Artur Pokropek
Uniwersytet Warszawski

Metody obliczania edukacyjnej wartości dodanej dla szkół kończących się egzaminem maturalnym

Wstęp

Wskaźnik edukacyjnej wartości dodanej (EWD) obliczany dla szkół kończących się egzaminem maturalnym, pod względem formalnym (konstrukcja wskaźnika, modelowanie zależności, typy modeli regresyjnych i estymacji) nie musi zasadniczo odbiegać od konstrukcji wskaźnika EWD stosowanego w Polsce dla szkół gimnazjalnych.[1] Tym, co wybitnie odróżnia EWD na poziomie szkoły maturalnej od EWD na poziomie gimnazjum, jest strukturalna różnorodność i złożoność samego egzaminu maturalnego, który – jako podstawa służąca do konstrukcji wskaźników efektywności pracy szkół – stawia szereg nowych problemów. Trzy cechy egzaminu maturalnego, w kontekście prac nad budową wskaźnika edukacyjnej wartości dodanej, można uznać za kluczowe. Są to:

1. Duża ilość i różnorodność egzaminów (na różnych poziomach trudności).
2. Możliwość swobodnego wyboru przedmiotów (oraz ich poziomów trudności) zdawanych na maturze.
3. Różne typy szkół prowadzące do tego samego egzaminu.

Pierwsza z poruszonych kwestii jest najtrudniejsza i na niej skupimy uwagę. Problem swobodnego wyboru egzaminu łączy się ścisłe z problemem ilości egzaminów i będzie rozważany w kontekście tego pierwszego. Kwestia trzecia jest niezwykle ważna, niemniej jednak w tej pracy będzie poruszona jedynie marginalnie.

I.

Na maturze uczeń może zdawać 17 przedmiotów (nie licząc języków mniejszości narodowych i języka regionalnego) na dwóch poziomach trudności, co *de facto* daje niebagatelną liczbę 34 egzaminów. Przy takiej liczbie możliwe do zrealizowania wydają się trzy strategie konstruowania wskaźnika efektywności pracy szkoły:

[1] Czyli wskaźnikiem EWD opartym na resztach regresji (patrz: R. Dolata (red.), *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie wyników egzaminów zewnętrznych,* CKE, Warszawa 2007.) lub efektach stałych regresji wielopoziomowej (patrz: S. W. Raudenbush i A. S. Bryk, *Hierarchical Linear Models,* Sage, Thousand Oaks–London–New Delhi 2002.).

1. Stworzenie 34 oddzielnych wskaźników EWD (lub dla mniejszej ilości).
2. Stworzenie jednego, sumarycznego, wskaźnika EWD na podstawie kilkudziesięciu egzaminów.
3. Stworzenie kilku wskaźników zdających relację z kilku bloków przedmiotów (np. przedmioty humanistyczne, ścisłe, artystyczne ...).

Gdyby przyjąć rozwiązanie (1), to pod względem formalnym wskaźnik EWD mógłby być tożsamy z EWD obliczanym dla gimnazjum, ale w tym wypadku zmienną zależną nie byłby egzamin gimnazjalny (w części matematyczno-przyrodniczej lub humanistycznej), lecz konkretny egzamin maturalny, a wynikiem wejściowym nie wynik sprawdzianu po szkole podstawowej, lecz wynik egzaminu gimnazjalnego. Taka strategia ma poważne mankamenty. Spójrzmy na Tabelę 1., która zawiera 17 najczęściej wybieranych egzaminów (w 2007 roku). Przedstawione zostały w niej statystyki opisowe, określające alokacje uczniów zdających dany egzamin maturalny w szkołach kończących się maturą – średnia liczebność szkół (średnia), mediana (med.) oraz łączna liczebność uczniów (suma).[2] Zakładając, iż dla oszacowania rocznego EWD dla szkoły wymagane jest minimum 20 uczniów zdających dany przedmiot – aby zapewnić przynajmniej zadowalającą jakość oszacowań – informacja zwrotna z poszczególnych cząstkowych wskaźników nie dotarłaby do większości szkół. Do więcej niż połowy szkół w Polsce dotarłaby informacja dotyczącą jedynie EWD skonstruowanego na podstawie wyników egzaminu maturalnego z j. polskiego i j. angielskiego, na poziomie podstawowym. Wszystkie inne wskaźniki EWD mogłyby być udostępniane mniej niż połowie szkół.

