IBE
INSTYTUT
BADAŃ
EDUKACYJNYCH

Marek Muszyński
Bartosz Kondratek
Agata Gajewska-Dyszkiewicz
Katarzyna Paczuska
Magdalena Szpotowicz

Właściwości psychometryczne egzaminu maturalnego 2015 - język angielski, poziom podstawowy.

Analizy IBE/02/2016

Warszawa 2016

Autorzy:
Marek Muszyński
Bartosz Kondratek
Agata Gajewska-Dyszkiewicz
Katarzyna Paczuska
Magdalena Szpotowicz

Recenzja:
dr Dorota Campfield

Wzór cytowania:
Muszyński, M., Kondratek, B., Gajewska-Dyszkiewicz, A., Paczuska, K., Szpotowicz, M., (2016). *Właściwości psychometryczne egzaminu maturalnego 2015 - język angielski, poziom podstawowy.* Analizy IBE, 2. Warszawa: Instytut Badań Edukacyjnych.

Wydawca:
Instytut Badań Edukacyjnych
ul. Górczewska 8
01-180 Warszawa
tel. (22) 241 71 00; www.ibe.edu.pl

Egzemplarz bezpłatny

Spis treści

Streszczenie

Celem niniejszego artykułu jest przebadanie własności psychometrycznych egzaminu maturalnego z języka angielskiego na poziomie podstawowym z 2015 roku, ocenienie przydatności danych egzaminacyjnych z Centralnej Komisji Egzaminacyjnej do badań naukowych oraz wskazanie możliwych przyczyn ewentualnych problemów poszczególnych pozycji testowych. Przeprowadzono, tak jak było to możliwe przy użyciu zastanych danych, badanie rzetelności testu, jego wymiarowości, powiązania pozycji testowych z wynikiem ogólnym, wielkości względnego, standardowego błędu pomiaru oraz zróżnicowanego funkcjonowania pozycji testowych ze względu na cechy respondenta (płeć, dysleksja, lokalizacja szkoły, typ szkoły). Wyniki analiz wskazują, że jakość zebranych danych jest wysoka, a własności psychometryczne testu są dobre. Test jest rzetelny, a jego wymiarowość zadowalająca. Nie wykryto również zbyt wielu pozycji testowych, miałyby różne charakterystyki, w różnych grupach zdających. Zidentyfikowano problematyczne pozycje testowe i skomentowano możliwe przyczyny problemów. Sporym problemem analizowanego testu wydaje się duże natężenie zgadywania poprawnych odpowiedzi przez uczniów. Wydaje się, że odpowiadają za to formaty poszczególnych zadań, szczególnie format prawda/fałsz.

Wprowadzenie

Centralna Komisja Egzaminacyjna publikuje sprawozdania z egzaminów, które przeprowadza. Dokumenty te zawierają informacje, dotyczące np. liczby uczniów, którzy podeszli do egzaminu oraz statystyki opisowe testów i zadań. Dane dotyczące konstruowania i pilotowania zadań, a także własności psychometrycznych testów egzaminacyjnych nie są publikowane. Informacje te są potencjalnie istotne dla środowisk edukacyjnych, w tym nauczycieli, badaczy edukacyjnych i decydentów, którzy na podstawie wyników egzaminacyjnych i informacji o wykonaniu poszczególnych zadań testowych wnioskują o poziomie umiejętności oraz słabych i mocnych stronach umiejętności uczniów. Dostarczenie powyższej wiedzy jest głównym celem artykułu. Analizą objęto egzamin podstawowy z języka angielskiego, ze względu na jego popularność wśród uczniów.

Jakikolwiek test, by był czymś więcej niż tylko „zbiorem pytań lub zadań" musi spełniać określone warunki: być obiektywny, wystandaryzowany, trafny, rzetelny i znormalizowany (Hornowska, 2007). Międzynarodowe normy stawiane testom mówią jeszcze o bezstronności narzędzi, a także stawiają wiele wymagań dotyczących m.in. procedury konstruowania pytań, testowania, oceniania i dokumentowania wyników, jak również praw i obowiązków osób zaangażowanych w cały proces testowania (AERA, APA i NCME, 2001). Szczególnie surowe wymagania powinny być stawiane testom, których konsekwencje mogą być dla testowanych długotrwałe i doniosłe. Niewątpliwie takim testem „wysokiej stawki" (*high stake*) jest w Polsce egzamin maturalny.

W artykule przedstawiamy wynik analizy i omówienie najważniejszych własności psychometrycznych testu maturalnego z języka angielskiego. Sprawdzone zostaną między innymi:

- rzetelność pozycji testowych i ich dopasowanie do mierzonego konstruktu (w tym wypadku poziomu umiejętności w języku angielskim),
- możliwy wpływ czynników behawioralnych (np. zgadywania) na wyniki uczniów,
- ewentualne zróżnicowane funkcjonowanie pozycji testowych ze względu na dane cechy uczniów (płeć, orzeczenie o dysleksji, lokalizację szkoły (miasto-wieś), rodzaj szkoły (publiczna-niepubliczna)),
- związane z nimi czynniki niepowiązane z mierzonym konstruktem
- wielkość względnego, standardowego błędu pomiaru.

Analizie poddana zostanie również wewnętrzna (latentna) struktura testu. W takim zakresie jak to było możliwe na podstawie posiadanych danych, ocenie poddane zostaną właściwości egzaminu maturalnego, do których odnoszą się międzynarodowe Standardy: 1) trafność, 2) rzetelność, 3) skale i normy oraz 7) bezstronność (AERA i in., 2001).

Wybrany został egzamin z języka angielskiego na poziomie podstawowym ze względu na jego popularność wśród zdających -ponad 90% wybiera angielski jako obowiązkowy język obcy. Drugim powodem był fakt wprowadzenia nowego arkusza egzaminacyjnego. Egzamin maturalny z 2015 roku był pierwszym, do którego przystępowali uczniowie realizujący program według „nowej" podstawy programowej z 2008 roku (MEN, 2008). W przypadku egzaminu z języków

obcych wiązało się z to ze zmianami arkuszy maturalnych (Smolik, 2012). Analizie poddano całą część pisemną egzaminu, czyli zadania sprawdzające sprawności rozumienia ze słuchu, rozumienia tekstu czytanego, znajomości struktur gramatycznych i słownictwa oraz tworzenia wypowiedzi pisemnej.

W artykule przyjęto nazywać pojedynczą pozycję testową pytaniem, pozycją testową lub itemem (kalka z angielskiego), natomiast grupę (wiązkę) pozycji testowych określa się jako zadanie. W literaturze anglojęzycznej stosuje się najczęściej pojęcia odpowiednio: *item* oraz *testlet*.

1. Rzetelność pozycji testowych i ich dopasowanie do mierzonego konstruktu

Analizę rozpoczęto od policzenia rzetelności całego testu, jak i poszczególnych jego zadań metodą alfy Cronbacha (Cronbach, 1951). Zgodność poszczególnych pozycji testowych z mierzonym przez cały test konstruktem policzono za pomocą korelacji pozycji testowej z ogólnym wynikiem całego testu (*item-test correlation, Rit*) oraz korelacji pozycji testowej z ogólnym wynikiem całego testu pomniejszonym o tę pozycję (*item-rest correlation*, Rir; Nunnally i Bernstein, 1994).

Wyniki dla całego testu i poszczególnych zadań wskazują, że matura z języka angielskiego z roku 2015 na poziomie podstawowym jest testem o bardzo wysokiej rzetelności, którego znaczna większość zadań mierzy ten sam lub zbliżony konstrukt (Tabela 1).

Tabela 1. Rzetelność i korelacje zadań testu.

Zadanie	Średnia	Max	Łatwość	Rit	Rir	Alfa testu bez zadania
1	3,75	5,00	0,75	0,63	0,55	0,90
2	3,21	4,00	0,80	0,76	0,71	0,89
3	4,13	6,00	0,69	0,83	0,77	0,89
4	3,35	4,00	0,84	0,80	0,76	0,89
5	2,34	3,00	0,78	0,76	0,72	0,90
6	4,17	5,00	0,83	0,84	0,80	0,89
7	2,33	3,00	0,78	0,79	0,75	0,89
8	3,04	5,00	0,61	0,66	0,59	0,90
9	4,16	5,00	0,83	0,79	0,75	0,89
10	7,86	10,00	0,79	0,86	0,75	0,91
Test	38,35	50	0,77	-	-	0,90

Wyniki przedstawione w Tabeli 1 pokazują także, że Zadanie 1 i Zadanie 8 zdecydowanie najniżej korelują z wynikiem całego testu (kolumny „Rit" i „Rir"). Może to oznaczać, że zadania te

mierzą inny konstrukt niż pozostałe lub też, że na ich wynik mają wpływ inne, zewnętrzne wobec testu zmienne, które nie mają takiego wpływu na pozostałe zadania.

W celu przyjrzenia się bliżej temu zjawisku przestudiowano analogiczne wskaźniki dla całego egzaminu w rozbiciu na pojedyncze pozycje testowe, a nie zsumowane zadania (Tabela 2). Wyższa rzetelność testu obliczona w przypadku analizy pojedynczych pozycji testowych w porównaniu do zsumowanych zadań, to nie tylko logiczna konsekwencja zwiększenia się „liczby" pozycji testowych (z 10 do 44), ale też wzięcie pod uwagę wysokich wzajemnych korelacji (interkorelacji) pozycji testowych w obrębie danego zadania. W tym przypadku rzetelność całego testu wynosi 0,9437 i zbliża się do ideału, tj. 0,95, wymaganego od testów, które mają nieść informacje o poziomie umiejętności indywidualnych uczniów (Nunnally i Bernstein, 1994).

Tabela 2. Rzetelność i korelacje pozycji testowych.

