dr Artur Pokropek
Instytut Filozofii i Socjologii PAN
Zespół Edukacyjnej Wartości Dodanej

Matura z języka polskiego
Wybrane problemy psychometryczne

Wprowadzenie

Od 2005 roku kilkaset tysięcy maturzystów corocznie zdaje egzamin z języka polskiego. Po napisaniu tego egzaminu każdy z nich dostaje swój wynik przedstawiony na skali procentowej, który ma odzwierciedlać jego umiejętności władania językiem ojczystym. Od tego wyniku w dużej mierze zależą losy maturzysty.

Jak pisze Ministerstwo Edukacji Narodowej, matura jest *formą oceny poziomu wykształcenia ogólnego, sprawdza wiadomości i umiejętności, ustalone w standardach wymagań będących podstawą do przeprowadzania egzaminu maturalnego określonych w rozporządzeniu Ministra Edukacji Narodowej*[1]. Centralna Komisja Egzaminacyjna wylicza natomiast zalety tego egzaminu, który zgodnie z deklaracją tej instytucji zapewnia:

a. jednolitość zadań i kryteriów oceniania w całym kraju,
b. porównywalność wyników,
c. obiektywizm oceniania (kodowane prace maturalne, oceniane przez zewnętrznych egzaminatorów),
d. rzetelność oceniania (wszystkie oceny są weryfikowane)
e. możliwość przyjęcia na uczelnię bez konieczności zdawania egzaminu wstępnego. [2]

Czy test maturalny z języka polskiego rzeczywiście jest dobrym narzędziem pomiarowym? W tym artykule matura z języka polskiego zostanie przeegzaminowana za pomocą narzędzi psychometrycznych po to, by odpowiedzieć na to pytanie. Przy czym zaznaczyć należy, iż nie będzie to kompletna diagnoza egzaminu maturalnego z języka polskiego, a raczej wypunktowanie kilku ważnych problemów, które mogą stać na przeszkodzie realizacji tego, o czym piszą Ministerstwo Edukacji Narodowej oraz Centralna Komisja Egzaminacyjna.

Przeanalizowane zostaną trzy kluczowe kwestie związane z jakością narzędzia pomiarowego używanego podczas egzaminu maturalnego z języka polskiego: Czy jest rzetelny? Czy jest obiektywny? Czy jest porównywalny?

[1] http://www.men.gov.pl/index.php?option=com_content&view=article&id=556&Itemid=2

[2] http://www.cke.edu.pl/images/stories/Inf_mat_od2008/inf_polski_a.pdf

Rzetelność

Rzetelność jest fundamentalną własnością testu, która decyduje o precyzji pomiaru umiejętności uczniów. Jeżeli test nie jest rzetelny, precyzja oceny umiejętności ucznia na jego podstawie jest niewielka. Wiąże się to oczywiście z zagadnieniami sprawiedliwości i obiektywności oceniania umiejętności uczniów. Jeżeli test jest nierzetelny, czyli ma małą precyzję oceniania umiejętności ucznia, nie można mówić o nim jako o narzędziu obiektywnym i sprawiedliwym. Na podstawie testu o niskiej rzetelności nie powinno się wypowiadać o osiągnięciach szkolnych ucznia, czy podejmować na jego podstawie decyzji, które mogą wpływać na indywidualne losy jednostki.

Istnieje kilka miar rzetelności testu, jednak zdecydowanie najpopularniejszą jest Alfa Cronbacha. Dzięki temu wskaźnikowi można oszacować przybliżoną wartość rzetelności testu. Obliczenie wskaźnika nie jest trudne i sprowadza się do użycia jednego wzoru (1):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_{suma}^2}\right)$$

gdzie:

k – liczba zadań w teście,

α_i^2 – wariancja dla *i*-tego zadania,

σ^2_{suma} – wariancja dla wyniku sumarycznego.

Za pomocą Alfy Cronbacha nie da się obliczyć dokładnej wartości rzetelności, gdyż podczas jej szacowania wykorzystuje się jedynie informacje na temat wewnętrznej spójności testu, nie mówi się tu o błędach, które wynikają z efektu egzaminatora, który – jak wiemy z nielicznych badań (Dolata i inni 2004) – jest dosyć wysoki, a przynajmniej był bardzo wysoki dla pierwszych ogólnopolskich egzaminów maturalnych. W przypadku egzaminów z pytaniami otwartymi Alfę Cronbacha można zatem traktować jako górną granicę rzetelności, a nie, jak chciałaby tego klasyczna teoria testów – nieuwzględniająca efektów egzaminatorów, za jej dolne ograniczenie (por. Brennan 2001).

O wysokich wskaźnikach rzetelności mówi się, gdy wskaźnik Alfa Cronbacha jest wyższy niż 0,9. W przedziale 0,70-0,89 uznaje się zwyczajowo, iż rzetelność nie jest wysoka, lecz jest do zaakceptowania. Gdy rzetelność testu znajduje się w przedziale 0,60-0,69, jest to sytuacja alarmująca, choć w określonych wypadkach do zaakceptowania (tylko gdy wynik testowy jest zmienną zależną). Testy charakteryzujące się rzetelnością poniżej 0,60 w zasadzie nie nadają się do dalszych analiz statystycznych, miast informacji niosą one przede wszystkim szum spowodowany błędem pomiaru (Jakubowski, Pokropek 2009).