Odwołując się do dodatkowych wyników (poza Tabelą 1.), można dodać, że 25% szkół w Polsce mogłoby liczyć również na wskaźniki EWD dla geografii na poziomie podstawowym oraz biologii na poziomie rozszerzonym. Gdybyśmy nieznacznie osłabili założenie tak, że EWD liczona byłaby dla szkół, w których zdawało co najmniej 19 uczniów, wskaźnik można by zaproponować również 25% szkół, których uczniowie zdawali język angielski, matematykę oraz chemię na poziomie rozszerzonym. Co gorsze, nawet jeżeli wskaźnik liczony byłby dla danych z trzech lat, nie poprawiłoby to znacząco sytuacji – nawet jeżeli pomnożymy wyniki przedstawione w tabeli razy trzy, ilość uczniów w szkole przypadająca na egzamin będzie nadal niezadowalająca.Sytuacja taka wydaje się nie do zaakceptowania, gdy myślimy o zastosowaniu EWD na szeroką skalę. Choć w określonych przypadkach może być zasadna. Informacja o wartości dodanej dla poszczególnych przedmiotów w przypadku dużych szkół, lub szkół ukierunkowanych na konkretny typ kształcenia, może okazać się informacją pożądaną i służącą pogłębionej ewaluacji wewnątrzszkolnej.

[2] Wartości statystyk opisowych dla poszczególnych przedmiotów liczone są dla populacji szkół, w których przynajmniej jedna osoba zdawała dany przedmiot.

Tabela 1. Liczba uczniów zdających egzamin maturalny przypadająca na szkołę. Statystyki opisowe dla 17 egzaminów maturalnych, na dwóch poziomach trudności; w szkołach, w których przynajmniej jeden uczeń zdawał egzamin z danego przedmiotu, rok 2007

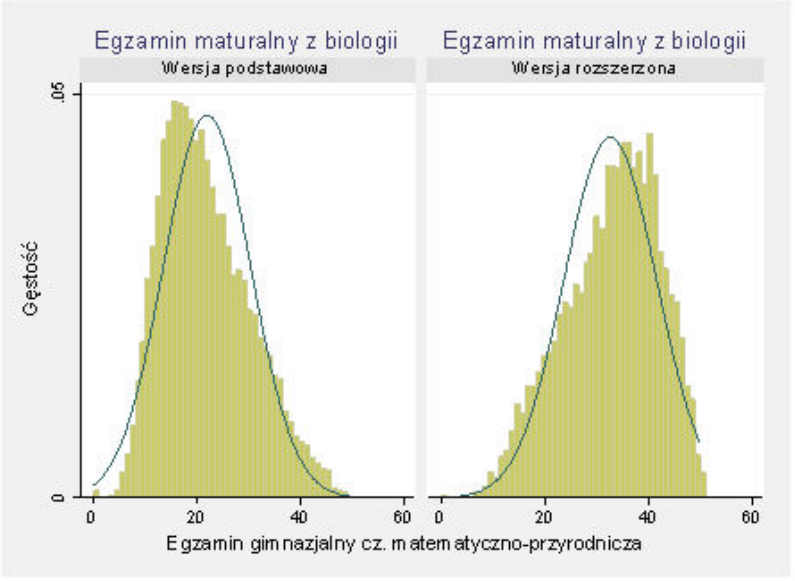
przedmiot egzaminu	poziom podstawowy			poziom rozszerzony		
	średnia	med.	suma	średnia	med.	suma
j. polski	49	29	303428	15	5	44634
j. angielski	37	22	221679	17	4	51131
j. niemiecki	11	7	52206	5	2	6828
j. rosyjski	8	5	14735	3	2	1513
j. francuski	3	2	2638	3	2	1203
j. hiszpański	2	1	245	2	1	204
j. włoski	2	1	234	2	1	174
historia	5	3	18596	11	3	32220
biologia	10	6	47826	15	4	44955
geografia	15	10	81845	11	4	38992
matematyka	8	4	32598	15	4	39952
chemia	3	2	4498	13	5	26826
fizyka	3	2	2839	11	4	20430
wos	9	6	43534	14	5	51423
historia muzyki	3	1	460	2	1	319
historia sztuki	2	1	857	3	2	3086
wiedza o tańcu	2	1	25	2	1	43

