jako że prawdopodobieństwo przybiera wartości od 0 do 1, naturalnym w tym wypadku wydaje się użycie funkcji logistycznej. Prawdopodobieństwo przynależności do klasy ukrytej modelowane jest za pomocą M zmiennych współwystępujących (w omawianym wypadku M=1) zgodnie z następującym wzorem:

$$\frac{\pi_{c|Z}}{1-\pi_{c|Z}} = \beta_0 + \sum_{m=1}^{M} \beta_m Z_{jm} \tag{4}$$

Gdzie Z_{jm} to m-ta zmienna współwystępująca, parametry β_0 i β_m są parametrami regresji logistycznej, czyli stałą i współczynnikiem kierunkowym.

Odpowiedź na trzecie pytanie badawcze będzie polegała na porównaniu współczynnika entropii dla modelu bez zmiennej współwystępującej oraz modelu ze zmienną współwystępującą. Współczynnik entropii wskazuje na stopień odseparowania od siebie klas ukrytych uzyskany przy pomocy modelu. Wskaźnik entropii wylicza się w sposób następujący (Clark i Muthén, 2009):

$$E_k = 1 - \frac{\sum_i \sum_k (-\hat{p}_{ik} \ln \hat{p}_{ik})}{n \ln K} \tag{3}$$

Gdzie $\hat{p}_{ik}$ oznacza oszacowane prawdopodobieństwo przynależności do klasy k przez jednostkę i, n wielkość próby, a K liczbę klas ukrytych. Absolutna wartość entropii przyjmuje wartości od 0 do ∞, a większa wartość wskazuje na lepsze odseparowanie od siebie klas. W Mplusie zaimplementowano wskaźnik względnej entropii, który przejmuje wartości z przedziału [0,1] (Clark i Muthén, 2009). Ponownie, wyższa wartość sugeruje lepsze odseparowanie od siebie klas ukrytych, jednak w literaturze brakuje jasnych kryteriów dotyczących tego, jaka wartość entropii jest zadowalająca.

Podsumowując, w rozdziale tym zaprezentowano wyniki analizy przeprowadzonej w trzech krokach:

1) Porównanie dopasowania do danych mieszanego modelu IRT z dwoma klasami z dopasowaniem do danych klasycznego modelu IRT, zakładającego homogeniczność badanej populacji oraz z dopasowaniem mieszanego modelu IRT z trzema klasami (zob. pytanie badawcze nr 1);
2) Interpretacja modelu klas ukrytych (jeśli okaże się lepiej dopasowany do danych niż klasyczne modele IRT) w kontekście motywacji testowej, tj. porównanie średniego wyniku SOS w obydwu klasach ukrytych (zob. pytanie badawcze nr 2);

3) Porównanie dopasowania do danych oraz jakości klasyfikacji do klas ukrytych modelu bez zmiennej współwystępującej oraz modelu z SOS jako zmienną współwystępującą (zob. pytanie badawcze nr 3).

6. Dane

Do analizy wykorzystano dane pochodzące z badań zrównujących przeprowadzonych przez ZAOU (Zespół Analiz Osiągnięć Uczniów Instytutu Badań Edukacyjnych) w 2014 roku, m. in. na losowej próbie uczniów gimnazjum (1617 osób). W analizie wykorzystano dane z testu umiejętności matematycznych oraz kwestionariusza motywacji – polskiej wersji Skali Opinii Uczniów (Sundre, 2007), przetłumaczonej za zgodą autorki przez ZAOU.