Tabela 1. Wartości wskaźnika Alfa Cronbacha i jego interpretacja

Alfa Cronbacha	Interpretacja
1,00-0,90	Wysoka rzetelność, test bardzo rzetelny
0,70-0,89	Rzetelność przeciętna, spójność testu do zaakceptowania
0,60-0,69	Rzetelność niska, w pewnych warunkach do zaakceptowania
0,00-0,59	Bardzo niska rzetelność, sugeruje, że test nie jest spójny i wskazuje na potrzebę jego analizy merytorycznej

W tabeli 2. wraz ze statystykami opisowymi testu maturalnego przedstawiono wskaźnik Alfa Cronbacha. Jak widać, matura w części podstawowej charakteryzuje się przeciętną rzetelnością (0,78), zaś matura na poziomie rozszerzonym rzetelnością niską (0,65). Biorąc pod uwagę to, iż egzamin maturalny z języka polskiego jest egzaminem państwowym o wielkiej wadze dla zdających, wartości rzetelności są w tym wypadku zdecydowanie niezadowalające. Po egzaminie tej wagi powinniśmy spodziewać się rzetelności na poziomie przynajmniej 0,90, co jest ogólnie przyjętą normą dla testów wysokiej stawki. Widoczne wartości Alfy Cronbacha nie w pełni uzmysławiają znaczenie faktu, iż pomiar umiejętności uczniów na maturze z języka polskiego charakteryzuje się niską rzetelnością, a tym samym niską precyzją określania wyników dla poszczególnych uczniów.

Tabela 2. Statystyki opisowe egzaminu maturalnego z języka polskiego, część podstawowa i rozszerzona rok 2010[3]

Matura z języka polskiego 2010	Część podstawowa		Część rozszerzona	
Alfa Cronbacha	0,78		0,68	
Liczba zdających	359 245		31 659	
	Punkty	Procenty	Punkty	Procenty
Błąd standardowy pomiaru	2,9	7,10%	1,66	11,19%
Średnia	23,19	56,56%	8,65	57,74%
Odchylenie standardowe	6,20	15,13%	2,94	19,78%
25% centyl	19	46,34%	7	46,66%
75% centyl	28	68,29%	11	73,33%

Aby uzmysłowić sobie wielkość błędu popełnianego podczas indywidualnego pomiaru, warto oszacować standardowy błąd pomiaru. Można to zrobić za pomocą prostego równania, jeśli znamy oszacowanie rzetelności – Alfę Cronbacha (α) oraz odchylenie standardowe wyniku sumarycznego (σ_{suma}):

[3] Dane zbierane na potrzeby projektu Edukacyjnej Wartości Dodanej. Dane różnią się nieznacznie od danych prezentowanych przez CKE. Różnice te nie przekraczają kilku procent. Różnice te nie wpływają na wartość oszacowania rzetelności oraz błędu standardowego pomiaru.

$$SEM = \sigma_{suma} * \sqrt{1-\alpha}$$

Błąd standardowy pomiaru jest wartością analogiczną do błędu standardowego liczonego dla średniej szkolnej, czy dla miary EWD. Za jego pomocą dowiadujemy się, jak precyzyjnie szacowany jest wynik poszczególnych uczniów. Podobnie jak w przypadku wyników średniej dla szkoły, tak też dla poszczególnego ucznia można oszacować przedział ufności, który z określonym prawdopodobieństwem (poziomem ufności) będzie zawierał wynik prawdziwy.

Gdy oszacujemy przedział ufności na poziomie 95%, czyli pozwolimy sobie na pomyłkę jedynie w 5 przypadkach na 100, przykładowo dla ucznia, który uzyskał wynik 60% z egzaminu maturalnego z języka polskiego na poziomie podstawowym, przedział ufności rozciągał się będzie w granicach +/- 13,9 punktów procentowych (+/-1,96*SEM) od empirycznego wyniku ucznia (sic!). Innymi słowy na podstawie 60% wyniku ucznia możemy powiedzieć tyle tylko, że na 95% jego wynik prawdziwy zawiera się w przedziale 46,1% – 73,9%. Patrząc na ten szeroki przedział ufności, warto zestawić go z wartościami centyli podanymi w tabeli 2. Otóż taka precyzja szacowania przy poziomie ufności 95% nie pozwala nam faktycznie stwierdzić, czy uczeń należy do pierwszego, czy czwartego kwartyla wyników. Warto to dobitnie powtórzyć: na podstawie wyniku z matury z języka polskiego na poziomie podstawowym przy 95% poziomie ufności nie wiemy, czy faktyczny stan umiejętności ucznia, który uzyskał 60% z matury kwalifikuje go do 25% uczniów najsłabszych, czy do 25% uczniów najlepszych. Sytuacja nie ulega znaczącej zmianie, jeżeli zmniejszamy poziom ufności z 95% na 90%. Wtedy przedział ufności szacowany jest jako +/- 11,6 punktów procentowych od wyniku ucznia (1,64*SEM). Nawet jeżeli ustalimy 68% przedział ufności (1*SEM), przedział pozostaje bardzo szeroki +/- 7,1 p.p. Abstrahując od tego, czy w przypadku matury możemy sobie pozwolić na 42 pomyłki na 100, wydaje się ciągle całkiem szeroki. Jego długość równa jest prawie jednemu odchyleniu standardowemu wyników matury w populacji. Sytuacja dla matury w wersji rozszerzonej jest jeszcze bardziej niepokojąca. Przyjęcie 95% poziomu ufności pozwala na konstrukcję przedziałów ufności oddalonych od wyniku ucznia o około +/- 21,9 p.p., czyli o długości około dwóch odchyleń standardowych. 90% poziom ufności daje przedział ograniczony przez +/- 18,4 p.p., a dla 68% poziomu ufności +/- 11,2 p.p., czyli powyżej jednego odchylenia standardowego.