Pozycja testowa	Średnia	Max	Łatwość	Rit	Rir	Alfa testu bez itemu
1_1	0,72	1,00	0,72	0,44	0,40	0,94
1_2	0,77	1,00	0,77	0,34	0,30	0,94
1_3	0,79	1,00	0,79	0,39	0,36	0,94
1_4	0,81	1,00	0,81	0,30	0,27	0,94
1_5	0,66	1,00	0,66	0,39	0,35	0,94
2_1	0,78	1,00	0,78	0,67	0,65	0,94
2_2	0,80	1,00	0,80	0,61	0,58	0,94
2_3	0,74	1,00	0,74	0,63	0,61	0,94
2_4	0,89	1,00	0,89	0,41	0,39	0,94
3_1	0,58	1,00	0,58	0,52	0,49	0,94
3_2	0,65	1,00	0,65	0,58	0,55	0,94
3_3	0,62	1,00	0,62	0,48	0,44	0,94
3_4	0,86	1,00	0,86	0,52	0,50	0,94
3_5	0,72	1,00	0,72	0,65	0,63	0,94
3_6	0,71	1,00	0,71	0,64	0,61	0,94
4_1	0,79	1,00	0,79	0,67	0,65	0,94
4_2	0,88	1,00	0,88	0,60	0,58	0,94
4_3	0,81	1,00	0,81	0,63	0,60	0,94
4_4	0,87	1,00	0,87	0,59	0,57	0,94
5_1	0,89	1,00	0,89	0,64	0,62	0,94
5_2	0,85	1,00	0,85	0,50	0,47	0,94
5_3	0,60	1,00	0,60	0,57	0,54	0,94
6_1	0,92	1,00	0,92	0,54	0,52	0,94

6_2	0,72	1,00	0,72	0,65	0,63	0,94
6_3	0,77	1,00	0,77	0,62	0,59	0,94
6_4	0,91	1,00	0,91	0,47	0,45	0,94
6_5	0,85	1,00	0,85	0,67	0,65	0,94
7_1	0,78	1,00	0,78	0,63	0,60	0,94
7_2	0,71	1,00	0,71	0,68	0,66	0,94
7_3	0,85	1,00	0,85	0,60	0,58	0,94
8_1	0,78	1,00	0,78	0,44	0,41	0,94
8_2	0,65	1,00	0,65	0,41	0,37	0,94
8_3	0,59	1,00	0,59	0,42	0,38	0,94
8_4	0,81	1,00	0,81	0,56	0,54	0,94
8_5	0,22	1,00	0,22	0,09	0,05	0,95
9_1	0,76	1,00	0,76	0,50	0,47	0,94
9_2	0,70	1,00	0,70	0,58	0,55	0,94
9_3	0,87	1,00	0,87	0,47	0,44	0,94
9_4	0,92	1,00	0,92	0,48	0,46	0,94
9_5	0,91	1,00	0,91	0,54	0,52	0,94
10_t	3,33	4,00	0,83	0,77	0,73	0,94
10_s	1,65	2,00	0,82	0,76	0,73	0,94
10_z	1,52	2,00	0,76	0,78	0,76	0,94
10_p	1,37	2,00	0,68	0,78	0,75	0,94
Test	38,35	50	0,77	-	-	0,94

Niskie umiejętności rozumienia tekstu i rozumowania matematycznego często współwystępują – niemal 15% osób to takie, które w badaniu osiągnęły co najwyżej 1. poziom zarówno w obszarze rozumienia tekstu, jak i rozumowania matematycznego.

Analiza korelacji poszczególnych pozycji testowych z wynikiem w całym teście pozwala ujawnić skąd wynikały wyraźnie niższe związki Zadania 1 i Zadania 8 z całym testem. W Zadaniu 1 jego poszczególne pozycje testowe nisko lub bardzo nisko korelują z wynikiem matury (po wykluczeniu analizowanej pozycji testowej), dwie z nich osiągają lub nawet są poniżej progu 0,3, który jest zalecanym najniższym poziomem wskaźnika Rit/Rir (Jakubowski i Pokropek, 2009). Kline (2005) uważa nawet, że powinno się unikać pozycji testowych z wartością Rit/Rir niższą niż 0,5. Pozostałe pozycje Zadania 1 oraz pytania 2_4, 8_2 i 8_3 również wykazują dość niskie związki z wynikiem całego testu. Najpoważniejszym problemem wydaje się jednak niezwykle niska kompatybilność pozycji 8_5 z innymi. To pytanie wydaje się mierzyć coś zdecydowanie innego (inny konstrukt), niż reszta testu.

Ocena na podstawie współczynników Rit i Rir dotyczy konkretnej grupy uczniów rozwiązujących test i konkretnego testu, w związku z czym jest zawsze względna. Ewentualna nieodpowiedniość pozycji testowych może być wnioskowana jedynie w odniesieniu do tych dwóch kryteriów. Pozycje testowe wykazujące niską korelację z wynikiem całego testu mogą wykazywać wysoką korelację, jeśli będą zastosowane dla innej grupy osób i/lub jeśli zostaną użyte

w teście złożonym z innych pozycji testowych. Jednak w wypadku analizowanego testu, wobec wszystkich pytań z Zadania 1 (1_2 i 1_4 w szczególności), pytania 2_4, pytań z Zadania 8 (oprócz 8_4) można przypuszczać, że w trakcie ich rozwiązywania przez uczniów miały znaczenie inne zewnętrzne czynniki związane z charakterystyką uczniów lub samych zadań. W przypadku pozycji testowej 8_5 można mieć pewność, że zadanie to nie pasuje do tego konkretnego testu i tych konkretnych uczniów.

Wyniki zaprezentowane w Tabeli 2 pozwalają zauważyć, że test składa się głównie z łatwych pozycji testowych. Kilka z nich wykracza poza przyjętą granicę, zgodnie z którą żadne pytanie nie powinno przekraczać 85-90% poprawnych odpowiedzi. W analizowanym egzaminie maturalnym aż 10 pozycji testowych ma łatwość większą niż 0,85, a kilka kolejnych zbliża się do tej granicy. Może to mieć związek z systematycznie rosnącymi umiejętnościami uczniów w zakresie języka angielskiego (Szaleniec i in., 2015), które być może już teraz przekraczają przewidziany przez podstawę programową poziom na tym etapie edukacyjnym. Kilka pozycji testowych wydaje się być zbyt łatwa w stosunku do poziomu uczniów, jednak trzeba pamiętać, że trudność testu musi być zgodna z obowiązującymi wymaganiami nakładanymi przez podstawę programową. Powodem niskiej trudności kilku pytań może być również to, że do matury 2015 w nowej formule podeszli w tym roku tylko uczniowie liceum (CKE, 2015). Należałoby więc przeprowadzić analogiczne analizy w roku przyszłym oraz dalej monitorować wzrost poziomu umiejętności uczniów za pomocą porównywalnych wyników egzaminów, by móc zdecydować o ewentualnym zwiększeniu wymagań na egzaminie podstawowym z języka angielskiego.

Jedna pozycja testowa ma z kolei łatwość poniżej zalecanego minimum 25-30% poprawnych odpowiedzi, co oznacza, że item ten może zawierać błędy, które uniemożliwiają jego wykonanie lub jest zbyt trudny dla danej populacji uczniów, np. wykracza poza materiał dla matury podstawowej (sugerowane przedziały dla współczynnika łatwości: np. Matlock-Hetzel, 1997; Schuwirth i Pearce, 2014). Wydaje się, że problemy z pozycją 8_5 są spowodowane właśnie jego wykroczeniem poza wymagania egzaminacyjne.

Zadania/pozycje testowe o niskiej korelacji z innymi obniżają rzetelność testu i mogą być również zagrożeniem dla źródeł jego trafności. Podobne efekty mogą dać zadania/pozycje testowe zbyt łatwe lub zbyt trudne dla danej populacji uczniów. Niska korelacja z innymi itemami sugerować może np. nieodpowiedniość pozycji testowej dla danej grupy uczniów (wynikająca z niezgodności z programem nauczania), niezgodność z badanym konstruktem (pozycja mierzy inne umiejętności, niż reszta testu) lub odpowiadanie przez uczniów w inny sposób niż na inne pytania (np. losowy wybór odpowiedzi; *random-guessing*). Wszystkie te czynniki stanowią zagrożenie dla trafności testu (Sireci, 2009), a w rezultacie mogą nawet uniemożliwić zastosowanie testu zgodnie z jego zakładanym przeznaczeniem. W tym wypadku jest nim sprawdzenie umiejętności w zakresie języka angielskiego w odniesieniu do zapisów podstawy programowej IV.1, a dla kontynuujących naukę, w zakresie podstawowym. W przypadku matury 2015 tylko kilka pozycji testowych wykazuje pewne możliwe problemy, więc zagrożenie dla trafności i rzetelności testu wydaje się nieznaczne. Bardziej szczegółowa analiza wyróżnionych pytań, przeprowadzona w dalszej części artykułu, pozwoli zrozumieć, jakie są możliwe przyczyny tego zagrożenia.

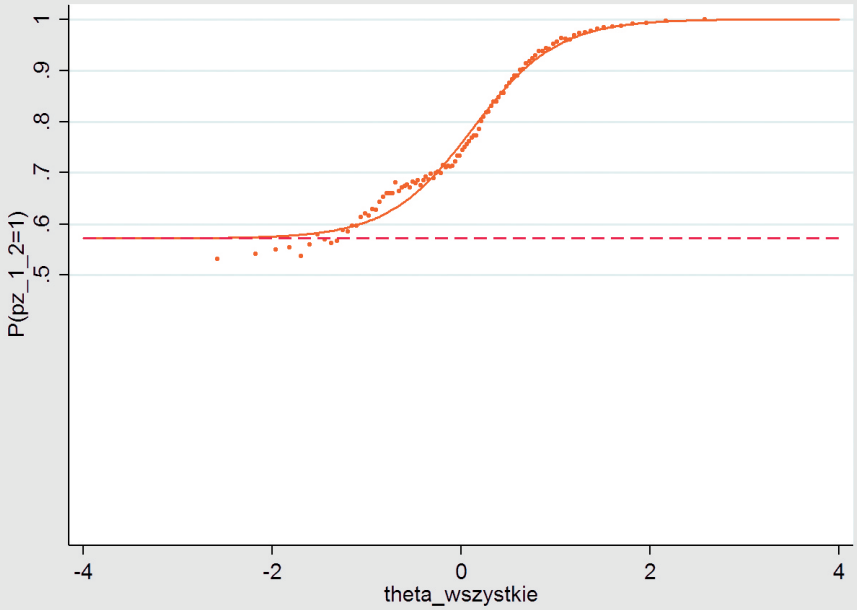
2. Analiza zadań wykazujących niskie korelacje z wynikiem w teście

Analizę tę rozpoczęto od przyjrzenia się tzw. krzywym charakterystycznym pozycji testowej (item characteristic curve; ICC). Uzyskano je poprzez dopasowanie trójparametrycznego modelu logistycznego (3PLM) z rodziny modeli teorii odpowiedzi na pozycje testowe (item response theory, IRT; patrz np.: Hambleton, Swaminathan i Rogers, 1991). Krzywe te informują o własnościach psychometrycznych poszczególnych pytań w sposób graficzny. Analizując je można łatwo dokonać interpretacji właściwości poszczególnych itemów i wychwycić pytania problematyczne. Krzywe niosą informacje o szacowanej trudności zadań (im bardziej środek krzywej przesunięty w lewo, tym pytanie łatwiejsze), ich mocy dyskryminacyjnej (im krzywa bardziej stroma tym pytanie lepiej odróżnia uczniów o różnym poziomie umiejętności) oraz wartości parametru pseudozgadywania (pseudoguessing), który jest parametrem korygującym możliwość zgadywania przez uczniów poprawnych odpowiedzi (Jakubowski i Pokropek, 2009), a którego interpretacja nastręcza dość poważnych trudności (Han, 2012).