Do analizy wykorzystano mieszany model IRT (2 PLM) dla zmiennych kategorialnych (szerzej w: Muthen i Muthen, 1998-2012; s. 198-199). Testowano modele z 1 klasą ukrytą (model bazowy), 2 klasami ukrytymi oraz 3 klasami ukrytymi. Analizę dobroci dopasowania przeprowadzono na podstawie wskaźników AIC, BIC, BIC skorygowanego ze względu na wielkość próby (*sample size adjusted* BIC) oraz dostępnego w programie Mplus testu Vuong-Lo-Mendell-Rubin opartego na ilorazie wiarygodności. Test ilorazu wiarygodności bazujący na wartości chi-kwadrat nie powinien być używany do testowania modeli z k-1 i k klasami z uwagi na to, że różnica funkcji wiarygodności dla obydwu tych modeli przemnożona przez 2 nie ma rozkładu chi kwadrat (Nylund, Asparouhov i Muthén, 2007). Test Voung-Lo-Mendell-Rubin (Lo, Mendell i Rubin, 2001) używa poprawnego rozkładu opartego na przemnożonej przez 2 różnicy funkcji wiarygodności. Opiera się na jednoczesnym testowaniu modelu z k klasami oraz k-1 klasami oraz wykorzystywaniu pochodnych obydwu modeli do wyliczenia poziomu istotności statystycznej (Asparouhov i Muthén, 2012).

7. Wyniki

7.1. Dobroć dopasowania mieszanego modelu IRT w porównaniu do klasycznego modelu IRT

W tabeli 1 przedstawiono statystyki dobroci dopasowania (AIC, BIC oraz BIC skorygowany na wielkość próby) dla modelu z: 1 klasą ukrytą (model bazowy – czyli klasyczny model IRT), 2 klasami ukrytymi oraz 3 klasami ukrytymi. Niższe wartości tych trzech współczynników wskazują na lepsze dopasowanie modelu do danych. Estymacja modeli trzyklasowych nie zakończyła się sukcesem, ze względu na słabe dopasowanie do danych oraz kłopoty z lokalnymi maksimami funkcji wiarygodności. W przypadku modelu z dwiema klasami ukrytymi wartości AIC i skorygowanego BIC wskazują na lepsze dopasowanie do danych niż analogicznego modelu jednoklasowego. W związku z tym, że wyniki nie są rozstrzygające, przeprowadzono także analizę wyników testu ilorazu wiarygodności Voung-Lo-Mendell-Rubin. Wynik testu jest istotny statystycznie (p=0,0360), co

oznacza, że model dwuklasowy dla matematyki jest lepiej dopasowany niż model jednoklasowy. Analiza wskaźników AIC, skorygowanego BIC oraz testu ilorazu wiarygodności pozwala na stwierdzenie, że w przypadku testu matematycznego można wyróżnić dwie klasy ukryte egzaminowanych. Kolejnym krokiem w analizie będzie więc próba ich interpretacji w kontekście motywacji testowej.

Tabela 1. Wskaźniki AIC, BIC oraz skorygowany BIC dla modeli z 1, 2 i 3 klasami ukrytymi.

