KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

IBE EDUKACYJNA WARTOŚĆ DODANA

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Analizy IBE/02/2013

Statystyczne modelowanie wskaźników edukacyjnej wartości dodanej – podsumowanie polskich doświadczeń

Autor:
Tomasz Żółtak

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

Publikacja opracowana w ramach projektu systemowego*: Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej,* współfinansowanego przez Unię Europejską ze środków Europejskiego Funduszu Społecznego, realizowanego przez Instytut Badań Edukacyjnych.

Egzemplarz bezpłatny

Spis Treści

1. Wprowadzenie

Raport ten ma na celu opisanie doświadczeń związanych z rozwojem metodologii wyliczania polskich wskaźników edukacyjnej wartości dodanej (EWD), w zakresie używanych do tego celu technik modelowania statystycznego. Idea wskaźników EWD opiera się na porównywaniu ze sobą wyników dwóch (lub więcej) egzaminów w celu określenia względnego postępu poszczególnych uczniów. Wskaźniki te mogą być traktowane jako szczególny sposób komunikowania wyników egzaminacyjnych dla grup uczniów, w odniesieniu do ich wcześniejszych osiągnięć (Dolata, 2007b).

Niniejszy tekst ma charakter w dużej mierze techniczny i został napisany z założeniem, że czytelnik posiada już ogólną orientację co do sposobu, w jaki konstruowane są wskaźniki EWD, rodzajów wyliczanych w Polsce wskaźników EWD oraz co do kształtu polskiego systemu egzaminów zewnętrznych. W razie konieczności uzupełnienia tej wiedzy cenne źródła informacji stanowić mogą raporty z badań nad trafnością polskich wskaźników EWD w gimnazjach (Dolata i in., 2013) i szkołach ponadgimnazjalnych (Karwowski, 2013), a także książka podsumowująca doświadczenia zebrane w początkowym okresie prac nad rozwojem wskaźników EWD w naszym kraju (Dolata, 2007a). Można tam również znaleźć krótkie omówienie podobnych rozwiązań stosowanych za granicą i ogólniejszych problemach metodologicznych związanych z wyliczaniem wskaźników EWD (Żółtak, 2013a).

Wskaźniki EWD wyliczane są w Polsce dla gimnazjów oraz dla szkół kończących się maturą – liceów ogólnokształcących i techników. Przy tym w przypadku techników należy pamiętać, że mierzą one wyłącznie z efektywności pracy szkoły w zakresie nauczania wiedzy ogólnej (przygotowania do matury). Ze względu na liczbę roczników absolwentów uwzględnianych przy wyliczaniu wskaźników wyróżniane są wskaźniki jednoroczne i trzyletnie. Te pierwsze z założenia mają być wykorzystywane przede wszystkim do ewaluacji wewnątrzszkolnej. Wskaźniki te nie są powszechnie dostępne. Parametry modelu EWD wyestymowanego na danych ogólnopolskich, niezbędne do wyliczania wartości wskaźników, zapisywane są w programie komputerowym działającym jako samodzielna aplikacja, która udostępniana jest użytkownikom. Po wczytaniu do tej aplikacji danych egzaminacyjnych możliwe jest prowadzenie różnorodnych analiz, w szczególności wyliczanie wskaźników EWD dla dowolnie zdefiniowanych grup uczniów, w zależności od potrzeb użytkownika. Z kolei wskaźniki trzyletnie skierowane są do znacznie szerszego grona odbiorców obejmującego nadzór pedagogiczny sprawowany przez kuratoria oświaty, organy prowadzące szkoły, rodziców i udostępniane są publicznie za pośrednictwem ogólnodostępnej strony internetowej. Wyliczane są one wyłącznie na poziomie szkół, a uwzględnienie w ramach pojedynczego wskaźnika informacji o wynikach trzech roczników absolwentów pozwala oceniać szkołę w sposób bardziej całościowy, nie prowokujący do zbyt pochopnych interpretacji.

Raport podzielony został na dwie części. W pierwszej omawiane są kwestie związane z metodami przekształcania wyników egzaminów zewnętrznych w ten sposób, aby były one bardziej użyteczne do wyliczania na ich podstawie wskaźników EWD. W drugiej omówiono kwestie związane z modelowaniem zależności pomiędzy wynikami egzaminu „na wejściu" a wynikami egzaminu „na wyjściu" przy pomocy modeli regresji i wyliczaniem na podstawie takich modeli wskaźników EWD. Raport kończą rekomendacje odnośnie kierunków dalszych prac, zarówno w zakresie ewentualnych zmian dotyczących egzaminów zewnętrznych, jak i samego wyliczania wskaźników EWD.

2. Przekształcanie wyników egzaminów

Idea wyliczania wskaźników EWD opiera się na porównywaniu ze sobą wyników egzaminów. W związku z tym wysoka jakość egzaminów jest podstawowym warunkiem koniecznym do uzyskania dobrych wskaźników EWD. W dotychczasowych publikacjach poświęconych tematyce trafności polskich egzaminów w kontekście wyliczania wskaźników EWD oceniane są one jako posiadające wystarczająco dobre własności (Jasińska i Żółtak, 2013; Pokropek, 2013), choć wskazywanych jest też wiele możliwych pól do poprawy (Grudniewska i Kondratek, 2012; Pokropek, 2012; Skórska, Koniewski i Majkut, 2013). Nie ulega przy tym wątpliwości, że surowe wyniki egzaminów, tj. proste sumy punktów uzyskanych za poszczególne zadania (albo równoważnie odsetki maksymalnej możliwej do uzyskania liczby punktów), nie mogą być bezpośrednio porównywane pomiędzy różnymi edycjami tego samego egzaminu.