Naturalna w tym przypadku wydaje się próba zwiększenia ilości egzaminowanych z danego przedmiotu poprzez zredukowanie dwóch wymiarów trudności – poziomu rozszerzonego i poziomu podstawowego. Zakładamy tutaj, że dany egzamin na dwóch różnych poziomach mierzy tę samą cechę. Jeden z nich na zagadnieniach łatwiejszych, drugi na trudniejszych. Zatem wystarczy „zrównać", czy może precyzyjniej „zestawić", dwa testy względem ich trudności (nie *equating* lecz *linking* – według terminologii anglosaskiej). Chodzi o znalezienie takiej relacji, która doprowadziłaby do porównywalności dwóch testów. Najprostszym sposobem do zestawienia (zrównania) jest metoda przekształcenia liniowego lub ekwipercentylowego[3].

Proste zrównanie ekwipercentylowe lub liniowe nie może wchodzić w tym przypadku w grę, ponieważ decyzja o pisaniu trudniejszej lub łatwiejszej wersji egzaminu należała do uczniów i jest oczywistym, iż owa decyzja nie jest przypadkowa, co łamie założenia prostego modelu zrównywania. Rysunek 1. przedstawia

[3] M. J. Kolen i R. L. Brennan, *Test Equating, Scaling and Linking Methods and Practices Second Edition,* Springer, New York 2004.

histogram wyników egzaminu gimnazjalnego z 2004 roku dla maturzystów zdających egzamin z biologii na poziomie podstawowym oraz rozszerzonym w 2007 roku. Widać wyraźnie, że są to zupełnie dwie grupy uczniów. Możemy przypuszczać, że do egzaminu z biologii w wersji podstawowej przystąpili po prostu uczniowie słabsi, a do egzaminu w wersji rozszerzonej uczniowie lepsi (podobny obraz, aczkolwiek mniej wyraźny, otrzymujemy dla części humanistycznej). Zestawianie dwóch wyników dla tak różnych populacji jest zadaniem niezwykle trudnym.

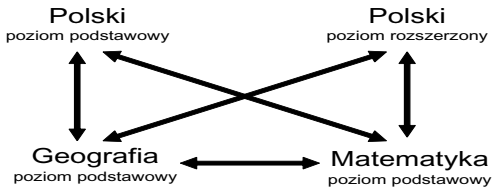


Rysunek 1. Histogramy wyników egzaminu gimnazjalnego, cz. matematyczno-przyrodnicza (2004), dla populacji maturzystów (2007) zdających egzamin na różnych poziomach trudności

W pewnych sytuacjach zrównywanie (zestawianie) wyników egzaminów dla tak różnorodnych grup jest stosunkowo łatwe, jednak sytuacja egzaminu maturalnego jest szczególnie trudna. Dwie wersje danego przedmiotu ustalone ze względu na jego trudność nie posiadają pytań wspólnych, po których dzięki skalowaniu za pomocą pytań kotwiczących można by dokonać procedury zestawiania (zrównania) egzaminów. Co gorsza, nie ma również uczniów, którzy zdawaliby obie wersje egzaminu, co również umożliwiłoby zrównanie (zestawienie) wyników dzięki skalowaniu. Na zakotwiczenie wyników maturalnych w wynikach egzaminu gimnazjalnego nie możemy sobie pozwolić, gdyż czas trzech lat jest długim okresem wypełnianym przez pracę szkół maturalnych. Operacja sprowadzania wyników do skali sprzed trzech lat prowadziłaby nieuchronnie do utraty informacji o wartości dodanej, a na to nie możemy sobie pozwolić.