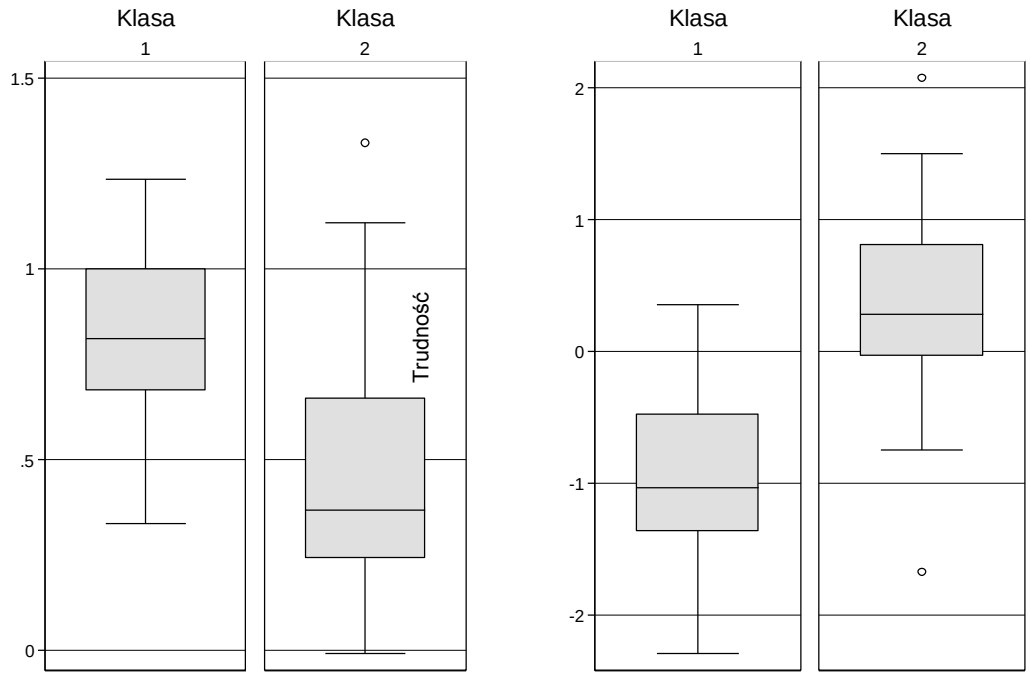
		Wartość wskaźnika
1 klasa	AIC	47920,657
	BIC	**48206,239**
	Skoryg. BIC	48037,867
2 klasy	AIC	**47710,763**
	BIC	48281,926
	Skoryg,BIC	**47945,183**
3 klasy	AIC	---
	BIC	---
	Skoryg.BIC	---

Uwaga: pogrubienie oznacza najniższą wartość wskaźnika, tym samym wskazuje preferowany model

7.2. Interpretacja klas ukrytych w kontekście motywacji testowej

Na rysunku 1. przedstawiono porównanie parametrów dyskryminacji oraz trudności zadań matematycznych dla 2 klas ukrytych. W pierwszej klasie parametry dyskryminacji zadań są znacznie wyższe od parametrów dyskryminacji estymowanych dla klasy 2. Wyraźną różnicę można także zauważyć w przypadku parametrów trudności: klasa 1 charakteryzuje się znacznie łatwiejszymi zadaniami, tymczasem w klasie 2 trudność zadań okazuje się większa. Interpretacja wielkości parametrów w kontekście motywacji testowej pozwala zakładać, iż klasa 1 powinna reprezentować uczniów o wyższej motywacji testowej, a 2 klasa uczniów o niższej motywacji testowej. W klasie 1 umiejętności uczniów są silniej związane z poprawną odpowiedzią na zadanie testowe, a znacząco niższa wartość parametrów trudności oznacza, iż uczniowie z tej klasy znacznie częściej odpowiadali poprawne na zadania niż uczniowie w klasie 2.

Oczywiście interpretacja różnic parametrów trudności i dyskryminacji w kategoriach motywacji testowej jest jedynie przypuszczeniem. Charakter różnic miedzy klasami ukrytymi może mieć różne podłoże, dlatego przeprowadziliśmy także analizę rozkładów wyników zmiennych pomocniczych w podziale na klasy ukryte. Wyniki w poniższej tabeli mogą przekonywać o możliwości interpretacji klas ukrytych w kategoriach motywacji testowej.



Rysunek 1. Wykres skrzynkowy parametrów zadań dla testu matematycznego

Tabela 2. Wybrane charakterystyki poszczególnych klas ukrytych

Klasy ukryte	% uczniów należących do danej klasy ukrytej	Średni wynik w kwestionariuszu SOS (max 50 pkt.)*	% dziewcząt w danej klasie ukrytej
1	**58,44**	**38,07**	51,32
2	**41,56**	**35,00**	49,11

W tabeli 2. przedstawiono proporcję dziewcząt i chłopców w poszczególnych klasach ukrytych – można przede wszystkim zauważyć, że płeć nie różnicuje w znaczący sposób przynależności do klas. Analiza wartości wskaźnika motywacji potwierdza nasze wcześniejsze przypuszczenia: mianowicie, pierwsza klasa ukryta ma wyższy średni wynik w kwestionariuszu SOS świadczący o wyższej motywacji. W klasie tej wyższe są współczynniki dyskryminacji oraz niższe parametry trudności dla zadań. Różnica pomiędzy klasami w średnim wyniku SOS [$t(1615)= 6,37$; $p< 0,00001$] jest istotna statystycznie.