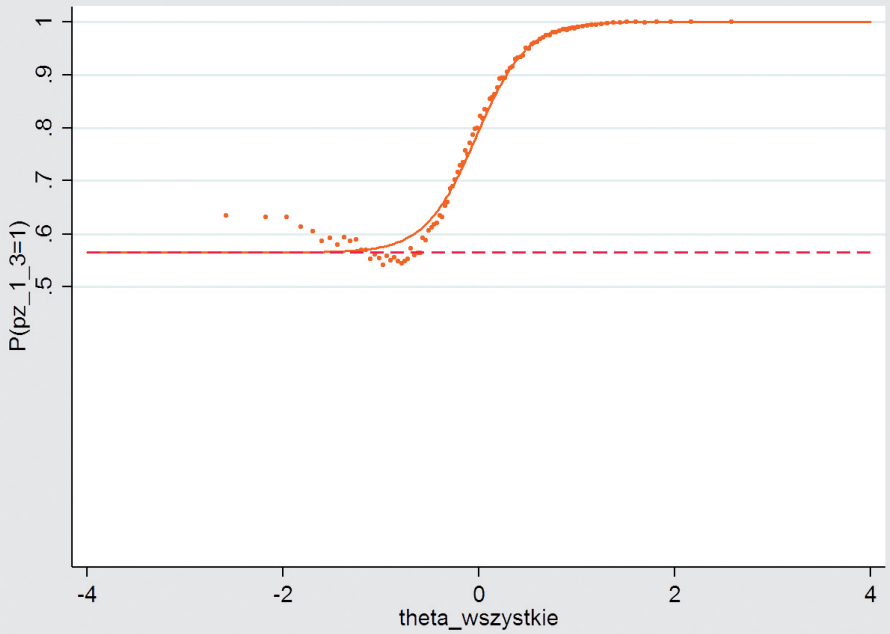
Krzywe niosą informacje o szacowanej trudności zadań, ich mocy dyskryminacyjnej oraz wartości parametru pseudozgadywania (pseudoguessing), który jest parametrem korygującym możliwość zgadywania przez uczniów poprawnych odpowiedzi (Jakubowski i Pokropek, 2009), a którego interpretacja nastręcza dość poważnych trudności (Han, 2012). Im bardziej środek krzywej przesunięty jest w lewo, tym pytanie jest łatwiejsze. Bardziej stroma krzywa oznacza, że dane pytanie lepiej odróżnia uczniów o różnym poziomie umiejętności/lepiej różnicuje pomiędzy uczniami o różnym poziomie umiejętności.

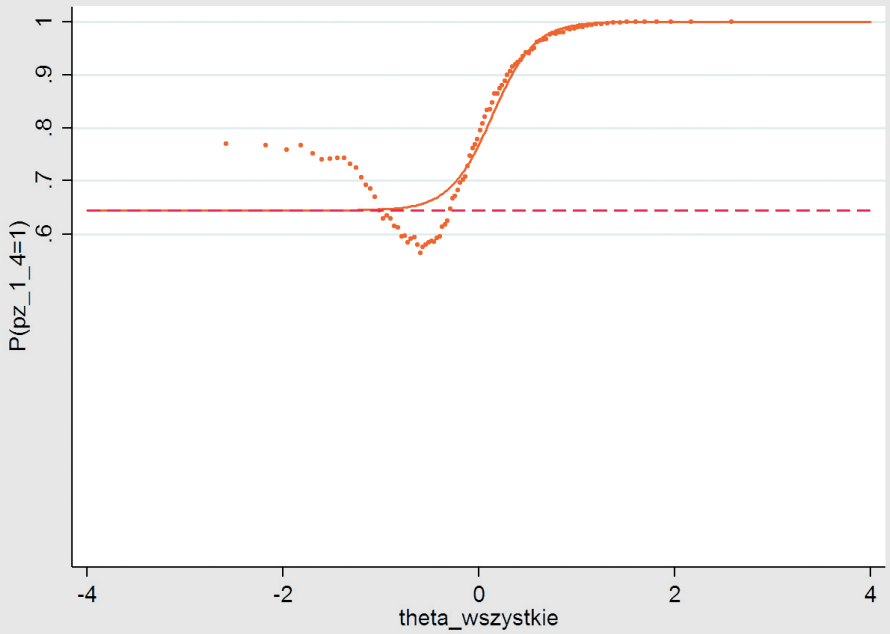
Niemniej jednak, im wyżej uniesiony jest lewy koniec krzywej, tym wyższa wartość parametru pseudozgadywania, co może oznaczać, że uczniowie zgadują lub dedukują odpowiedzi (Rysunek. 1).

Rysunek 1. Krzywe charakterystyczne wybranych pytań z Zadania 1[1] .



[1] Na osi pionowej rysunków oznaczone jest szacowane prawdopodobieństwo poprawnego rozwiązania danego zadania przez ucznia o określonym poziomie umiejętności. Szacowany poziom umiejętności ucznia jest oznaczony na osi poziomej, na standaryzowanej skali o średniej 0 i odchyleniu standardowym 1.


P(pz_1_3=1)
.5
.6
.7
.8
.9
1
-4
-2
0
2
4
theta_wszystkie


P(pz_1_4=1)
.6
.7
.8
.9
1
-4
-2
0
2
4
theta_wszystkie

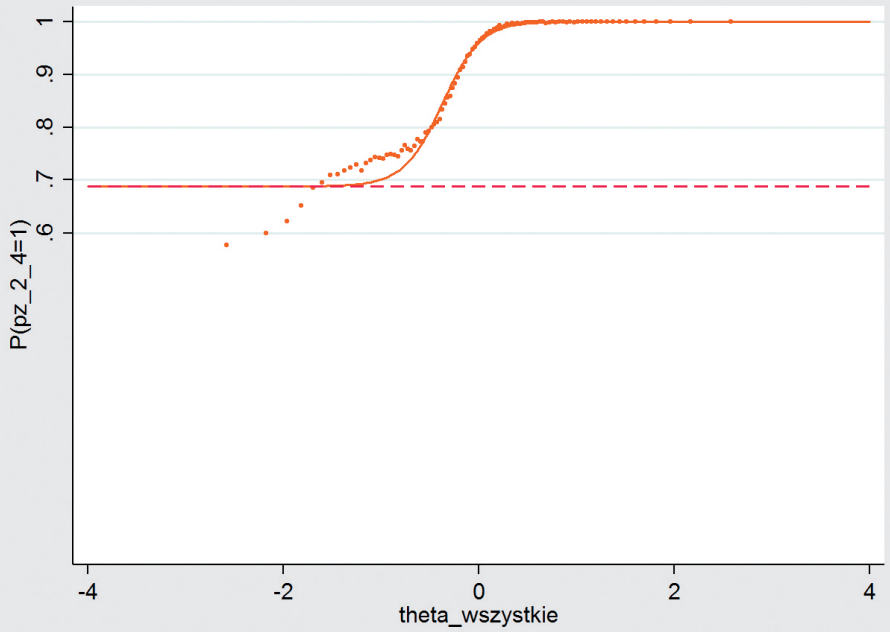
Przyjrzenie się trzem spośród pięciu pozycji testowych Zadania 1 pozwala na zauważenie pewnego istotnego problemu - uczniowie o niskim szacowanym poziomie umiejętności (lewa strona wykresu) mają wysokie szacowane prawdopodobieństwo udzielenia poprawnej odpowiedzi na te pytania. Jeśli założyć, że uczniowie o niższych umiejętnościach zgadywali losowo (*random guessing), to prawdopodobieństwo udzielenia przez nich poprawnej odpowiedzi powinno wynosić w tym wypadku 0,5 (50%), gdyż zadanie miało dwie opcje do wyboru. W prezentowanych powyżej przykładach widać, że to prawdopodobieństwo jest wyższe niż 0,5, w przypadku itemu 1_4* jest znacząco wyższe.. Taki wygląd krzywych charakterystycznych może sugerować, że uczniowie byli w stanie odgadnąć poprawną odpowiedź na podstawie danych charakterystyk pytania lub nawet całego zadania. Analiza treści poszczególnych pytań nie wskazuje niczego, co mogłoby uczniom pomóc w odgadnięciu poprawnej odpowiedzi. Być może więc uczniowie kierowali się odpowiedziami na inne pytania. Badania wskazują, że uczniowie kierują się niekiedy różnorodnymi strategiami, np. unikają dawania kilku odpowiedzi jednego typu (np. „A") pod rząd (Koniewski, Majkut i Skórska, 2014) lub starają się „odgadnąć" wzór, w jaki układają się odpowiedzi. W przypadku Zadania 1 być może uczniowie wykorzystali tę strategię, gdyż wzór poprawnych odpowiedzi na to pozwalał, układając się w naprzemienną sekwencję (dla arkusza A-fałsz, prawda, fałsz, prawda, fałsz).

Dodatkowo, warto zauważyć, że dla pozycji testowych 1_3 i 1_4 uczniowie najsłabsi mają wyższe szacowane prawdopodobieństwo udzielenia poprawnej odpowiedzi, niż uczniowie o średnim poziomie umiejętności (krzywa obrazująca rzeczywiste obserwacje układa się w „U"). Taki wygląd krzywej jest niepożądany, znamionuje on problemy z niską rzetelnością pytania i jest zagrożeniem dla trafności testu (Jakubowski i Pokropek, 2009). Sytuację taką może powodować wiele czynników: niedopasowanie pytania do poziomu uczniów, jego błędna konstrukcja, czy zawartość wskazówek, które pozwalają uczniom odgadnąć poprawną odpowiedź.

W przypadku omawianych pytań z Zadania 1 nic nie wskazuje, by zawierały one błędy merytoryczne lub metodologiczne. Być może są one dla uczniów trudne, gdyż udzielenie poprawnej odpowiedzi na nie wymaga połączenia kilku informacji, (np. pytanie 1_3: *„Cindy studied for 10 hours at the weekend just before the competition"- wymaga zrozumienia i zapamiętania liczby godzin nauki, pory tygodnia kiedy to zdarzenie miało miejsce oraz dodatkowego umieszczenia go w czasie, a więc aż trzech informacji). Uczniowie mogli więc* uciec się do zgadywania, gdyż pytanie okazało się dla nich zbyt trudne, lub gdyż jego format zachęcał do tego. Pytania typu prawda/fałsz od dawna są przedmiotem krytyki psychometryków (Storey, 1966), którzy wskazywali na ich niższą rzetelność i moc różnicującą niż pytania z większą liczbą opcji odpowiedzi (dystraktorów) do wyboru (Ebel, 1971; Frisbie, 1973; Hancock, Thiede, Sax i Michael, 1993). Co prawda cześć badaczy próbowała wykazać, że w pewnych określonych warunkach pytania typu prawda/fałsz działają równie dobrze jak pytania wielokrotnego wyboru (Ebel, 1972; Dudley, 2006), jednak nagromadzona przez lata wiedza wskazuje, że tego rodzaju pytania są trudne do poprawnego skonstruowania, są często mylące dla uczniów, charakteryzują się również gorszymi własnościami psychometrycznymi oraz dość łatwo jest na nie uczniom udzielić poprawnej odpowiedzi poprzez zgadywanie (Haladyna i Downing, 1989; Rodriguez, 2005). Ze względu na to, być może warto pomyśleć o odejściu od stosowania tego formatu w polskim systemie egzaminów zewnętrznych lub też zmianie sposobu punktowania tego typu zadań.

Kolejne pytanie wykazujące niską korelację z innymi (a więc zmniejszające rzetelność i, prawdopodobnie, trafność testu) to pozycja testowa 2_4.

Rysunek 2. Krzywa charakterystyczna pytania 2_4.