II.

Strategia zestawiania dwóch poziomów trudności w takim kształcie egzaminu maturalnego, jaki obecnie funkcjonuje, jest trudna, lecz nie jest niemożliwa. Uczniowie zdają przecież wiele przedmiotów.



Rysunek 2. Hipotetyczne pary przedmiotów zdawanych na maturze przez część populacji uczniów

Na Rysunku 1. przedstawiony został przykład porównań. Strzałka wskazuje na to, że istnieje pewna grupa uczniów, która zdawała dany przedmiot, brak strzałki mówi, że taka grupa nie istnieje. Na rysunku nie ma, co prawda uczniów zdających język polski na dwóch poziomach, są natomiast uczniowie, którzy zdawali matematykę (poziom podstawowy) i język polski (poziom podstawowy) oraz są tacy uczniowie, którzy zdawali język polski (poziom rozszerzony) i matematykę (poziom podstawowy). Za pośrednictwem matematyki możemy ustalić relatywną trudność przedmiotów z dwóch poziomów trudności, a tym samym zestawić (zrównać) ich wyniki. Trzeba również zauważyć, że mamy dodatkowe informacje – inne przedmioty, np. na Rysunku 1. dodatkowo mamy geografię na poziomie podstawowym. Do porównań możemy wykorzystać, również uczniów zdających język polski (poziom podstawowy) i geografię (poziom rozszerzony) oraz korygować określanie relatywnych trudności dzięki, zdającym zarówno geografię (poziom podstawowy) i matematykę (poziom podstawowy).

Używając wiedzy o istniejących parach przedmiotów, można zestawić (zrównać) nie tylko wyniki egzaminu maturalnego z różnych poziomów trudności, lecz również z różnych egzaminów.

Strategia zestawiania (zrównywania) jedynie poziomów trudności wydaje się mało efektywna. Połączenie dwóch poziomów trudności w zdecydowanej większości szkół nadal nie będzie dawało odpowiedniej liczebności uczniów, która gwarantowałaby zadowalającą precyzję oszacowań. Racjonalnym w tym wypadku wydaje się zrównanie większej ilości egzaminów poprzez określenie ich relatywnej trudności.
Istnieje kilka sposobów określania relatywnej trudności egzaminów: proste porównanie parami, metoda Kelly'ego, gdzie nie porównuje się w jedynie par przedmiotów, lecz uwzględnia się średnią wszystkich innych zdawanych egzaminów (model stosowany w Szkocji), metody oparte na analizie wariancji ANOVA, jak AMS (*average marks scaling*) stosowana w Australii. Ostatnio jednak szczególną uwagę zwraca się w kierunku zrównywania za pomocą modelów IRT (*Item Response Theory*), które

z kilku względów wydają się najodpowiedniejszym rozwiązaniem.[4]
Klasyczne modele IRT[5] stworzone zostały do analizowania zmiennych binarnych (zerojedynkowych). Ogólny model IRT można zdefiniować jako:

$$P(x_{im} = 1 \mid z_m) = c_i + (1 - c_i)g[a_i(z_m - B_i)] \quad (1)$$

gdzie lewa część równania oznacza prawdopodobieństwo wystąpienia *i*-tego zdarzenia dla *m*-tego badanego, np. odpowiedzenia poprawnie na pytanie testowe przy określonej indywidualnej cesze z_m (umiejętności) dla *m*-tego respondenta. Po prawej stronie równania: c_i oznacza parametr określający możliwość „zgadywania" odpowiedzi, a_i oznacza moc dyskryminacyjną pytania, a B_i jego trudność, *g* natomiast funkcję łączącą (zazwyczaj logit albo probit). Gdy w modelu estymowany jest tylko jeden parametr B_i, mamy do czynienia z modelem Rascha (dla każdego *i*: a_i=1 oraz c_i=0), gdy a_i pozostaje stałe dla wszystkich pytań, lecz jego wartość estymujemy mamy do czynienia z modelem jednoparametrycznym (1PL). Gdy dodatkowo szacowane są parametry a_i (dla każdego pytania), mówimy, że model jest dwuparametryczny (2PL), a jeżeli estymowane są parametry c_i, model określamy jako trójparametryczny (3PL). Przedstawiony model można estymować (szczególnie chodzi tutaj o parametry B_i) w taki sposób, by brane były pod uwagę wszystkie wyniki poszczególnych pytań, nie zaś całych egzaminów (uwzględniając trudności innych pytań), które nie są przecież oceniane w skali 0 – 1. Po estymacji trudności wszystkich pytań zadanych na maturze w danym roku można stworzyć jedną skalę, na której wyniki z poszczególnych przedmiotów będą porównywalne. A tym samym poprzez odpowiednie ich przeważenie można stworzyć jeden sumaryczny wynik dla szkoły określający coś, co konceptualnie moglibyśmy nazwać „umiejętnościami szkolnymi". Można tworzyć też kilka wskaźników, np. „umiejętności humanistyczne" i „matematyczno-przyrodnicze", etc.

Estymacja modelu dla poszczególnych pytań jest jednak dość skomplikowana, wymaga dostępności szczegółowych danych oraz ogromnego zaplecza technicznego. Alternatywą w tym wypadku jest użycie modelu GRM (*Granded Response Model*) zaproponowanego już w 1969 roku przez Samejima. Model ten jest analogiczny do klasycznego IRT z tym, że przeznaczony jest dla zmiennych porządkowych o większej ilości kategorii. Model można sformułować następująco:

$$P(x_{im} = k \mid z_m) = g(n_{ik}) - g(n_{i,k+1}) \quad (2)$$

[4] R. Cole (red.), *Relative difficulty of examinations in different subjects,* CEM Centre, Durham University 2008.

[5] A. Skrondal i S. Rabe-Hesketh, *Generalized Latent Variable Modeling*, A CRC Press Company, Boca Raton–London–New York–Washington, D.C. 2004.

gdzie:

$$n_{ik} = a_i(z_m - B_{ik}), \quad dla \quad k = 1, \ldots, K_i, \tag{3}$$

Lewa strona równania (2) oznacza prawdopodobieństwo odpowiedzi o kategorii k na dane pytanie x_i (dla zmiennej porządkowej), po prawej stronie równania mamy różnicę funkcji dla dwóch kolejnych kategorii odpowiedzi, gdzie n_{ik} określone jest w równaniu (3), a_i oznacza parametr dyskryminacyjny pytania i, z_m cechę ukrytą respondenta m, a B_{ik} „trudność" k-tej kategorii dla i-tego pytania. Podobnie jak w przypadku poprzednich modeli, gdy estymujemy jedynie parametr B (dla każdego a_i=1) – uzyskujemy model analogiczny do modelu Rascha, możemy również uzyskać analogiczny do tego modelu model jednoparametryczny (1PL) gdy a jest estymowane, lecz pozostaje stałe dla wszystkich pytań; natomiast gdy estymujemy dwa parametry – otrzymujemy model dwuparametryczny (2PL).

Kategoria k z równania (2) może być określona jako punkt z egzaminu maturalnego lub przedział punktowy z danego egzaminu. Właśnie tak zostały estymowane modele dla polskich danych (prezentowane w tym artykule), gdzie wynik każdego egzaminu został podzielony na 5 równych przedziałów (gdzie pierwszą kategorią jest 20% najgorszych wyników z każdego przedmiotu, a ostatnią 20% najlepszych wyników).[6]