Z psychometrycznego punktu widzenia należy zatem stwierdzić, iż niska rzetelność testu maturalnego z języka polskiego dla obydwu poziomów nie pozwala go traktować jako wiarygodnego źródła informacji o poziomie osiągnięć szkolnych uczniów.

Obiektywność

Przez pomiar obiektywny rozumiemy taki pomiar, który uczniom o tym samym poziomie umiejętności gwarantuje, jeżeli test byłby doskonale rzetelny, możliwość uzyskania takiego samego wyniku. Postulat ten sprowadza się zasadzie do bardzo prostego wymagania. Każdy z uczniów powinien otrzymać test

o takiej samej trudności (chyba, że zastosowana zostanie procedura skalowania zrównująca skale testów, co w przypadku polskiej matury nie ma miejsca).

Czy można powiedzieć, że matura z języka polskiego jest obiektywna? Problem zostanie przeanalizowany na podstawie egzaminu z języka polskiego na poziomie podstawowym. Znaczną część tego egzaminu stanowi wypracowanie. Uczeń może wybrać jeden spośród dwóch tematów. Niestety okazuje się, iż tematy tych wypracowań różnią się od siebie pod wieloma względami, a przede wszystkim pod względem trudności. Doskonałym przykładem jest tutaj egzamin z roku 2010, gdzie nawet bez odwołania się do opinii eksperckich czy analiz statystycznych, można stwierdzić, iż egzaminy fasadowo różnią się złożonością poruszanych zagadnień i trudno jest postulować między nimi ekwiwalentność:

Temat 1. Na podstawie podanych fragmentów komedii Moliera *Świętoszek* scharakteryzuj głównego bohatera oraz omów postawy Orgona, Kleanta i Elmiry wobec tytułowej postaci.

Temat 2. Na podstawie podanego fragmentu utworu Hanny Krall *Zdążyć przed Panem Bogiem* przedstaw przemyślenia Marka Edelmana o możliwościach godnego życia w czasach Zagłady i różnych poglądach na temat godnej śmierci.

W bazach danych Centralnej Komisji Egzaminacyjnej nie ma dostępnej informacji, który temat został wybrany przez danego ucznia. Informacją o wyborze tematu dysponują tylko okręgowe komisje egzaminacyjne, które zdecydowały się dla własnego użytku zapisywać taką informację. Wśród takich komisji jest Okręgowa Komisja Egzaminacyjna w Łodzi, która udostępniła dane do analiz statystycznych mających skwantyfikować problem wpływu swobodnego wyboru tematów wypracowań na konstrukcję skali osiągnięć maturzystów z języka polskiego.

W tabeli 3. przedstawione zostały procentowe rozkłady poprawnych odpowiedź uczniów w zależności od tego, który temat z wypracowania został wybrany (dane dla OKE Łódź). Uczniowie wybierający temat pierwszy uzyskali z pozostałej części testu 73,5% poprawnych odpowiedzi, a uczniowie wybierający temat drugi o 3,4 p.p. mniej. Sugeruje to, iż uczniowie lepsi wybierali temat fasadowo łatwiejszy. Sytuacja ta nie tłumaczy jednak diametralnych różnic w procencie poprawnych odpowiedzi dla poszczególnych kryteriów oceny wypracowania. Uczniowie wybierający temat pierwszy średnio rzecz biorąc uzyskiwali za rozwinięcie tematu ponad dwa razy więcej punktów niż uczniowie wybierający temat drugi. Różnica między tematem pierwszym a drugim dla kompozycji wynosiła ponad 10%, różnice między pozostałymi kryteriami oceny wypracowania (nie licząc szczególnych walorów pracy) wahają się między 6,7% a 7,9%. Wyniki te sugerują, iż występuje przynajmniej słaby proces autoselekcji (uczniowie o wyższych umiejętnościach wybierają temat 1.) oraz istotne różnice w trudności tematów: temat 1. jest łatwiejszy od tematu 2.