Podobnie jak w przypadków pytań z Zadania 1 krzywa charakterystyczna informuje , że szacowana wartość parametru pseudozgadywania jest bardzo wysoka. Analiza treści tego itemu wskazuje na występowanie w odpowiednim fragmencie nagrania wielu polsko-angielskich internacjonalizmów, które rozpoznają na pewno także uczniowie o niższym poziomie znajomości angielskiego (pogrubione w transkrypcji poniżej):

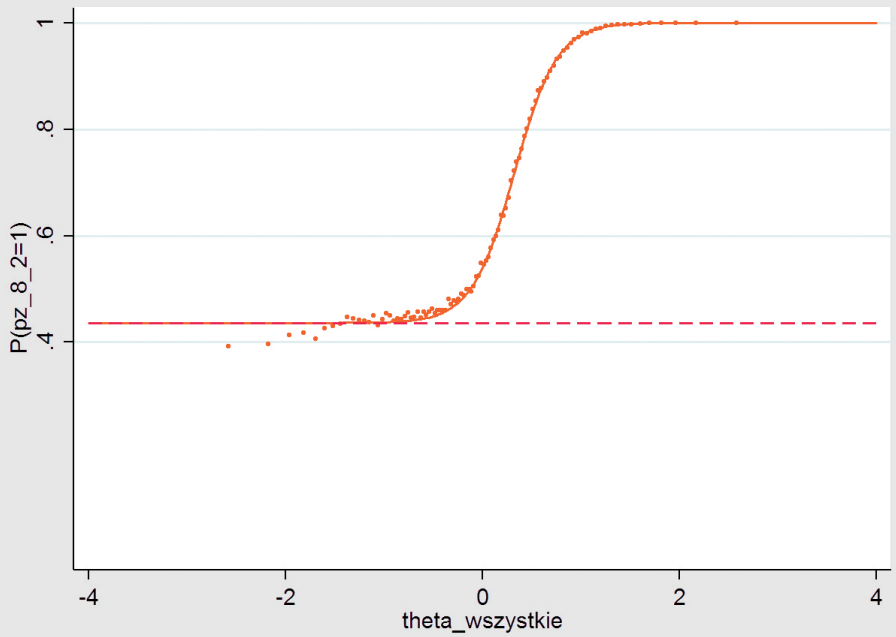
*„Can you imagine trying to **relax** with planes flying low over your head day and night? This happened to me when we were away. The **hotel** we stayed in was fine, but when booking the trip we didn't know we would have to stay so close to the airport. Believe it or not, we had to stop talking when a plane was landing or taking off. Even car alarms sometimes went off when a 747 **jumbo jet** went over. We went there to **relax** and there was no chance!"*
(Fragment zaadoptowany przez CKE z tripadvisor.com).

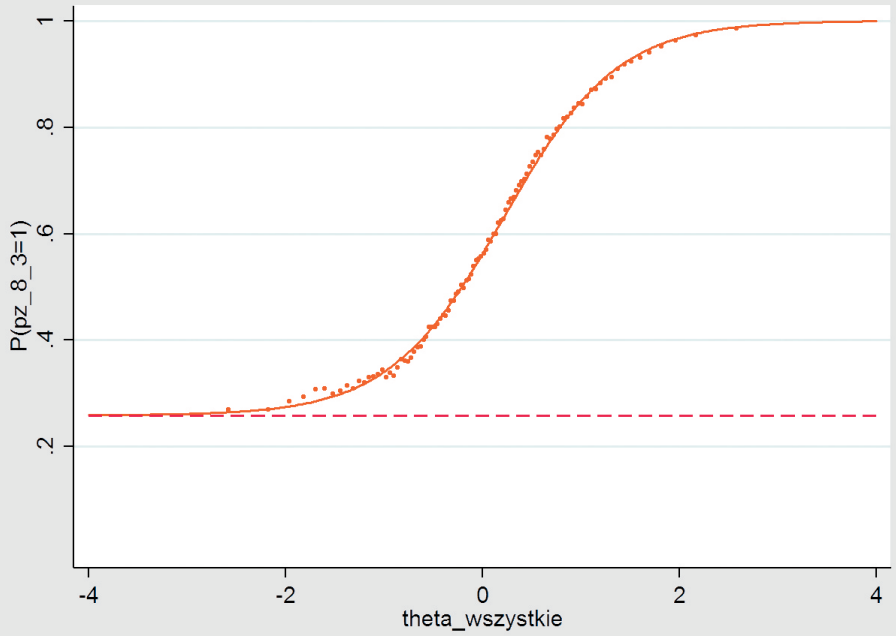
Wskazane we fragmencie powyżej słowa mogły aktywować kontekst wypoczynku i podróży. Tylko jeden dystraktor traktował o wakacjach, odpowiedź poprawna wydawała się więc oczywista. Być może uczniowie dedukowali odpowiedź na bazie tylko kilku zasłyszanych słów, które występują również w języku polskim i na bazie eliminacji nie pasujących opcji. Niewykluczone też, że na kształt krzywej miał wpływ ponownie format zadania. W tym zadaniu uczniowie mieli odpowiedzieć na cztery pytania, wybierając jedną z pięciu podanych opcji. Tak więc odpowiadając poprawnie na kolejne pytania, uczniowie zwiększali prawdopodobieństwo na kolejną poprawną odpowiedź, w miarę jak malał zbiór niewykorzystanych dystraktorów. Prawdopodobieństwo losowego zgadnięcia poprawnej odpowiedzi w pierwszym pytaniu wynosi 0,20, ale, w przypadku udzielenia poprawnej odpowiedzi, rośnie, gdyż kolejna odpowiedź wybierana jest już z mniejszej puli dystraktorów. Fakt, że pytanie 2_4 było ostatnie skłania do przypuszczania, że odpowiedź na nie mogła zostać po prostu wydedukowana na podstawie odrzucenia innych dystraktorów wcześniej. Być może więc, należałoby w przyszłości zadbać o większą proporcję opcji do wyboru do pytań w zadaniach tego formatu, tak jak jest np. w Zadaniu 4 z tego samego egzaminu, gdzie na 4 pytania przypada już 6 dystraktorów i w zgromadzonych danych brak

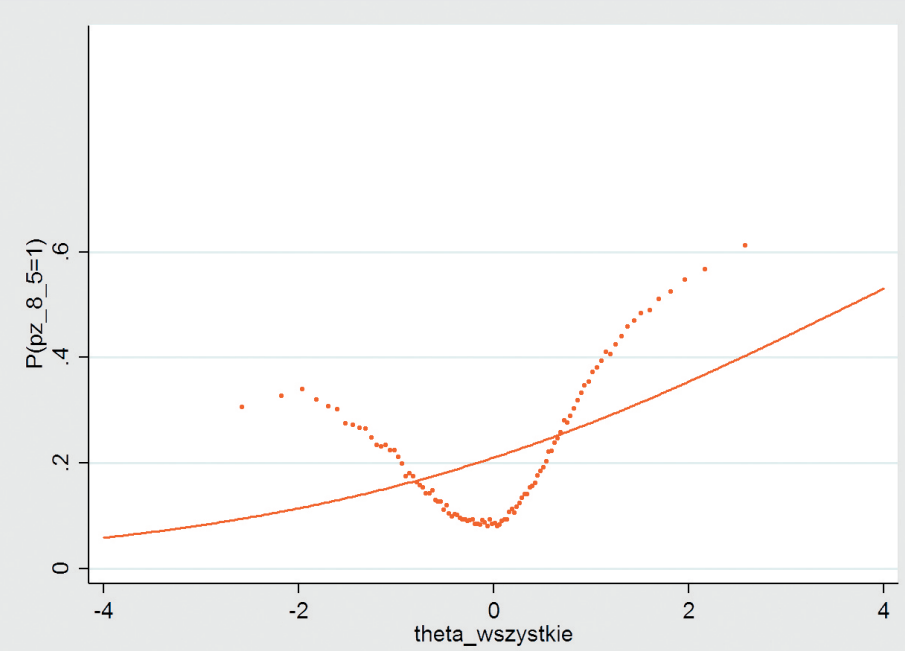
wskaźników sugerujących problemy z nadmiernie łatwym zgadywaniem odpowiedzi przez uczniów.

Kolejne problemy z psychometrycznymi właściwościami dotyczą pytań z Zadania 8- 8_2, 8_3 i zwłaszcza 8_5. Ponownie, spojrzenie na ich krzywe charakterystyczne może pomóc w odnalezieniu źródła problemów (Rysunek 3).

Rysunek 3. Krzywe charakterystyczne wybranych pytań z Zadania 8.







Krzywa dla pytania 8_2 wskazuje, że zgadywanie przez uczniów poprawnych odpowiedzi może mieć wpływ na obniżenie korelacji pytania z testem, jednak w przypadku pozycji testowej 8_3 takie problemy wydają się nie występować. Analiza treści obu pytań nie pozwala rozstrzygnąć, co może powodować obserwowane niskie korelacje. Krzywa dla pytania 8_5 wskazuje na bardzo poważne zaburzenie dyskryminacji przez to pytanie. Uczniowie słabsi mają tutaj większe prawdopodobieństwo poprawnego odpowiedzenia na to pytanie niż uczniowie o średnim poziomie umiejętności (pytanie ma lokalnie ujemną dyskryminację). Jest to oczywiście sytuacja wysoce niepożądana, która nie powinna mieć miejsca. Pytania takie należy wykrywać i poddawać drobiazgowej analizie ze względu na np. treść, możliwe błędy w kluczu, niezgodność z wymaganiami lub błędy w konstrukcji zadania, jako stanowiące poważne zagrożenie dla rzetelności i prawdopodobnie trafności pomiaru (Jakubowski i Pokropek, 2009).

Pozycja testowa 8_5 wydaje się po prostu niedopasowana do wymaganego na poziomie matury podstawowej stopnia znajomości języka. Analiza wybierania dystraktorów wskazuje, że uczniowie wybierali głównie odpowiedź B (*broke up), jako zapewne jedyny znany im spośród dystraktorów czasownik złożony. Znajomość poprawnej odpowiedzi (went off) wykracza poza sprawdzany na maturze podstawowej poziom B1 (za: English Profile, 2015; Rada Europy, 2001), a użycie zbyt trudnego dla uczniów pytania jest zagrożeniem dla trafności treściowej testu (Martone i Sireci, 2009). Więcej informacji na temat tego pytania można też przeczytać w podrozdziale 2.4 publikacji: Szpotowicz, Muszyński, Gajewska-Dyszkiewicz, Paczuska i Kondratek, 2016.*

Psychometryczne zastrzeżenia do pytań z matury 2015 dotyczą tylko kilku pozycji testowych, a zaledwie w jednym wypadku są to zastrzeżenia, które powinny doprowadzić do wymiany pytania w teście (8_5). Problemy pozycji testowych z Zadania 1 oraz pytania 2_4 są zapewne skutkiem użycia formatu zadania, które sprzyja pozamerytorycznym strategiom ich rozwiązywania. Natomiast pytanie 8_5 (i tylko ono) wykracza poza wymagania egzaminacyjne i nie powinno się prawdopodobnie znaleźć w tym teście.