7.3. Interpretacja mieszanego modelu IRT ze zmienną współwystępującą

Weryfikację tego, do jakiego stopnia dodatkowa zmienna (SOS) pozwala przewidzieć przynależność do danej klasy ukrytej, umożliwiła następna część analizy, w której przeprowadzono modelowanie klas ukrytych z uwzględnieniem zmiennej współwystępującej. W tabeli 3 zaprezentowano statystyki dobroci dopasowania (AIC, BIC oraz BIC skorygowany na wielkość próby) dla modeli z 1, 2 i 3 klasami ukrytymi, z uwzględnieniem wyników Skali Opinii Uczniów (SOS).

Wskaźniki BIC i skorygowany BIC wskazują na najlepsze dopasowanie do danych z testu matematycznego modelu dwuklasowego, natomiast wskaźnik AIC sugeruje lepsze dopasowanie modelu trzyklasowego. W przypadku modelu trzyklasowego estymacja nie przebiegła jednak do końca stabilnie ze względu na problemy z lokalnymi maksimami funkcji wiarygodności, które pojawiały się nawet w sytuacji zwiększenia liczby losowych wartości startowych. Zestawienie dobroci dopasowania modeli z SOS o danej liczbie klas ukrytych przedstawiono w poniższej tabeli. Test ilorazu wiarygodności Voung-Lo-Mendell-Rubin może być w tym wypadku nierozstrzygający – przyjmując tradycyjne kryterium istotności statystycznej ($p < 0,05$), można stwierdzić, że model jednoklasowy lepiej odzwierciedla strukturę danych niż dwuklasowy ($p=0,06$). Jednakże wynik testu jest na granicy istotności statystycznej, która także jest uwarunkowana wielkością próby, więc w tym wypadku przyjmiemy wskaźniki BIC i skorygowanego BIC za rozstrzygające.

Tabela 3. Porównanie dopasowania modelu IRT z modelami dwuklasowymi i trzyklasowymi z zmienną współwystępującą (SOS)

1 klasa	AIC	47920,657
	BIC	48206,239
	Skoryg. BIC	48037,867
2 klasy z SOS	AIC	47654,611
	BIC	**48231,162**
	Skoryg,BIC	**47891,243**
3 klasy z SOS	AIC	**47598,914**
	BIC	48466,435
	Skoryg.BIC	47954,967

Uwaga: pogrubienie oznacza najniższą wartość wskaźnika, tym samym wskazuje preferowany model

Ze względu na dopasowanie modelu dwuklasowego możliwa jest analiza wpływu zmiennej niezależnej na przynależność do klas. Poniższa tabela zawiera liczebność klas oraz współczynniki regresji (i ilorazy szans) określające relację między wynikami SOS a przynależnością do danej klasy. Interpretacja współczynników regresji w kategoriach logitu jest mało intuicyjna, dlatego do raportowania wyników posłużymy się ilorazem szans, który wyznacza efekt wpływu zmiennej współwystępującej na szanse przynależności do klasy c w stosunku do klasy referencyjnej C, związany z jednostkowym przyrostem zmiennej współwystępującej (Collins i Stanza, 2010). Dla

zmiany wyniku w kwestionariuszu SOS o jedno odchylenie standardowe, szansa na przynależność do klasy bardziej zmotywowanej rośnie ponad dwukrotnie (2,39), jednocześnie szanse przylepności do klasy zdemontowanej wraz ze wzrostem SOS spadają o 0,418 razy (czyli o 58,2%).

Wykorzystanie zmiennej współwystępującej w mieszanym modelu IRT wskazuje, że wynik w kwestionariuszu SOS jest istotnym predyktorem dla przynależności do klas ukrytych.