Tabela 3. Procent rozwiązanych zadań, j. polski poziom podstawowy, matura 2010, dla uczniów, którzy wybrali różne tematy maturalne, OKE Łódź.

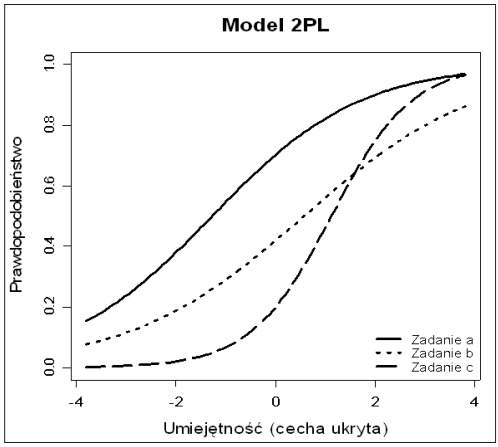
Wskaźnik	temat 1	temat 2	różnica
PYTANIA ZAMKNIĘTE	73,5%	70,1%	3,4
Wypracowanie:			
I.ROZWINIĘCIE TEMATU	73,9%	32,5%	41,3
II. KOMPOZYCJA (maksymalnie 5 punktów)	74,2%	62,5%	11,6
III. STYL (maksymalnie 5 punktów)	72,4%	65,7%	6,7
IV. JĘZYK (maksymalnie 12 punktów)	64,4%	55,8%	8,6
V. ZAPIS (maksymalnie 3 punkty)	62,4%	54,5%	7,9
VI. SZCZEGÓLNE WALORY PRACY	3,9%	3,2%	0,8
N	9 415	14 538	

Aby potwierdzić te przypuszczenia, wykorzystano modelowanie IRT. Wyestymowano 4 modele IRT. Do skalowania wykorzystano dwuparametryczny model IRT dla odpowiedzi częściowej:

$$P_{ik}(\theta_n) = \frac{\exp(\theta_n - a_i(\beta_i + \delta_k))}{1 + \exp(\theta_n - a_i(\beta_i + \delta_k))}$$

W modelu IRT modeluje się prawdopodobieństwo uzyskania poprawnej odpowiedzi na i-ty podpunkt, k-tego pytania dla n-tego ucznia o wyznaczonej wartości cechy ukrytej θ_n. Parametr β_i oznacza trudność pytania *i*. Wyznaczany jest jako punkt na skali umiejętności w miejscu, gdzie prawdopodobieństwo poprawnej odpowiedzi definiowane przez funkcję logistyczną wynosi 0,5. Parametr a_i nazywamy dyskryminacją i mówi on, jak bardzo strome są krzywe prawdopodobieństwa dla danych zadań. Jest to parametr, który w klasycznej teorii testu odpowiada parametrowi mocy różnicującej i w taki sposób można go też interpretować. Parametr δ_k określa trudność każdego ktego podpunktu zadania. Na rysunku 1. przedstawiono trzy zadania o różnych parametrach trudności i różnych parametrach dyskryminacji (przedstawiono tutaj zadania jednopunktowe, czyli w uproszczonej sytuacji, gdzie nie estymuje się parametru δ_k).

Zadanie „*a*" jest najłatwiejsze, zadanie „*b*" jest trudniejsze od zadania „*a*", najtrudniejszym zadaniem jest zadanie „*c*". Nie jest to regułą, lecz w tym konkretnym przypadku okazało się, że najtrudniejsze zadanie ma też najwyższy parametr dyskryminacji, czyli krzywa charakterystyczna jest najbardziej stroma. Natomiast najłatwiejsze zadanie *a* okazało się mieć krzywą charakterystyczną bardziej stromą niż zadanie *b* (zadanie środkowe pod względem trudności). Najmniej stroma spośród przedstawionych krzywych jest krzywa charakterystyczna zadania *b*, dlatego też zadanie to ma najniższą moc dyskryminacyjną.



Rysunek 1. Krzywe charakterystyczne trzech zadań o różnej trudności i różnej dyskryminacji

Dzięki parametrowi dyskryminacji przestawiony model uzyskuje bardzo dobre dopasowania dla danych, a tym samym precyzyjniejsze oszacowania umiejętności uczniów, nawet w wypadku, gdy poszczególne pytania charakteryzują się stosunkowo słabymi właściwościami pomiarowymi. Model bowiem w procesie estymacji każdemu pytaniu nadaje wagę proporcjonalnie do jego dyskryminacji (dyskryminację pytania można uznać za wskaźnik jakości pomiarowej). Innymi słowy, im pytanie jest „lepsze pomiarowo", tym jego udział w oszacowaniu parametru mówiącego o umiejętności n-tego ucznia θ_n jest większe.