3. Wewnętrzna struktura testu

Jednym ze źródeł trafności testu jest również wiedza o jego wewnętrznej strukturze (Sireci, 2009), to znaczy o relacjach pozycji testowych względem siebie oraz wobec postulowanej cechy ukrytej (latentnej), która leży u podstaw obserwowalnej wariancji pytań (patrz np. Konarski, 2004). W testach edukacyjnych, których zsumowany wynik wszystkich pytań ma informować o poziomie umiejętności uczniów oczekuje się jednowymiarowości testu (Gerbing i Anderson, 1988). Oznacza to, że u podłoża obserwowanych umiejętności uczniów będzie leżeć tylko jeden ukryty czynnik. W przypadku matury z języka angielskiego oczekuje się więc, że o obserwowanych wynikach uczniów będzie decydować tylko umiejętność posługiwania się językiem angielskim, a inne wpływy będą marginalne. Pewne pojęcie o strukturze testu daje Tabela 2. Zaleca się jednak przeprowadzenie analizy wymiarowości testu większą liczbą metod niż tylko obliczenie miar Rit i Rir (Gerbing i Anderson, 1988).

W tym przypadku zdecydowano o przeprowadzeniu analizy latentnej struktury testu metodą eksploracyjnej analizy czynnikowej. Jako punkt wyjścia analizy użyto obliczonej wcześniej macierzy korelacji polichorycznych, dających znacznie lepsze oszacowanie korelacji między zmiennymi zero-jedynkowymi i porządkowymi (Holgado–Tello, Chacón–Moscoso, Barbero–García, i Vila–Abad (2010).). Analiza wskazała, iż należy wyodrębnić dwa czynniki (na podstawie wartości własnej >1; Preacher i MacCallum, 2003), z których jeden jest czynnikiem zdecydowanie dominującym, odpowiadającym za ponad 89% wariancji wspólnej pozycji testowych (Tabela 3).

Tabela 3. Wartości własne i wariancja wyjaśniona czynników.[2]

Czynnik	Wartość własna	Wariancja wyjaśniona	Skumulowana wariancja wyjaśniona
1	22,54	0,89	0,89
2	1,89	0,07	0,97
3	0,85	0,03	1,00

Większość pozycji testowych ma zadowalająco wysokie ładunki czynnikowe na pierwszym czynniku, jedynie pozycje 1_4 i 8_5 mają je poniżej często przyjmowanego progu 0,4 (Tabachnick i Fidell, 2007). Jedynie kilka pozycji testowych posiada dość wysokie ładunki na czynniku 2 (głównie z Zadania 1 i pozycja 8_5). Wartości wariancji wyjątkowej są najwyższe dla pozycji testowych z Zadania 1 i Zadania 8 (Tabela 4).

Tabela 4. Ładunki czynnikowe i proporcja wariancji wspólnej.

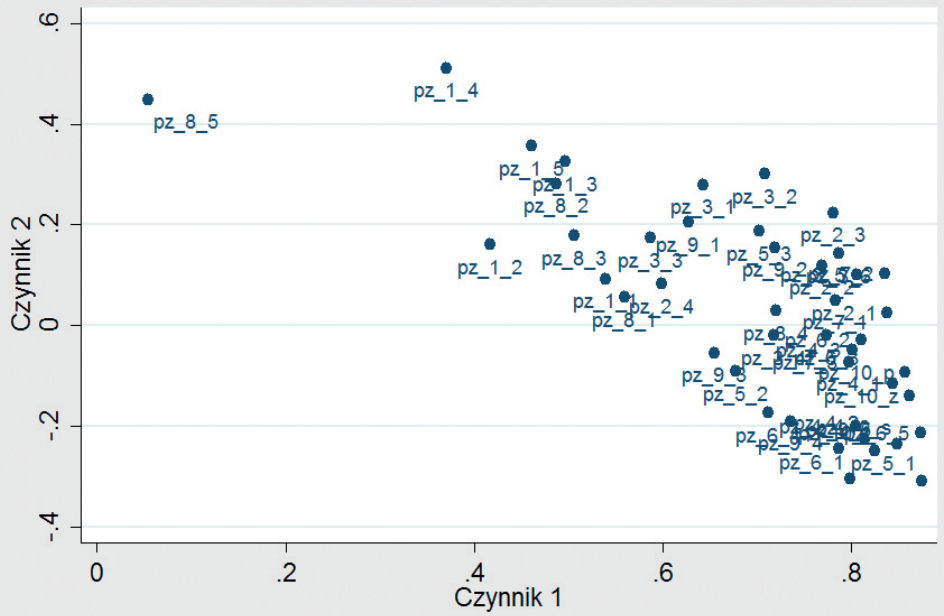
Zadanie	Czynnik1	Czynnik2	Czynnik3	Wariancja wspólna
1_1	0,54	0,09	0,01	0,30
1_2	0,42	0,16	0,00	0,20
1_3	0,50	0,33	0,01	0,35
1_4	0,37	0,51	0,14	0,42

[2] Rozwiązanie uzyskane za pomocą metody iteracyjnego czynnika głównego (iterated principal factor) z rotacją Varimax.

1_5	0,46	0,36	0,14	0,36
2_1	0,84	0,03	0,18	0,73
2_2	0,77	0,12	0,18	0,64
2_3	0,78	0,22	0,23	0,71
2_4	0,60	0,08	0,13	0,38
3_1	0,64	0,28	0,02	0,49
3_2	0,71	0,30	0,13	0,61
3_3	0,59	0,17	0,01	0,38
3_4	0,72	-0,02	0,16	0,54
3_5	0,81	0,10	0,09	0,67
3_6	0,79	0,14	0,10	0,65
4_1	0,84	-0,12	0,09	0,73
4_2	0,81	-0,23	0,06	0,72
4_3	0,80	-0,05	0,08	0,65
4_4	0,80	-0,20	0,09	0,69
5_1	0,87	-0,31	0,03	0,86
5_2	0,68	-0,09	0,01	0,47
5_3	0,70	0,19	0,10	0,54
6_1	0,80	-0,31	0,03	0,73
6_2	0,81	-0,03	0,08	0,66
6_3	0,77	-0,02	0,03	0,60
6_4	0,71	-0,17	0,04	0,54
6_5	0,87	-0,21	0,05	0,81
7_1	0,78	0,05	0,05	0,62
7_2	0,84	0,10	0,06	0,71
7_3	0,80	-0,07	0,08	0,65
8_1	0,56	0,06	0,05	0,32
8_2	0,49	0,28	0,10	0,33
8_3	0,51	0,18	0,09	0,30
8_4	0,72	0,03	0,03	0,52
8_5	0,06	0,45	0,26	0,27
9_1	0,63	0,20	0,03	0,44
9_2	0,72	0,16	0,02	0,54
9_3	0,65	-0,06	0,05	0,43
9_4	0,74	-0,19	0,01	0,58
9_5	0,79	-0,25	0,01	0,68
10_t	0,83	-0,25	0,31	0,84

10_s	0,85	-0,24	0,32	0,88
10_z	0,86	-0,14	0,36	0,89
10_p	0,86	-0,09	0,37	0,88

Rysunek 4. Wykres ładunków czynnikowych



Rysunek 4 obrazuje graficznie strukturę latentną testu. Czynnik 1 można interpretować jako umiejętność posługiwania się językiem angielskim, Czynnik 2 natomiast jako zmienną niepowiązaną z mierzonym przez ogół zadań konstruktem, która wyjaśnia najwięcej wariancji pozostałej po wyodrębnieniu Czynnika 1. Wykres informuje, że Czynnik 2 jest najsilniej związany z pytaniami, które wykazywały szereg problemów we wcześniejszych analizach. Ponownie widać, iż pozycja 8_5 oraz pozycje z Zadania 1 mają najmniej wspólnego z pozostałymi pytaniami i najsłabiej ładują mierzony konstrukt.

Należy więc stwierdzić, że test spełnia założenie jednoczynnikowości w zadowalającym stopniu, choć oczywiście kilka pozycji testowych może wymagać pewnej korekty.

4. Zróżnicowane funkcjonowanie pozycji testowej

O zróżnicowanym funkcjonowaniu pozycji testowej (*differential item functioning, DIF) mówimy, gdy osoby o tym samym szacowanym poziomie umiejętności mierzonej przez test mają różne prawdopodobieństwa udzielenia poprawnej odpowiedzi na tę pozycję. Aby stwierdzić DIF należy więc znaleźć warunkowe* względem mierzonej umiejętności różnice międzygrupowe w odpowiadaniu na to pytanie (Kondratek, Skórska i Świst, 2015). Uzyskanie istotnego DIF najczęściej świadczy o tym, że odpowiedzi na dany item zależą od jakiegoś zewnętrznego, niezwiązanego z mierzoną umiejętnością czynnika, którego wartość jest zróżnicowana między badanymi grupami (Kondratek i in., 2015). Na skutek dalszej oceny pytania, np. jego treści lub jej zgodności z wymaganiami testu, identyfikacja DIF może prowadzić do uznania danej pozycji testowej za stronniczą wobec pewnej grupy i nawet do jej eliminacji z puli pytań egzaminacyjnych (Hu i Dorans, 1989). Zawarcie stronniczych pytań w teście stanowi zagrożenie dla trafnego odnoszenia jego wyników do danej grupy i powinno być wychwycone i wyeliminowane na etapie pilotowania pozycji testowych.

W tej części przeanalizowane zostało ewentualne występowanie DIF ze względu na płeć uczniów, stwierdzoną dysleksję rozwojową, lokalizację oraz typ szkoły. Tabela 5 obrazuje, jaki procent pozycji testowych wykazuje DIF ze względu na daną cechę. Ze względu na wielkość próby zrezygnowano z kierowania się miarą istotności statystycznej przy stwierdzaniu DIF i wykorzystywano w tym celu tylko wielkość efektu zróżnicowanego funkcjonowania pozycji testowej. W Tabeli 5 i w dalszych analizach uwzględniono tylko pytania dla których wybrana miara wielkości efektu DIF (tzw. P-DIF) przekroczyła 0,05 („średni DIF) lub 0.1 („silny DIF"; porównaj: Monahan, McHorney, Stump i Perkins, 2007). Do obliczenia zróżnicowanego funkcjonowania pozycji testowej wybrano test ilorazu wiarygodności w modelowaniu IRT (IRT-LR; Thissen, Steinberg i Wainer, 1993).

Tabela 5. Liczba pozycji testowych wykazujących DIF[3].

Kryterium	Liczba pozycji testowych	
	DIF średni (P-DIF<0.1 i >0.05	**DIF silny (P=DIF>0.1)**
płeć	2	0
dysleksja	1	0
lokalizacja szkoły	0	0
typ szkoły (publiczna vs. niepubliczna)	10	1

Tylko znikomy procent pozycji testowych z analizowanego testu wykazuje DIF ze względu na płeć lub stwierdzenie dysleksji rozwojowej, a żadna - ze względu na lokalizację szkoły. Natomiast aż 11 pozycji testowych (jedna czwarta testu) wykazało DIF ze względu na typ szkoły.