Tabela 4. Regresja wyników SOS na przynależność do klas ukrytych

	Klasa 1 (zmotywowana)		Klasa 2 (zdemotywowana)		Klasa 1	Klasa 2
	Współczynnik regresji	Iloraz szans	Współczynnik regresji	Iloraz szans	Liczebność w klasach (%)	
Wynik SOS	0,872	2,390	-0.872	0,418	881 (54%)	736 (46%)

* wszystkie współczynniki regresji są istotne statystycznie

By porównać jakość klasyfikacji modelu ze zmienną współwystępującą oraz podstawowego modelu bez zmiennych współwystępujących, w poniższej tabeli (tabela 5) prezentujemy wartości współczynnika entropii dla obydwu tych modeli.

Tabela 5. Wskaźnik entropii i odsetek uczniów w bardziej zmotywowanej klasie w porównywanych modelach

Model	Entropia	% uczniów w bardziej zmotywowanej klasie
Bez zmiennych współwystępujących	0,493	58,44
Z wynikiem w SOS jako zmienną współwystępującą	0,526	54,00

Możemy zauważyć, że model ze zmienną współwystępującą cechuje się wyższym wskaźnikiem entropii, czyli klasy ukryte są od siebie lepiej odseparowane. Po wykorzystaniu wyniku w kwestionariuszu SOS, do klasy pierwszej zostaje zaklasyfikowane 54% uczniów, czyli o prawie 4,5 punktów procentowych mniej niż w przypadku modelu bez zmiennych współwystępujących. Można wiec wnioskować o tym, że wykorzystanie zmiennej współwystępującej polepsza jakość klasyfikacji uczniów do klas ukrytych, biorąc pod uwagę wskaźnik entropii, jednak sama klasyfikacja nie zmienia się znacznie.

8. Dyskusja wyników

Wydaje się, że mieszane modele IRT mogą mieć zastosowanie m.in. w próbnym testowaniu (tj. w warunkach obniżonej motywacji testowej) zadań, które mają zostać później wykorzystane w testach wysokiej stawki. Wyodrębnienie klasy ukrytej, reprezentującej wyższy poziom motywacji podczas testowania próbnego, może pozwolić na oszacowanie parametrów zadań bliższych parametrom z

prawdziwej sytuacji egzaminowania. Może to tym samym umożliwić rzetelniejszą walidację narzędzi i dać trafniejsze informacje, które mogą następnie zostać zdeponowane w banku zadań.

Wykorzystanie mieszanych modeli IRT do detekcji niezmotywowanych uczniów może także stanowić punkt wyjścia do dalszych analiz mających na celu zwiększenie trafności wniosków wyprowadzanych na podstawie rozwiązywanych przez uczniów testów umiejętności. Uzyskane wyniki można wykorzystać do walidacji rezultatów innych metod – określenie spójności ich wyników z wynikami miar samoopisowych, ale także nieuwzględnionych w tym opracowaniu miar zachowania uczniów (np. RTE i liczba zadań otwartych opuszczonych w rozwiązywanym przez ucznia teście). W dalszych badaniach można także zastanowić się, jakie konsekwencje dla wyników analiz ma filtrowanie danych uczniów niezmotywowanych, np. dla szacowanego poziomu umiejętności uczniów. Należy zwrócić uwagę, że do klasy uczniów mniej zmotywowanych należy około 45% uczniów. Konsekwencje usunięcia ich z analizy dla jej wyników powinny stać się przedmiotem kolejnych analiz.

Literatura

Asparouhov, T. i Muthén, B. (2012). Using Mplus TECH11 and TECH14 to test the number of latent classes. *Mplus Web Notes, 14*, 1-17.

Asparouhov, T. i Muthén, B. (2013). Auxiliary variables in mixture modeling: 3-step approaches using Mplus. *Mplus web notes, 15*, 1-24.