Do rozwiązania interesującego nas problemu wpływu wyboru tematu na oszacowanie poziomu umiejętności oszacowanych zostało kilka modeli typu 2PL. W pierwszym modelu język polski został wyskalowany z pominięciem informacji o wyborze tematu (1). W modelu drugim (2) i trzecim (3) testy wyskalowane zostały osobno dla uczniów, którzy wybrali temat pierwszy, i osobno dla uczniów, którzy wybrali temat drugi. Czwarty model (4) to model selekcji. W tym modelu do skalowania wykorzystano dane wszystkich uczniów i wszystkie pytania oraz dodano parametr selekcji. Plan skalowania przedstawiono w tabeli 4.

Tabela 4. Plan skalowania umiejętności uczniów w modelu selekcji

	Pytania			Parametr selekcji
Uczniowie	Pytania testu bez wypracowania	Wypracowanie temat 1.	(…)	1
		(…)	Wypracowanie temat 2.	0

W modelu selekcji częścią wspólną skalowania są pytania sprawdzające rozumienie przeczytanego tekstu. Sześć kryteriów oceniania wypracowania traktowanych jest jako niezależne zadania (w sumie 12 zadań). Sześć pierwszych odnosi się do tematu 1., sześć kolejnych do tematu 2. Uczniowie, którzy wybrali temat pierwszy, w miejsce wartości dla zadań do tematu drugiego mieli braki danych. Podobnie uczniowie, którzy wybrali temat 2. w miejsce zadań

odnoszących się do tematu 1. mieli braki danych. Każdy z uczniów oznaczony został również zmienną 0-1, która wskazywała, czy wybrany został temat 1., czy nie (czyli wybrał temat 2.). Zmienna ta potrzebna jest do oszacowania parametru selekcji, który opisuje relacje między wyborem tematu a umiejętnościami ucznia. Zaletą modelu selekcji jest to, iż szacuje on parametry dla dwóch wypracowań na jednej skali, czego nie możemy być pewni podczas osobnej kalibracji w dwóch grupach uczniów wybierających różne tematy.

W tabeli 5. podano parametry czterech kolejno szacowanych modeli. W tabeli podano parametry dyskryminacji (a) oraz parametry trudności pytań (b). Część tabeli oznaczona szarym kolorem odnosi się do parametrów pytań wypracowania.

Tabela 5. Parametry pytań, j. polski poziom podstawowy, matura 2010, dla uczniów, którzy wybrali różne tematy maturalne, OKE Łódź.

	model (1) razem		model (2) temat 1.		model (3) temat 2.		model (4) m. selekcji (t1/t2)			
	a	b	a	b	a	b	a		b	
p1	0,34	-2,21	0,48	-1,68	0,41	-1,91	0,35		-2,19	
p2	0,26	-4,51	0,32	-3,69	0,30	-3,90	0,26		-4,50	
p3	0,21	-0,01	0,26	0,00	0,24	0,00	0,21		0,00	
p4	0,41	-1,75	0,53	-1,43	0,47	-1,56	0,41		-1,75	
p5	0,26	-2,15	0,32	-1,77	0,28	-1,98	0,26		-2,14	
p6	0,34	-2,25	0,46	-1,73	0,40	-1,95	0,34		-2,21	
p7	0,28	-3,63	0,39	-2,70	0,33	-3,06	0,28		-3,57	
p8	0,21	-6,47	0,29	-4,73	0,27	-5,10	0,21		-6,29	
p9	0,25	-4,80	0,37	-3,40	0,31	-3,94	0,26		-4,74	
p10	0,24	-4,98	0,31	-3,91	0,28	-4,27	0,24		-4,94	
p11	0,26	-4,07	0,30	-3,63	0,31	-3,50	0,26		-4,06	
p12	0,37	-1,96	0,53	-1,45	0,45	-1,67	0,37		-1,94	
p13	0,52	1,40	0,63	1,22	0,58	1,29	0,52		1,40	
p14	0,34	-0,37	0,47	-0,27	0,40	-0,31	0,34		-0,36	
	Wypracowanie						t1	t2	t1	t2
Temat	0,88	0,38	0,91	-0,60	0,96	1,07	0,85	0,94	-0,51	1,00
Kompozycja	1,52	-0,93	1,40	-1,20	1,55	-0,75	1,34	1,54	-1,12	-0,84
Styl	1,75	-1,03	1,63	-1,14	1,78	-0,92	1,62	1,82	-1,04	-1,01
Język	1,73	-0,44	1,63	-0,60	1,76	-0,33	1,58	1,81	0,11	-0,42
Zapis	0,54	-0,50	0,55	-0,71	0,54	-0,35	0,52	0,53	-0,62	-0,44
Walory	0,65	3,43	0,70	3,18	0,66	3,36	0,67	0,65	3,47	3,32
parametr selekcji (t1=1)							0,24		0,64	

a – dyskryminacja; b – trudność zadania (kryterium)

Rzut oka na wyniki dla modeli (2) i (3) prezentowane w tabeli 5. pozwala stwierdzić, iż parametry poszczególnych pytań różnią się między estymowanymi modelami skalowania, szczególnie jeżeli chodzi o wypracowanie. Nie ma ścisłych reguł co do tego, jak duże różnice w estymowanych parametrach pytań mogą wpływać na właściwości parametrów uczniów. „Regułą kciuka" stosowaną w modelu Rascha, którą i tutaj przyjmujemy, są wartości powyżej 0,3 logita dla parametru trudności (Wright i Masters 1982). Dla parametru dyskryminacji, charakteryzującego się mniejszym rozrzutem wyników, można przyjąć wartość 0,1 jako wartość graniczną.