Analiza treści wskazanych pytań ujawniła, że w jednym przypadku - pozycji 3_2 wydaje się, że to właśnie treść mogła być przyczyną wykrycia DIF ze względu na płeć i dysleksję. Pozycja ta mierzy rozumienie tekstu słuchanego w nagraniu, które dotyczy udzielania wskazówek topograficznych

[3] Obliczenie DIF dla pozycji testowej 8_5 okazało się być niemożliwe. Może to mieć związek z bardzo niskim dopasowaniem modelu 3PLM do tej pozycji.

i orientacji w terenie. Badania wskazują, że zadania na orientację w terenie i wykorzystywanie wskazówek do poruszania się w nim są trudniejsze dla kobiet (Kelly, McNamara, Bodenheimer, Carr i Riese, 2009). Co więcej, dyslektycy również mają problemy z orientacją w przestrzeni, szczególnie z relacją lewo-prawo (Quercia, Feiss i Michel, 2013). Wydaje się więc, że konieczność orientowania się w wydawanych wskazówkach powoduje niezwiązaną z mierzonym konstruktem wariancję międzygrupową, która prowadzi do identyfikacji DIF. Można oczywiście rozważyć unikanie umieszczanie tego typu pytań na egzaminach, z drugiej jednak strony uzyskiwanie i rozumienie wskazówek przestrzennych jest bardzo ważną umiejętnością, tak w języku rodzimym, jak i obcym (porównaj: Zieky, 2003).

Wyjaśnienie DIF między uczniami szkół publicznych i niepublicznych nastręcza większych trudności. Wydaje się, że identyfikacja tak dużej liczby różnie funkcjonujących pytań może wynikać z dwóch przyczyn: dużej różnicy w poziomie umiejętności między badanymi grupami (DeMars i Wise, 2010) lub też różnicy w zgadywaniu poprawnych odpowiedzi (Finch i French, 2014). Wydaje się, że z powodu wyraźnie niższego poziomu umiejętności uczniów szkół niepublicznych mogą oni częściej uciekać się do zgadywania odpowiedzi, także w sposób losowy, co może prowadzić do międzygrupowych różnic w wartości parametru pseudozgadywania. To z kolei prowadzi do specyficznej formy DIF (tzw. c-DIF, Finch i French, 2014). Warto jednak podkreślić, że większość cytowanych badań opiera się na wynikach egzaminów o niskiej stawce.

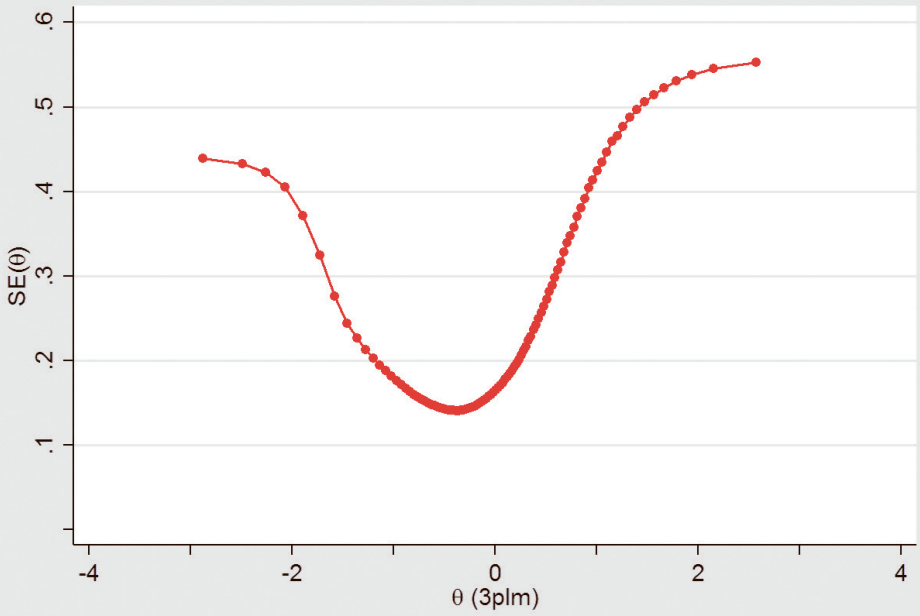
Odrębnym wyjaśnieniem może być wzięcie pod uwagę specyfiki szkół niepublicznych, w których prawie połowa uczniów to osoby dorosłe. Być może więc duża liczba pozycji testowych wykazujących DIF jest związana ze specyfiką tej grupy zdających egzamin. Oprócz wyraźnie niższego poziomu umiejętności, wpływ może mieć także inny poziom motywacji, doświadczenia życiowe inne niż młodzieży licealnej lub różnice w materiale, który nauczyciele są w stanie zrealizować w szkołach tego typu.

5. Wielkość warunkowego błędu pomiaru

Kolejnym cennym źródłem wiedzy o teście jest warunkowy standardowy błąd pomiaru (*conditional standard error of measurement; CSEM)* (Kolen, Zeng i Hanson, 1996). Mówi on o precyzji pomiaru umiejętności uczniów w zależności od poziomu umiejętności uczniów - im większy błąd, tym precyzja mniejsza. Przy stosowaniu podejścia teorii odpowiedzi na pozycje testowe (IRT) warunkowa precyzja pomiaru jest wyrażona na skali wyników przeskalowanych (*scale scores). Precyzja pomiaru umiejętności w danym punkcie umiejętności* (skala theta na osi x, Rysunek 5) zależy od tego, na ile zadania testu potrafią dobrze różnicować uczniów w tym punkcie, co jest silnie związane ze stromością krzywych charakterystycznych zadań w tym punkcie. Na przykład, CSEM jest wysoki dla uczniów o poziomie umiejętności 2 i więcej na skali theta, jest jednak niski dla uczniów w okolicach 0 na tej skali. W testach składających się w znacznej mierze z zadań zamkniętych z wyborem, często obserwuje się spadek precyzji pomiaru dla uczniów o najniższym poziomie umiejętności. Jest to związane z występowaniem zjawiska zgadywania, powodującym, że dla najniższego poziomu umiejętności krzywe charakterystyczne zadań są wypłaszczone, wskazując na jednostajne prawdopodobieństwo udzielenia odpowiedzi poprawnej równe parametrowi pseudozgadywania c. Precyzja pomiaru testem zależy więc jednocześnie od parametrów poszczególnych pozycji testowych w zestawieniu z rozkładem umiejętności uczniów.

Rysunek 5 pokazuje, że dla znacznej większości uczniów precyzja pomiaru jest zadowalająca (SE < 0,4; Thompson, 2008). Wartość błędu pomiaru dla testów przesiewowych, w których chodzi o wykrycie uczniów, którzy nie osiągnęli pewnego minimalnego poziomu umiejętności (a takim jest matura podstawowa z języka angielskiego) powinna być najniższa w rejonie progu zdawalności testu (*cut score)*. W przypadku analizowanego egzaminu próg odcięcia wynosił 15 punktów, co odpowiada wartości około –1,90 na skali wyników standardowych. Jest to rejon, w którym precyzja oszacowania umiejętności spada, zapewne w skutek wysokiej wartości parametru c dla wielu pozycji testowych. W przyszłości można by zadbać o zmniejszenie możliwości zgadywania, co wpłynęłoby pozytywnie na precyzję pomiaru umiejętności uczniów wokół liczby punktów oznaczającej zdanie testu. Test ma dość słabą precyzję jeśli chodzi o szacowanie umiejętności uczniów osiągających wysokie wyniki. Nie jest to jednak powód do zmartwień, gdyż egzamin maturalny na poziomie podstawowym nie służy do różnicowania wśród uczniów dobrych.

Rysunek 5. Wielkość warunkowego błędu pomiaru



6. Podsumowanie i wnioski

Wyniki uczniów w zakresie wszystkich czterech obszarów – rozumienia ze słuchu, rozumienia wypowiedzi pisemnych, znajomości środków językowych i tworzenia wypowiedzi pisemnych - badanych na egzaminie maturalnym z języka angielskiego na poziomie podstawowym są wysokie (więcej informacji- Szpotowicz i in., 2016). Sufitowy efekt, charakterystyczny dla egzaminu w starej formule, uwidacznia się także w egzaminie przeprowadzonym w 2015 roku, choć w dużej mierze może być związany z przystąpieniem do niego tylko części absolwentów liceów ogólnokształcących. Do rozważań na temat poziomu trudności egzaminu należy zatem powrócić w przyszłym roku, gdy przystąpią do niego także absolwenci innych typów szkół.

Właściwości psychometryczne testu są dobre, jedynie kilka pytań wykazuje problematyczne charakterystyki, a właściwie tylko jedno powinno zostać wymienione, ze względu na uzasadnione podejrzenie, że jest ono dla uczniów zbyt trudne Jedynie kilka pozycji testowych wykazuje zróżnicowane funkcjonowanie ze względu na dane różnice międzygrupowe. Oznacza to, że DIF jest znikomym zagrożeniem dla trafności analizowanego testu i że jego autorzy dobrze poradzili sobie z odfiltrowaniem treści potencjalnie stronniczych wobec pewnej grupy uczniów. W kolejnych odsłonach egzaminu należy utrzymać jego wysoki standard, można jednak rozważyć rezygnację lub zmianę pewnych formatów zadań, które premiują zgadywanie (np. Zadanie 1 i Zadanie 2) oraz lepsze monitorowanie treści pytań, tak by nie zawierały one słownictwa lub struktur wykraczającego poza wymagany na egzaminie poziom. Pozycja testowa 8_5 jest dobrym przykładem problemów, które powodują pozycje testowe wykraczające poza wymagania testowe. Test ma dobrą precyzję pomiaru, można jedynie zalecić ograniczenie możliwości zgadywania. Powinno to spowodować spadek błędu pomiaru w okolicy progu zdawalności matury (15 punktów).

Przeprowadzone analizy dostarczyły informacji na temat wysokich własności psychometrycznych jednego z najpowszechniej zdawanych egzaminów maturalnych w Polsce. Wyniki pokazują również potencjał badawczy, jaki leży w wynikach zbieranych i udostępnianych przez CKE. Mogą one posłużyć do badania np. wpływu formatu zadania na jego własności psychometryczne, wpływu danych czynników na pojawianie się c-DIFu, czy też psychometrycznych konsekwencji stosowania zadań niedostosowanych poziomem trudności do uczniów. Wysoka jakość zebranych danych, będąca między innymi wynikiem zastosowania ścisłych procedury zbierania wyników i wielkości próby powinna zachęcić badaczy do ich wykorzystywania w celach naukowych. Podsumowując:

- własności psychometryczne testu są dobre
- jakość zebranych danych jest wysoka
- test jest narzędziem rzetelnym, choć niewątpliwie jego rzetelność jest podwyższona przez stosowanie wiązek pozycji testowych
- wymiarowość testu jest zadowalająca, większość pozycji testowych dobrze mierzy umiejętności uczniów w zakresie języka angielskiego
- zróżnicowane funkcjonowanie pozycji testowej (DIF) nie stanowi zagrożenia dla analizowanego testu

- zidentyfikowano kilka pozycji testowych, których własności psychometryczne są słabe, w większości jednak przypadków odpowiada za to format pytań
- udział zgadywania jest dość duży, powodem są najprawdopodobniej ponownie formaty pytań w rodzaju prawda/fałsz, czy łączenie odpowiedzi ze sobą, gdzie występuje mała liczba dystraktorów
- powyższe przyczyny wpływają na to, że warunkowy, standardowy błąd pomiaru (CSEM) jest nieco zbyt wysoki w rejonie progu zdawalności matury

Literatura cytowana

AERA, APA i NCME (2001). *Standardy dla testów stosowanych w psychologii i pedagogice. Gdańsk: GWP.*

Centralna Komisja Egzaminacyjna [CKE] (2015). Sprawozdanie Ogólne z egzaminu maturalnego. Warszawa: CKE.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, s. 297–334.