Na Rysunku 3. przedstawione zostały wyniki oszacowań trudności egzaminów, czyli średnia parametrów B_{ik} dla każdego i-tego egzaminu. Na rysunku przedstawione zostały wyniki dwóch modeli – jednoparametrycznego (1PL) oraz dwuparametrycznego (2PL). Z modeli wyłączone zostały egzaminy, do których przystępowało mniej niż 350 osób – wyniki estymacji dla tych egzaminów okazywały się szczególnie niestabilne. Jak widać na Rysunku 3., potoczne oczekiwania pokrywają się z wynikami modelowania. Wszystkie egzaminy z poziomów podstawowych są łatwiejsze od tych z wersji rozszerzonej. Do najtrudniejszych przedmiotów można zaliczyć rozszerzony egzamin z języka francuskiego, fizyki, matematyki. Najłatwiejsze egzaminy to: rosyjski, geografia i wiedza o społeczeństwie (oczywiście na poziomie podstawowym). Ponadto wyniki estymacji dla dwóch różnych modeli są niezwykle podobne. Wyraźne różnice widać tylko w przypadku przedmiotów, których trudność estymowana była na małej ilości zdających (np. historia sztuki).

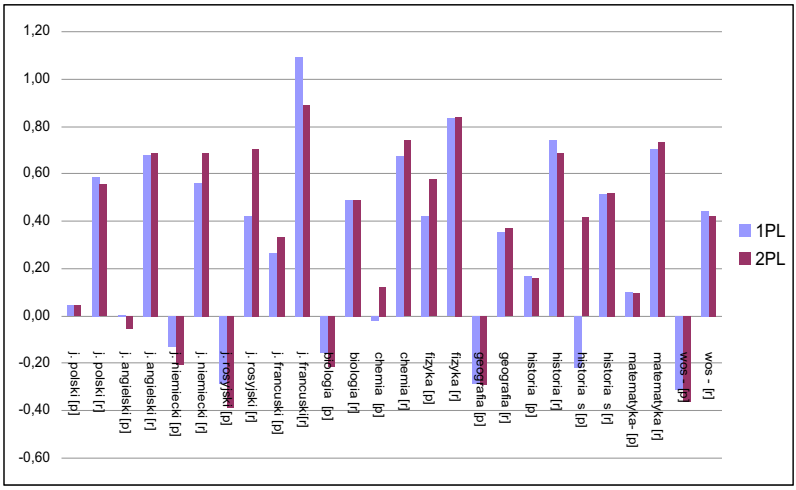
Na podstawie prezentowanych wyników można stworzyć jeden lub kilka wskaźników na podstawie maturalnych wyników egzaminacyjnych, traktując prezentowane wyniki jako swoiste wagi w przekształceniach surowych wyników.

[6] Zrównywanie wyników dla polskich danych wciąż jest w fazie testów, opisywany tutaj podział nie jest podziałem docelowym, a raczej wybranym z powodów technicznych – możliwości obliczeniowych komputerów.

III.

Jak pokazałem, możliwe jest stworzenie jednego lub kilku wskaźników będących podstawą do liczenia EWD dla szkół maturalnych. Pozostaje pytanie, ile i jakie wskaźniki konstruować. Innymi słowy, jaką wartość dodaną liczyć dla szkół maturalnych. Czy ogólną wartość dodaną – mierzącą przyrost „umiejętności szkolnych"? Czy wskaźniki bardziej precyzyjne, np. wartość dodaną wiedzy humanistycznej (opartej na wskaźniku stworzonym z przedmiotów humanistycznych, takich jak język polski, historia, historia sztuki, etc.) oraz matematyczno-przyrodniczej (opartą na matematyce, fizyce, chemii, etc.)?