Różnice między wartościami parametrów dyskryminacji i trudności w modelach (2) i (3) przekraczające ustalone kryteria zostały pogrubione w wierszach. Jeżeli patrzymy na zadania egzaminacyjne poza wypracowaniem, różnice w estymowanych parametrach dla uczniów wybierających pierwszy i drugi temat są do zaakceptowania (w przypadku p7 i p8 różnice w trudności między uczniami, którzy wybrali temat 1. a tymi, którzy wybrali temat 2., przekraczają 0,3 logita, lecz odnoszą się do pytań skrajnie łatwych i o bardzo niskiej dyskryminacji, czyli niskim wkładzie do całej skali). Jeżeli chodzi o wypracowania, okazuje się, że kryteria oceniania funkcjonują różnie w zależności od wybranego tematu. A co za tym idzie, porównanie modeli skalowanych osobno dla uczniów, którzy wybrali różne tematy (podobnie jak procent rozwiązań), wskazuje, iż temat pierwszy okazał się oceniany znacznie łagodniej niż temat drugi pod względem oceny rozwinięcia tematu, kompozycji i zapisu. Wyniki te sugerują również, że temat drugi charakteryzuje się silniejszymi wartościami dyskryminacji dla kompozycji, stylu oraz języka.

Jak powiedziane było to już wcześniej, estymacja na dwóch oddzielnych próbach może prowadzić do błędów wywołanych procesem autoselekcji. Jeżeli uczniowie zdolniejsi wybierają łatwiejsze wypracowanie, to parametry trudności pytań mogą być negatywnie obarczone, tj. mogą wydawać się łatwiejsze niż są w rzeczywistości. Ponadto różnice w rozkładzie umiejętności uczniów wybierających różne wypracowania mogą przekładać się na różne wartości parametrów dyskryminacji.

Wyniki modelu selekcji, które uważane są za najbardziej trafne, wskazują, iż pod względem trudności tylko rozwinięcie tematu dla pierwszego wypracowania było łatwiejsze (należy podkreślić, że znacznie łatwiejsze). Istnieją pewne przesłanki, iż kompozycja w temacie drugim oceniana była surowiej – różnica między tematem pierwszym a drugim wynosi 0,28 logita, co jest bardzo bliskie założonego kryterium. Lecz co dla wcześniejszych analiz nie było zauważane, w przypadku tematu 2. znacznie łagodniej oceniany był język. Jeżeli spojrzeć globalnie na wszystkie kryteria, średnia różnica między tematami wynosi 0,22 logita, tak iż temat drugi jest łatwiejszy. O ile wartości 0,22 logita dla jednego pytania nie uznalibyśmy za różnicę istotnie zmieniającą wynik ucznia, to taka różnica rozpatrywana dla grupy pytań musi oznaczać systematyczny błąd w ocenianiu i/lub konstrukcji zadania. Innymi słowy, wynik dla uczniów, którzy wybrali temat pierwszy i temat drugi nie jest ściśle porównywalny. Należy też podkreślić, iż łatwiejszy temat wybierany był przez zdolniejszych uczniów, co sugeruje parametr selekcji. Uczniowie o wyższych

umiejętnościach mierzonych maturą z języka polskiego wybierali łatwiejszy, pierwszy temat wypracowania.

W tabeli 6. można odnaleźć wyniki końcowe uczniów zdających egzamin maturalny z języka polskiego na poziomie podstawowym. Wyniki przedstawiono na skali stosowanej w EWD, czyli skali, dla której średnia wynosi 100, a odchylenie standardowe 15. Przedstawiono średnie wyniki zarówno dla uczniów, którzy wybrali temat pierwszy, jak i drugi. Przedstawiono wyniki dla 6 modeli skalowania. W pierwszym wypadku (1) mamy do czynienia z sumarycznym wystandaryzowanym wynikiem całego testu. W drugim (2) mamy wystandaryzowany wynik oparty o dwuparametryczny model skalowania. W (3) i (4) przypadku widać wynik sumaryczny (3) i wynik modelowania IRT (4) dla zadań z wyłączeniem wypracowania. W kolejnym wierszu (5) podano wyniki uczniów, gdzie odpowiedzi na pytania skalowano modelem IRT osobno dla uczniów wybierających temat pierwszy i osobno temat drugi. Ostatni wynik w tabeli (6) wyraża wyniki uzyskane na podstawie modelu selekcji.