DeMars, C. E., i Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning?. *International Journal of Testing*, *10*(3), 207-229.

Dudley, A. (2006). Multiple dichotomous-scored items in second language testing: Investigating the multiple true-false item type under norm-referenced conditions. *Language Testing*, *23*(2), 198-228.

Ebel, R. L. (1971). *The Comparative Effectiveness of True-False and Multiple Choice Achievement Test Items.* Paper presented at Annual Meeting of the American Educational Research Association, New York, USA.

Ebel, R. L. (1972). *Essentials of educational measurement.* Oxford: Prentice-Hall.

EnglishProfile (2015). *English Vocabulary Profile.* Dostępny na: www.englishprofile.org/wordlists (dostęp z: 30 października 2015).

Finch, W. H., i French, B. F. (2014). The impact of group pseudo-guessing parameter differences on the detection of uniform and nonuniform DIF. *Psychological Test and Assessment Modeling*, *56*, 25-44.

Frisbie, D. A. (1973). Multiple choice versus true-false: A comparison of reliabilities and concurrent validities. *Journal of Educational Measurement*, *10*(4), 297-304.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of marketing research*, 186-192.

Haladyna, T. M., i Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, *2*(1), 37-50.

Hambleton, R. K., Swaminathan, H., i Rogers, J. H. (1991). *Fundamentals of Item Response Theory* (Measurement Methods for the Social Science). Sage Publications.

Han, K.T. (2012). Fixing the c Parameter in the Three-Parameter Logistic Model. *Practical Assessment, Research and Evaluation,*17(1), s. 1-24.

Hancock, G. R., Thiede, K. W., Sax, G., & Michael, W. B. (1993). Reliability of Comparably Wriften Two-Option Multiple-Choice and True-False Test Items.*Educational and Psychological Measurement*, *53*(3), 651-660.

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., i Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*(1), 153-166.

Hornowska, E. (2007). *Testy psychologiczne. Teoria i praktyka.*Warszawa: Scholar.

Hu, P. G., i Dorans, N. J. (1989). The effects of deleting items with extreme differential item functioning on equating functions and reported score distributions. Referat wygłoszony na: *The annual meeting of the American Educational Research Association, San Francisco*.

Jakubowski, M. i Pokropek, A. (2009). *Badając egzaminy. Podejście jakościowe w badaniach edukacyjnych*. Warszawa: CKE.

Kelly, J. W., McNamara, T. P., Bodenheimer, B., Carr, T. H., i Rieser, J. J. (2009). Individual differences in using geometric and featural cues to maintain spatial orientation: Cue quantity and cue ambiguity are more important than cue type. *Psychonomic bulletin & review*, *16*(1), 176-181.

Kline, T.J. B. (2005). *Psychological Testing: A Practical Approach to Design and Evaluation*. Sage Publications.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*,*33*(2), 129-140.

Konarski, R. (2004). Model cechy latentnej w analizie psychometrycznej testów i pozycji testowych [w:] B. Niemiecko i H. Szaleniec (red.) *Standardy wymagań i normy testowe w diagnostyce edukacyjnej. Kraków: Polskie Towarzystwo Diagnostyki Edukacyjne.*

Kondratek, B., Skórska, P. i Świst, K. (2015). Wprowadzenie do zróżnicowanego funkcjonowania pozycji testowej. [w:] A. Pokropek (red.) *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania*. Warszawa: Instytut Badań Edukacyjnych.

Koniewski, M., Majkut, P., i Skórska, P. (2014) Zróżnicowane funkcjonowanie zadań testowych ze względu na wersję testu. *Edukacja, 1(126),* 68-83.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, *79*(4), 1332-1361.

Matlock-Hetzel, S. (1997). *Basic Concepts in Item and Test Analysis*. Referat wygłoszony na: The Annual Meeting of the Southwest Educational Research Association.

MEN [Ministerstwo Edukacji Narodowej] (2008). Rozporządzenie Ministra Edukacji Narodowej z dnia 23 grudnia 2008 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół (Dz. U. nr 4 z dn. 15 stycznia 2009). Warszawa: Kancelaria Prezesa Rady Ministrów.

Monahan, P. O., McHorney, C. A., Stump, T. E., i Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics*, *32*(1), 92-109.

Nunnally, J. C., i I. H. Bernstein. (1994). *Psychometric Theory. 3rd ed.* New York: McGraw–Hill.

Preacher, K.J., MacCallum, R.C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2(1), 13-43.

Quercia, P., Feiss, L., i Michel, C. (2013). Developmental dyslexia and vision. *Clinical ophthalmology (Auckland, NZ)*, *7*, 869.

Rada Europy (2001). *Common European Framework of Reference for Languages: Learning, Teaching and Assessment.* Cambridge: Cambridge University Press.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3-13.

Schuwirth, L. i Pearce, J. (2014), *Determining the Quality of Assessment Items in Collaborations: Aspects to Discuss to Reach Agreement Developed by the Australian Medical Assessment Collaboration*. AMAC.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. [w:] Robert W. Lissitz (red.), *The concept of validity: Revisions, new directions and applications* (s.19-37). Charlotte: IAP.

Smolik, M. (2012), Języki obce na egzaminach zewnętrznych: innowacje w latach 2012-2015. *Języki Obce w Szkole: 1/2012*.

Storey, A. G. (1966). A Review of Evidence or the Case Against the True-False Item. *The Journal of Educational Research*, *59*(6), 282-285.

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.*

Szpotowicz, M., Muszyński, M., Gajewska-Dyszkiewicz, A., Paczuska, K. i Kondratek, B. (2016). *Umiejętności maturzystów 2015. Obowiązkowy egzamin maturalny z języka angielskiego w 2015 roku.* Warszawa: Instytut Badań Edukacyjnych.

Tabachnik, B. G. i Fidell, L. S. (2007). Principal Components and Factor Aalysis. [W:] B. G. Tabachnik and L. S. Fidell. *Using Multivariate Statistics* (5th Ed.), Ch. 13. New York: Pearson Education Inc.

Thissen, D., Steinberg, L., i Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. [w:] P.W. Holland i H.Wainer (red.), *Differential item functioning* (s. 67-115). Hillsdale: Lawrence-Earlbaum.

Thompson, T. D. (2008). Growth, precision, and CAT: An examination of gain score conditional SEM. [w:] *Annual meeting of the National Council on Measurement in Education, New York,* (s. 1-31).

Zieky, M. (2003). *A DIF primer.* Princeton: Educational Testing Service.

Załącznik 1.Macierz korelacji polichorycznych pozycji testowych.

	pz_1_1	pz_1_2	pz_1_3	pz_1_4	pz_1_5	pz_2_1	pz_2_2	pz_2_3	pz_2_4
pz_1_1	1								
pz_1_2	.53668407	1							
pz_1_3	.2385	.49584417	1						
pz_1_4	.2662323	.02601295	.47407275	1					
pz_1_5	.18511114	.29305415	.11066831	.61707441	1				
pz_2_1	.44182309	.33978135	.4268422	.29767408	.3738719	1			
pz_2_2	.39804482	.30368006	.40492747	.31405734	.36140443	.67364423	1		
pz_2_3	.4149294	.32913541	.44974175	.37302266	.40429891	.78023795	.89213787	1	
pz_2_4	.29167307	.21354186	.29397349	.24298768	.28438677	.60708909	.575408	.73224058	1
pz_3_1	.40036813	.30862335	.40684939	.35714413	.40474059	.53467844	.51225152	.54114738	.3962225
pz_3_2	.41239141	.34736754	.46303883	.38356408	.40958995	.62682365	.58328354	.61704747	.4271943
pz_3_3	.34427878	.26887741	.35156239	.27908715	.34216888	.49254967	.45385338	.47157052	.35276066
pz_3_4	.38423059	.28271168	.33269953	.2370476	.31943305	.63869721	.55906824	.55912469	.4324847
pz_3_5	.44153981	.34928629	.44928134	.34146385	.40386693	.68668443	.61951834	.63520279	.4826784
pz_3_6	.43325768	.35077806	.45556433	.35266254	.40117426	.67646094	.62112534	.63763466	.47322294
pz_4_1	.45114735	.32696049	.39099569	.26088826	.34772571	.70431156	.63236288	.62214092	.47131637
pz_4_2	.43250876	.30861443	.32015874	.17667599	.29383338	.66914999	.59012258	.57069331	.45694175
pz_4_3	.4243432	.31888656	.38908009	.26519338	.34038771	.67094272	.60798676	.61027207	.44892113
pz_4_4	.41622039	.31327869	.32167498	.1833911	.30557221	.67118185	.59309663	.581339	.45747515
pz_5_1	.46641322	.33501746	.32101093	.12626768	.30019668	.73193361	.62819101	.59672863	.47924539
pz_5_2	.36368949	.27231201	.3019785	.20104571	.28865552	.55179847	.50074256	.49145101	.38968201
pz_5_3	.40375637	.30970316	.40676459	.34522817	.40901173	.56651298	.53742625	.55273001	.4082176
pz_6_1	.41324154	.29831739	.27502893	.12802685	.28022463	.65528406	.5749671	.54251583	.45728413
pz_6_2	.44480501	.32281239	.4080633	.2817845	.37306359	.68072452	.61335773	.61159755	.45732297
pz_6_3	.42477263	.3012622	.39092799	.2764836	.35831874	.64153571	.57839238	.58101187	.44652106
pz_6_4	.36216571	.25707779	.27123674	.18628978	.27999666	.5841767	.51784551	.50930458	.44214505
pz_6_5	.46179768	.33048987	.3705115	.1916629	.31904788	.73146794	.64301444	.62528159	.47711945
pz_7_1	.4107189	.30726804	.41500211	.32123463	.36955827	.65141444	.61192203	.61823777	.46319575
pz_7_2	.44424494	.351965	.47098539	.36999884	.41653693	.69554909	.6493808	.66368617	.492292
pz_7_3	.41190472	.30350543	.37340808	.25289136	.3220982	.67020943	.59976233	.59518036	.45677071
pz_8_1	.29629703	.23432495	.28612426	.22704949	.27992693	.46541088	.41527558	.41965112	.32516759
pz_8_2	.29418868	.26001091	.31165884	.27967703	.32116198	.37262858	.36719517	.39498667	.28927669
pz_8_3	.29470664	.23482021	.28834019	.24519005	.28812147	.40257825	.38198268	.39366777	.291829
pz_8_4	.3712913	.28248073	.35964615	.26096555	.33027757	.60021933	.54284194	.55350569	.41323173