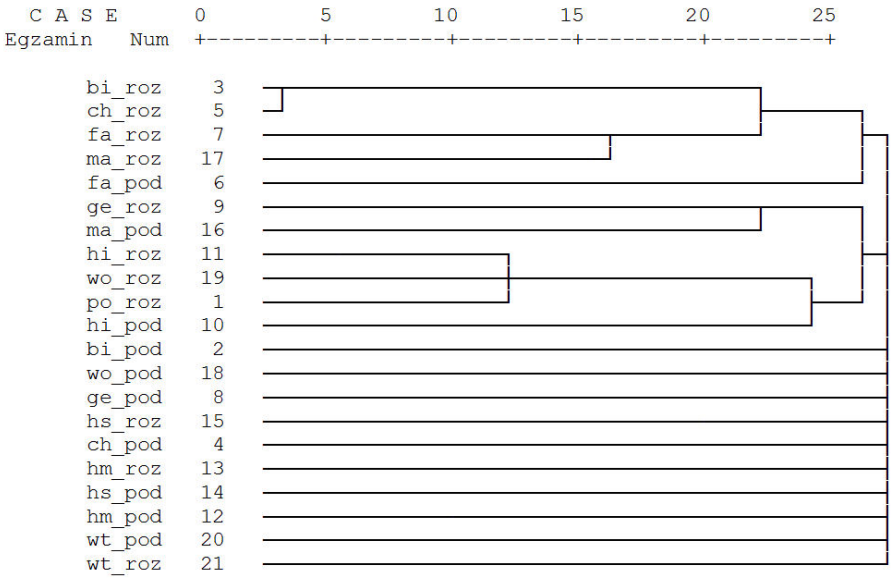
Jeżeli chcemy stworzyć kilka wskaźników efektywności opartych na różnych egzaminach, to egzaminy te powinny być do siebie podobne. Jednym ze sposobów patrzenia na podobieństwo jest odwołanie się do wyborów uczniów. Jeżeli wybory egzaminów układają się w konkretny wzór – są uczniowie nastawieni na przedmioty humanistyczne i takie wybierają na maturze, są również maturzyści wybitnie nastawieni na przedmioty przyrodnicze i wybierają tylko przedmioty przyrodnicze. Wtedy możemy mówić o podobieństwie.



Rysunek 3. Relatywne trudności egzaminów maturalnych – wartości w skali logitowej; wartości dla modelu GRM jednoparametrycznego (1PL) i dwuparametrycznego (2PL); [p] – oznacza poziom podstawowy, [r] rozszerzony przedmiotów zdawanych na egzaminie maturalnym

Jeżeli struktura zdawania jest spójna i krystalizuje kilka typów przedmiotów, należy wówczas zwrócić się ku koncepcji stworzenia kilku wskaźników.

Próbę uchwycenia wzorów zdawania podejmiemy za pomocą hierarchicznej analizy skupień. Analiza taka nie jest pozbawiona mankamentów (np. uzyskana klasyfikacja nie musi być klasyfikacją najlepszą, a jedynie lokalnie najlepszą, poza tym kształt uzyskanej klasyfikacji silnie uzależniony jest od – niekiedy arbitralnych – decyzji przy wyborze odległości czy centrów skupień), niemniej jednak może być ona pomocna jako narzędzie ułatwiające wgląd w analizowane dane. Wynik analizy skupień, gdzie analizowane były wybory danych egzaminów (zdawał – nie zdawał), przedstawia Rysunek 4. Zastosowana tu została binarna metoda hierarchizacji, dystansów Jaccarda, przy algorytmie maksymalizującym odległości międzygrupowe (najdalszego sąsiada).



Rysunek 4. Hierarchiczna analiza skupień – wybory egzaminów (bi – biologia, ch – chemia, fa – fizyka, ma – matematyka, ge – geografia, wo – wiedza o społeczeństwie, po – język polski, hm– historia muzyki, hs – historia sztuki, wt – wiedza o tańcu, pod – poziom podstawowy, roz – poziom rozszerzony)