Tabela 6. Wyniki uczniów zdających j. polski na poziomie podstawowym, matura 2010, dla różnych tematów maturalnych, skala 100; 15, OKE Łódź

Wyniki:	temat 1	temat 2	różnica
1) Suma, cały test	105,22	96,62	8,59
2) IRT cały test	104,86	96,85	8,01
3) Suma bez wypracowania	102,04	98,68	3,36
4) IRT bez wypracowania	102,00	98,70	3,30
5) IRT osobno	101,69	98,90	2,79
6) Model selekcji	103,86	97,50	6,36
Liczba zdających	9 415	14 538	

Rzut oka na wyniki pozwala stwierdzić, iż wypracowanie w języku polskim daje silny wkład do różnicowania wyniku uczniów. Skalowanie bez wypracowania (3) i (4) wskazuje na różnice około 3 punktów między uczniami, którzy wybrali różne tematy. Gdy zostanie uwzględnione wypracowanie (1) i (2) różnica ta wzrasta prawie trzykrotnie. Fakt występowania problemu autoselekcji – uczniowie słabsi wybierają trudniejszy temat – sprawia, iż różnicowanie to jest sztucznie zawyżane. Średnia różnica w wartościach między uczniami dla wyniku sumarycznego i wyniku opartego na modelowaniu IRT wynosi około 8 punktów, podczas gdy model selekcji sugeruje, iż różnica ta nie powinna przekraczać 6,4 punktu.

Można zatem powiedzieć, iż uczniowie wybierający temat 1. dostają premię w postaci około 1/10 odchylenia standardowego. Nie jest to błąd bardzo duży, ale znaczący i mogący wpływać na dalsze losy jednostki, jeżeli wynik z matury będzie brany pod uwagę podczas rekrutacji na uczelnie wyższe.

Porównywalność

Zagadnienie porównywalności ma kilka wymiarów. Można zapytać o porównywalność regionalną, można zapytać o porównywalność matur dla uczniów różnej płci. W tym artykule chciałbym się skupić na porównywalności egzaminu maturalnego w kolejnych latach. Najlepiej sprawdzić to empirycznie. Czy wynik egzaminu maturalnego ucznia z języka polskiego z roku 2010 jest porównywalny z wynikiem ucznia w roku 2009? W tabeli 7. przedstawione zostały procentowe wyniki matur z roku 2010 i 2009 na poziomie podstawowym i rozszerzonym.

Tabela 7. Wyniki uczniów zdających j. polski na poziomie podstawowym oraz rozszerzonym, matura 2010 i 2009, dane CKE, skala procentowa

Matura z języka polskiego	Poziom podstawowy		Poziom rozszerzony	
Rok	2010	2009	2010	2009
Średnia	57,24%	52,21%	60,00%	65,00%
Odchylenie standardowe	17,42%	15,06%	16,80%	14,85%

Jak widać, różnica między wynikami maturzystów z tych dwóch edycji egzaminu maturalnego wynosi około 5 p.p. Mało prawdopodobne jest to, żeby cała populacja z roku na rok średnio rzecz biorąc poprawiła swoje umiejętności o około 1/3 odchylenia standardowego. Trudno byłoby też bronić tezy, iż kierunki zmian są różne dla różnych poziomów egzaminu. Na dodatek wyniki innych egzaminów (nieprzedstawianych tutaj) nie sugerują takiej zależności. Z tej perspektywy rozsądnym założeniem jest to, że egzamin maturalny na poziomie podstawowym był w roku 2009 trudniejszy niż w roku 2010, na odwrót zaś w przypadku poziomu rozszerzonego. Różnica trudności na pierwszy rzut oka wydaje się niewielką – 5 p.p. Jednak, jak wspomniano już powyżej, te 5 p.p. to blisko 1/3 odchylenia standardowego wyników surowych. Taka różnica powoduje, iż nie można powiedzieć, że egzamin maturalny z języka polskiego jest porównywalny w kolejnych latach.

Tabela 8. Wartości wybranych centyli (zakładając rozkład normalny) dla wyników uczniów zdających j. polski na poziomie podstawowym, matura 2010 i 2009, skala procentowa

Centyl	Rok 2010		Rok 2009
5	26,9%		27,2%
10	34,1%		32,3%
15	38,6%		36,2%
20	42,1%		39,2%
25	45,2%		42,0%
30	47,9%		44,1%
35	50,5%		46,4%
40	52,6%		48,1%
45	54,9%		49,9%

Centyl	Rok 2010		Rok 2009
50	56,9%		52,0%
55	59,5%		54,1%
60	61,5%		56,0%
65	63,6%		58,2%
70	66,5%		60,3%
75	69,0%		62,1%
80	71,6%		64,7%
85	75,6%		67,4%
90	80,1%		70,4%
95	85,8%		76,1%