pz_8_5	.10526755	.12132535	.1299163	.17158656	.19188798	-.00418008	.03236568	.0884853	.04285783
pz_9_1	.34304535	.28847509	.37598338	.31807009	.3393623	.52873579	.49808506	.51745941	.36852728
pz_9_2	.39109767	.32226084	.41126814	.33551149	.37823631	.59909965	.56301211	.57667023	.41444058
pz_9_3	.34428875	.2502266	.29732339	.22636261	.29221073	.53316297	.48891474	.48460979	.39243535
pz_9_4	.36765143	.2637671	.2913096	.17581823	.28152043	.60753003	.53536655	.51884222	.42714796
pz_9_5	.40987893	.29336742	.29339716	.15492189	.28220438	.65036433	.57049675	.55268663	.44126221
pz_10_t	.42228341	.29778163	.34508393	.23643608	.33241969	.65277768	.58176327	.55803627	.47132402
pz_10_s	.43598653	.31203726	.35944651	.24576331	.34575757	.67258558	.59758939	.57679554	.48274702
pz_10_z	.43779892	.32665905	.39887175	.30894352	.38629989	.67177085	.61705937	.6007927	.49207749
pz_10_p	.43990958	.33123928	.41058041	.32554646	.39616777	.66599538	.61282881	.6014502	.49516755

	pz_3_1	pz_3_2	pz_3_3	pz_3_4	pz_3_5	pz_3_6	pz_4_1	pz_4_2	pz_4_3
pz_3_1	1								
pz_3_2	.5735811	1							
pz_3_3	.40511677	.55354459	1						
pz_3_4	.46699577	.57014995	.42966753	1					
pz_3_5	.53692382	.64460866	.50719546	.65850525	1				
pz_3_6	.55248676	.6181354	.50386954	.6417478	.71041298	1			
pz_4_1	.51536235	.57065462	.47127799	.60701527	.6638537	.64597311	1		
pz_4_2	.47026233	.51064171	.44519401	.59514015	.63515041	.61166943	.82393992	1	
pz_4_3	.51059628	.56509041	.45046677	.56949202	.63681582	.62635226	.76301121	.67352676	1
pz_4_4	.45916809	.52114467	.44162494	.61763893	.63143015	.62087619	.72407831	.75601845	.72890351
pz_5_1	.48185324	.52739465	.48733591	.64844398	.70036145	.65755434	.77025582	.766058	.70572743
pz_5_2	.41644146	.44882535	.38221032	.47981797	.52909044	.50737554	.57025631	.56665931	.53775537
pz_5_3	.51773942	.51539353	.43557398	.46068204	.56031757	.55517447	.57274701	.53688723	.56311148
pz_6_1	.43131138	.46027322	.42407912	.58679608	.6302761	.59033228	.69871026	.70973457	.64159887
pz_6_2	.52143857	.5754891	.46452193	.59756527	.64141919	.62594009	.69493418	.6766417	.66667653
pz_6_3	.49533988	.54375295	.46292912	.55209707	.61951167	.59978077	.6588609	.63400204	.61891947
pz_6_4	.40632222	.43886561	.39144049	.52097672	.55832837	.53732225	.60755287	.61931117	.56289394
pz_6_5	.50029092	.57024609	.4917495	.62913947	.69186638	.65999385	.76467085	.74996447	.71311425
pz_7_1	.51609456	.56168326	.45284666	.53830429	.62626685	.6172923	.65867079	.61860757	.63477428
pz_7_2	.5527454	.62605499	.49528914	.5876477	.67404108	.66708885	.69246835	.65631625	.66870479
pz_7_3	.47989396	.55726553	.44318225	.57265817	.63332359	.62358215	.68497186	.65833134	.65056984
pz_8_1	.37214301	.38760333	.33093815	.40234286	.43902743	.44211869	.45430712	.44772229	.44227609
pz_8_2	.40988421	.40298979	.33963284	.32678419	.40997171	.40604973	.36173864	.32958439	.36345909
pz_8_3	.37407441	.37935422	.31432836	.34067282	.41651566	.40539386	.4045636	.38005194	.3992385
pz_8_4	.45998407	.50896772	.42947899	.49814002	.58120727	.56418346	.61039787	.5714356	.58111152
pz_8_5	.19469953	.13520377	.14475715	-.02982143	.0667625	.08535207	-.01214606	-.06975723	-.00121086

pz_9_1	.43701662	.49860588	.38892252	.41828729	.52070686	.52285117	.50617049	.44749268	.49576277
pz_9_2	.49893119	.55956574	.43989027	.51645728	.5967706	.58832621	.58524279	.55178819	.57387009
pz_9_3	.41384189	.42776092	.36182615	.43566749	.50677917	.49032217	.55212058	.52699461	.52464417
pz_9_4	.41604811	.44606632	.40427973	.51960496	.56276432	.54594703	.63849972	.63100816	.58396706
pz_9_5	.43634601	.47770618	.41489957	.56583909	.61729933	.59216233	.68898317	.68330878	.63696635
pz_10_t	.46144983	.47524785	.43907104	.55588284	.61245437	.58264621	.68475392	.69951163	.63668128
pz_10_s	.48251754	.50702971	.45710064	.57539529	.63825415	.60629932	.69998744	.71192244	.64927968
pz_10_z	.51400236	.52229162	.46878586	.56803133	.64132975	.61661845	.69807815	.7056165	.66127777
pz_10_p	.52236766	.53226444	.46887413	.56706714	.64082616	.61693195	.68819431	.69755179	.65540688

	pz_4_4	pz_5_1	pz_5_2	pz_5_3	pz_6_1	pz_6_2	pz_6_3	pz_6_4	pz_6_5
pz_4_4	1								
pz_5_1	.75193812	1							
pz_5_2	.56795638	.70130695	1						
pz_5_3	.51803311	.58100585	.52716853	1					
pz_6_1	.70120902	.78609244	.5786478	.49740577	1				
pz_6_2	.66628079	.74095777	.5582561	.56785234	.71951114	1			
pz_6_3	.61568733	.69159168	.53050502	.54428436	.63731616	.63635945	1		
pz_6_4	.61135288	.66613362	.50831736	.46002175	.6586561	.59672091	.60171844	1	
pz_6_5	.72940063	.82921996	.61548212	.58203621	.76193418	.75624828	.71568712	.63976065	1
pz_7_1	.60712987	.65951272	.52234346	.56246954	.60537624	.64294169	.61696468	.55503162	.6723349
pz_7_2	.64793394	.69633682	.5477202	.59655666	.63213698	.67097429	.65122969	.58278647	.71071521
pz_7_3	.65885389	.70545029	.53499213	.53551884	.65321231	.661014	.6215821	.59452703	.71031257
pz_8_1	.44734607	.48542366	.37303793	.39872936	.4525929	.43723544	.4216431	.409921	.48187212
pz_8_2	.33440893	.33501348	.31152456	.43062046	.31282134	.38801128	.36925	.31069668	.34586409
pz_8_3	.37827566	.40013689	.32457172	.40390112	.35823385	.40531929	.39015312	.32692741	.41881403
pz_8_4	.57405979	.6241129	.48323034	.50811653	.56489941	.58132122	.56384095	.50538065	.63334002
pz_8_5	-.07280545	-.1182949	.00842833	.20430379	-.11425567	.05682324	.03182437	-.05949806	-.08594361
pz_9_1	.44685049	.45653521	.3961573	.48927396	.41126014	.49515644	.48368588	.39960759	.49880625
pz_9_2	.55752993	.58541058	.47457456	.53040885	.52899413	.56955821	.55746107	.48111091	.60342866
pz_9_3	.515849	.56957067	.43971582	.47427392	.52594956	.52570434	.50346463	.47629988	.57547611
pz_9_4	.61580652	.68979242	.52265065	.49712263	.63670422	.60200166	.57345471	.55119324	.67124455
pz_9_5	.66901713	.74452405	.5484007	.49895864	.68957475	.65437862	.61230543	.59847178	.73642204
pz_10_t	.67660236	.78418905	.55984768	.54253205	.7206337	.62715953	.61850603	.60522545	.75410657
pz_10_s	.69284471	.78815719	.57703567	.55944512	.72619993	.65046964	.63637784	.6195795	.76545002
pz_10_z	.68560659	.78729187	.57408474	.59210169	.71764551	.64546495	.64318496	.62057137	.76493054
pz_10_p	.67671872	.77191047	.56793092	.60043926	.70600622	.63784795	.63744559	.61104979	.75395391

	pz_7_1	pz_7_2	pz_7_3	pz_8_1	pz_8_2	pz_8_3	pz_8_4	pz_8_5	pz_9_1
pz_7_1	1								
pz_7_2	.74195684	1							
pz_7_3	.69498688	.75875804	1						
pz_8_1	.43069022	.46100124	.44417108	1					
pz_8_2	.38431862	.42399395	.35503403	.45437386	1				
pz_8_3	.39808205	.43444499	.38627233	.2622988	.42335248	1			
pz_8_4	.57784876	.61503818	.58194357	.39877199	.38264732	.43478524	1		
pz_8_5	.03684079	.06852195	-.03213792	.09809652	.17861975	.20334055	.09729375	1	
pz_9_1	.50925454	.55667218	.48577995	.36180942	.36174597	.35060772	.45687461	.12977484	1
pz_9_2	.57119834	.62269298	.56428876	.39665176	.39910734	.38456209	.52816689	.10302809	.60436551
pz_9_3	.51574662	.55410924	.52910382	.3787665	.29928813	.33412224	.4753416	.02675383	.49348376
pz_9_4	.56409818	.58716016	.59438407	.41378224	.29078503	.34457082	.52375625	-.07288583	.46002058
pz_9_5	.60121991	.63632512	.64595915	.44582521	.31411358	.36806172	.56703807	-.09611412	.44151891
pz_10_t	.61222611	.63414823	.64444165	.44576774	.33099068	.37878574	.58793312	-.00267608	.46519403
pz_10_s	.62720733	.6584093	.65791422	.46301263	.35008185	.39680039	.60424437	-.01835309	.4778531
pz_10_z	.64780194	.67126178	.66274257	.47016627	.4004343	.42133	.61487036	.08686463	.52224304
pz_10_p	.64579704	.67356434	.6619738	.47275804	.41673165	.4235329	.61535867	.13335248	.53468371

	pz_9_2	pz_9_3	pz_9_4	pz_9_5	pz_10_t	pz_10_s	pz_10_z	pz_10_p	
pz_9_2	1								
pz_9_3	.3632891	1							
pz_9_4	.53715641	.57873645	1						
pz_9_5	.55823741	.59384338	.64464005	1					
pz_10_t	.53708778	.55171713	.64838227	.69856392	1				
pz_10_s	.55788483	.5609697	.65713899	.71051452	.91269317	1			
pz_10_z	.58619791	.57566096	.65541946	.70310679	.8937235	.90613522	1		
pz_10_p	.58991191	.57863478	.6506232	.69529214	.84766203	.90448801	.89622743	1	