Clark, S. L. i Muthén, B. (2009). *Relating latent class analysis results to variables not included in the analysis*. Pobrano z https://www.statmodel.com/download/relatinglca.pdf.

Collins, L. M. i Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley i Sons.

Dolata, R.i Pokropek, A. (2010). 10). Motywacja a wynik testu z nauk przyrodniczych Studium na przykładzie PISA 2006. W B. Niemierko i M.K. Szmigel (red.). *Teraźniejszość i przyszłość oceniania szkolnego: XVI Krajowa Konferencja Diagnostyki Edukacyjnej, Toruń, 22 - 24 października 2010 r*. (s. 86-97).Kraków: Grupa Tomami.

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement, 66*(4), 643–656.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*, 311-326.

Eklöf, H. (2008). Test-taking motivation on low-stakes tests: A Swedish TIMSS example. W *Issues and Methodologies in Large-Scale Assessments*. (s.9-23). IERI Monograph Series, Vol. 1. Hamburg: IEA-ETS Research Institute.

Lo, Y., Mendell, N. R. i Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767-778.

Messick, S. (1989). Validity. W R. Linn (red.), *Educational measurement* (wydanie trzecie., s. 13–103). Washington, DC: American Council on Education.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Muthén, L.K. i Muthén, B.O. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén i Muthén

Nylund, K. L., Asparouhov, T., i Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling, 14*(4), 535-569.

Setzer, J.C., Wise, S.L., van den Heuvel, J.R. i Ling, G.(2013). An Investigation of Examinee Test-Taking Effort on a Large-Scale Assessment. *Applied Measurement in Education, 26*(1), 34-49.

Skórska, P., Świst, K. i Pokropek, A. (2014). Indywidualne i grupowe efekty motywacji testowej uczniów. W B. Niemierko i M.K. Szmigel (red.). *Diagnozy edukacyjne. Dorobek i nowe zadania. XX Krajowa Konferencja Diagnostyki Edukacyjnej. Gdańsk, 18 - 20 września 2014 r.*(s. 146-156). Kraków: Grupa Tomami.

Sundre, D.L. (1999). *Does examinee motivation moderate the relationship between test consequences and test performance*? (Report No. TM029964). Harrisonburg, Va.: James Madison University. (ERIC Documentation Reproduction Service No. ED432588.)

Sundre, D.L. (2007). The Student Opinion Scale (SOS): *A measure of examinee motivation. Test Manual*. Pobrano z http://www.jmu.edu/assessment/resources/Overview.htm

Sundre D. L. i Wise S.L. (2003). *'Motivation filtering' : An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the annual meeting of National Council on Measurement in Education, Chicago.

Sundre, D.L. i Kitsantas, A.L. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*(1), 6-26.

Swerdzewski, P.J., Harmes, J.C. i Finney, S.J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*, 162-188.

Swerdzewski, P., Finney, S.J. i Harmes, J.C. (2007). *Skipping the Test: Using Evidence to Inform Policy Related to Those Students Who Avoid Taking Low-Stakes Assessments in College.* Paper presented at the annual meeting of the Northeastern Education Research Association, Rocky Hill, CT.

Thelk, AD., Sundre, D.L., Horst, S.J. i Finney, S.J. (2009). Motivation matters: Using the Student Opinion Scale (SOS) to make valid inferences about student performance. *Journal of General Education, 58,* 129-166.

Węziak-Białowolska, D. (2011). Skala motywacji testowej – analiza właściwości psychometrycznych. W B. Niemierko i M.K. Szmigel (red.). *XVII Konferencja Diagnostyki Edukacyjnej. Ewaluacja w edukacji: koncepcje, metody, perspektywy. Kraków 23-25 września 2011* (s. 451-462). Kraków: Grupa Tomami.

Wise, S.L. i DeMars, C.E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

Wise, S.L. i Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 162- 183.

Wolf, L.F. i Smith, J.K. (1993). *The effects of motivation and anxiety on test performance*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Wolf, L.F. i Smith, J.K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242.