Z lewej strony rysunku mamy listę przedmiotów (z wyłączeniem języka polskiego na poziomie podstawowym, który był zdawany przez wszystkich uczniów oraz języków obcych – gdyż wymuszony jest tu wybór przynajmniej jednego). Linie prowadzące od wyszczególnionych przedmiotów można traktować jako graficzne reprezentacje odległości między przedmiotami. Dlatego krótkie linie świadczyć będą o podobieństwie wyborów, długie zaś o jego braku. Interpretując wykres, możemy powiedzieć, że maturzyści, którzy zdawali biologię na poziomie

rozszerzonym, najczęściej jako kolejny przedmiot wybierali chemię na poziomie rozszerzonym, do tych przedmiotów dołączają również fizyka i matematyka (poziom rozszerzony). Wyraźne skupienie tworzą też przedmioty takie jak historia, wiedza o społeczeństwie i j. polski na poziomie rozszerzonym, do których dołącza historia na poziomie podstawowym oraz, co jest pewnym zaskoczeniem, dwa przedmioty – geografia na poziomie rozszerzonym oraz matematyka na poziomie podstawowym. Jest to prawdopodobnie spowodowane specyficznymi wymaganiami niektórych uczelni.

Żadnej struktury nie pokazują natomiast większość przedmiotów z poziomu podstawowego. Nie powoduje to ogromnego zdziwienia, ta część przedmiotów wybierana jest prawdopodobnie ze względu na subiektywną ocenę trudności przez uczniów zdających maturę, którzy nie mają planów łączących się z konkretnym kierunkiem studiów. Istnieje zatem potwierdzenie, choć nie pełne, dla koncepcji budowania przynajmniej dwóch wskaźników edukacyjnej wartości dodanej – humanistycznej oraz matematyczno przyrodniczej. Dodatkowo zdroworozsądkowym podejściem wydaje się stworzenie również osobnego wskaźnika dla języków obcych, który w sposób oczywisty mierzyć będzie efektywność nauczania dla wyodrębniającej się grupy przedmiotów.

Wnioski

Istnieje kilka możliwych kierunków tworzenia wskaźników edukacyjnej wartości dodanej dla szkół maturalnych. Żaden z nich nie jest pozbawiony wad, o których ledwo wspomnieliśmy, jednak większość podejść, przynajmniej na poziomie konceptualnym, wydaje się, z różnych perspektyw i dla różnych potrzeb, niezwykle użyteczne. W przypadku szkoły maturalnej równolegle powinno powstać kilka modeli edukacyjnej wartości dodanej:

1. Model ogólny – mierzący wartość dodaną dla „ ogólnych umiejętności szkolnych" stworzony na podstawie wszystkich egzaminów maturalnych.
2. Modele cząstkowe dla bloków przedmiotów: humanistycznego matematyczno-przyrodniczego i bloku języków obcych.
3. Modele cząstkowe dla poszczególnych przedmiotów na różnych poziomach.
4. Modele cząstkowe dla poszczególnych przedmiotów po zestawieniu (zrównaniu) wyników dla różnych poziomów trudności.

Modele z pierwszych dwóch punktów powinny mieć szerokie zastosowanie i być w przyszłości dostępne dla dyrekcji, nauczycieli, rodziców, uczniów i władz samorządowych. Modele z dwóch ostatnich punktów powinny służyć jedynie jako narzędzia ewaluacji wewnątrzszkolnej – dla szkół i przedmiotów, które osiągną wystarczającą dla stabilnego oszacowania wskaźnika liczebność zdających maturzystów. Na koniec warto jeszcze wspomnieć, że modele powinny być szacowane osobno dla różnych typów szkół prowadzących do egzaminu maturalnego, na co wskazuje specyfika pracy tych szkół (uniemożliwiająca bezpośrednie porównywanie efektywności ich pracy).

Bibliografia:

1. Cole R. (red.), *Relative difficulty of examinations in different subjects,* CEM Centre, Durham University 2008.
2. Dolata R. (red.), *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie wyników egzaminów zewnętrznych,* CKE, Warszawa 2007.
3. Raudenbush S.W. i Bryk A. S., *Hierarchical Linear Models,* Sage, Thousand Oaks–London–New Delhi 2002.
4. Skrondal A. i Rabe-Hesketh S., *Generalized Latent Variable Modeling*, A CRC Press Company, Boca Raton–London–New York–Washington, D.C. 2004.
5. Kolen M. J. i Brennan R. L., *Test Equating, Scaling and Linking Methods and Practices Second Edition,* Springer, New York 2004.