Stopień nieporównywalności wyników łatwo sobie wyobrazić, odnosząc się do konkretnych przykładów. Aby to zrobić, w tabeli 8. przedstawiono wartości wybranych centyli dla matury z poziomu podstawowego z języka polskiego osobno w roku 2010 i 2009. Wyobraźmy sobie hipotetyczną sytuację, w której uczeń, który zdawał egzamin maturalny w roku 2009, przystępuje do rekrutacji na studia wyższe w roku 2010. Jeżeli o dostaniu się na studia decyduje jedynie egzamin z języka polskiego, jego sytuacja jest trudna. Jeżeli na egzaminie w 2009 roku nasz uczeń uzyskał trochę powyżej 76%, oznacza to, że znalazł się wśród 5% uczniów o najwyższych wynikach, lecz te same 76% oznacza co innego na skali z roku 2010. Jako że egzamin w roku 2010 był łatwiejszy, to 76% wynik klasyfikowany jest w granicach 85 centyla. Oznacza to, iż wynik procentowy ucznia zdającego egzamin w 2009, a przystępującego do rekrutacji w 2010, niesłusznie kategoryzowany jest o około 10 centyli niżej. Gdyby nasz hipotetyczny uczeń uzyskał w roku 2009 wynik nieznacznie powyżej 67%, oznaczałoby to, że jego wynik jest lepszy od 85% innych wyników. Gdy poda ten wynik na rekrutacji w 2010 roku, zostanie on jednak uznany za wynik lepszy od pozostałych 70% najlepszych wyników. Wszystko dlatego, że zdawał egzamin trudniejszy niż jego koledzy w roku 2010. Nie wymaga to większego komentarza – wyniki procentowe są nieporównywalne i w określonych wypadkach niesprawiedliwe.

Podsumowanie i rekomendacje

Już z krótkiego przeglądu podstawowych problemów związanych z maturą z języka polskiego widać, że z psychometrycznego punktu widzenia egzamin ten pozostawia wiele do życzenia. Bez głębszych zmian nie będzie można traktować tego egzaminu jako rzetelnego, obiektywnego, sprawiedliwego czy choćby w pełni porównywalnego. Wprowadzenie ogólnopolskiej jednolitej matury było ważnym elementem budowy systemu racjonalnej rekrutacji na studia i ewaluacji pracy szkół. Jednak egzamin ten zdecydowanie wymaga kolejnych zmian, które mogłyby zbliżyć go do przyjętych na świecie standardów testowania. Na zakończenie przedstawiamy kilka wniosków i rekomendacji, które przychodzą na myśl po przedstawieniu w tym artykule podstawowych problemów dotyczących egzaminu maturalnego z języka polskiego.

1. W obecnym kształcie matura nie nadaje się do szacowania umiejętności poszczególnych uczniów, jest narzędziem zbyt mało rzetelnym i przemyślanym, generującym zbyt duże błędy w wypadku ocen indywidualnych.
2. W obecnym kształcie matura z języka polskiego z pewnymi zastrzeżeniami pozwala jedynie na ocenę pracy szkół – agregowane miary charakteryzują się bowiem znacznie mniejszym błędem pomiaru, który w trakcie agregacji „znosi się" i pozwala oszacować dosyć dokładne statystyki. Miary te trzeba jedna traktować ostrożnie.
3. Z uwagi na bardzo niską precyzję pomiaru próg zdawalności powinien być zniesiony. Przy dotychczasowym kształcie matury nie pozwala on zadowalająco odróżnić uczniów o niewystarczających kompetencjach z języka polskiego od tych, posiadających kompetencje dostateczne do uzyskania certyfikatu wykształcenia średniego.
4. Jeżeli egzamin maturalny ma przynosić informacje o poziomie osiągnięć indywidualnych uczniów, liczba pytań na egzaminie powinna ulec przynajmniej potrojeniu, a psychometryczne właściwości pytań powinny zostać lepiej dobrane.
5. Gruntownej zmiany wymaga wypracowanie. Temat powinien być jeden lub procedury standaryzacyjne powinny dobierać tematy o takiej samej trudności. Wypracowanie – na wzór egzaminów z USA – powinno zostać skrócone i powinno je sprawdzać przynajmniej dwóch niezależnych egzaminatorów.
6. Powinna zostać wprowadzona procedura skalowania wyników, która umożliwiałaby kontrolę jakości końcowych miar.
7. W egzaminie maturalnym powinny funkcjonować takie mechanizmy, które umożliwiłyby zrównywanie wyników rok do roku (por. Kohen i Brennan 2004).

Bibliografia:

1. Brenna R.R. *Generalizability Theory*, Springer, Nowy Jork 2001.
2. Dolata R., Putkiewicz E., Wiłkomirska A., *Reforma egzaminu maturalnego – oceny i rekomendacje*, Instytut Spraw Publicznych, Warszawa 2004.
3. Jakubowski M. i Pokropek A., *Badając egzaminy*, CKE, Warszawa 2009.
4. Kolen M.J. i Brennan R.L., *Test Equating, Scaling and Linking Methods and Practices Second Edition*. Springer, Nowy Jork 2004.
5. Scheerens J., Glas C.A.W., Thomas S.M., *Educational evaluation, assessment, and monitoring: a systemic approach*. Swets and Zeitlinger, Lisse 2003.
6. *Wright,* B.D. *i Masters G.N., Rating Scale Analysis. MESA Press, Chicago 1982.*