Tabela 1 zawiera zestawienie średnich i odchyleń standardowych wyników surowych sprawdzianu oraz części humanistycznej i matematyczno-przyrodniczej egzaminu gimnazjalnego z lat 2005-2011 (zamieszczane w sprawozdaniach Centralnej Komisji Egzaminacyjnej). Przytoczone liczby wskazują, że rozkłady wyników surowych różnią się znacząco między niektórymi latami, zarówno co do średniej, jak i zróżnicowania wyników. Wydaje się nieprawdopodobnym, aby te różnice miały odwzorowywać zróżnicowanie istotnie wsytępujące pomiędzy poszczególnymi rocznikami absolwentów. Dużo bardziej prawdopodobne, że ich źródłem jest po prostu zróżnicowanie własności arkuszy egzaminacyjnych, wykorzystywanych w poszczególnych latach.

Tabela 1. Średnie i odchylenia standardowe wyników surowych sprawdzianu, części humanistycznej i części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2005-2011.
źródło: CKE

	sprawdzian		część. hum. egz. gimn.		część. mat.-przyr. egz. gimn.	
rok	średnia	odch. stand.	średnia	odch. stand.	średnia	odch. stand.
2011	25,27	7,51	25,31	9,34	23,63	9,37
2010	24,56	8,03	30,34	8,38	26,71	9,32
2009	22,64	7,63	31,67	8,70	26,03	11,02
2008	25,80	7,52	30,75	9,84	27,07	10,65
2007	26,60	7,82	31,48	9,78	25,31	10,22
2006	25,32	8,56	31,39	8,39	23,90	10,30
2005	29,50	7,43	33,18	8,71	24.26	10,15

W związku z tym wysoce wskazane wydaje się przekształcanie surowych wyników egzaminów tak, aby uczynić je bardziej porównywalnymi pomiędzy kolejnymi latami, a w przypadku chęci wykorzystania wyników egzaminów do wyliczania trzyletnich wskaźników EWD, uwzględniających trzy kolejne roczniki absolwentów szkoły, jest to wręcz konieczne. Jednocześnie trzeba zaznaczyć, że nie stawiamy tu sobie za cel wyrażenia wyników egzaminów na tej samej skali, pozwalającej śledzić zmiany średniego poziomu i zróżnicowania wyników w skali kraju w kolejnych latach. Choć w ogólności jest to cel możliwy do osiągnięcia, jednak jego realizacja jest bardzo skomplikowana i kosztowna, gdyż wymaga prowadzenia dodatkowych badań zrównujących (Szaleniec i in., 2013).

W kontekście wyliczania wskaźników EWD cel może być dużo skromniejszy – odniesienie wyników uzyskanych na egzaminie przez danego ucznia do wyników wszystkich zdających w tym samym roku. Zmiana takiego relatywnego położenia uczniów w rozkładzie wyników jest wystarczająco dobrą podstawą do wyliczania wskaźników EWD, bowiem z założenia są one miarami względnymi (Żółtak, 2013a). Aby ten cel osiągnąć, konieczne jest jednak odpowiednie przekształcenie surowych wyników egzaminacyjnych.

Z teoretycznego punktu widzenia metody przekształcania wyników egzaminów stosowane w procesie wyliczania wskaźników EWD możemy podzielić na te, które zakładają określony model pomiarowy oraz metody, które nie odwołują się do takich założeń. Pierwsze z nich stanowią modele IRT, reprezentantem drugiej grupy jest zaś metoda normalizacji ekwikwantylowej. Z praktycznego punktu widzenia zasadnicze znaczenie ma jednak inne rozróżnienie, a mianowicie, czy metoda wykorzystuje wyłącznie sumaryczny wynik surowy (równoważnie odsetek zdobytych punktów), czy też wymaga informacji o punktacji uzyskanej za wykonanie poszczególnych zadań egzaminu. W tej pierwszej grupie możemy umieścić normalizację ekwikwantylową oraz model Rascha, w drugiej pozostałe modele IRT. Oczywiście rozróżnienie to ma sens tylko w sytuacji, gdy wszyscy uczniowie rozwiązują ten sam zestaw zadań. Jeśli jest inaczej, metody oparte wyłącznie na sumie nie znajdują zastosowania.

Możliwość posługiwania się wyłącznie wynikiem sumarycznym ma duże znaczenie w przypadku jednorocznych wskaźników EWD[1], które wyliczane są przez użytkowników w zewnętrznej aplikacji – Kalkulatorze EWD - na podstawie samodzielnie wczytanych danych. Metody wymagające informacji o punktacji za poszczególne zadania byłyby bardzo skomplikowane do zaimplementowania w tej aplikacji (konieczność implementacji modułu statystycznego do estymacji oszacowań z danych wejściowych na podstawie zadanych parametrów modelu IRT lub konieczność stosowania olbrzymiej wielkości tablic przeliczeniowych). Z drugiej strony użytkownicy mogliby mieć trudności z pozyskaniem wyników egzaminacyjnych w tej formie, a proces wczytywania ich do Kalkulatora byłby potencjalnym źródłem ogromnej liczby przekłamań i błędów.

W praktyce spośród dwóch wymienionych metod operujących wyłącznie na wyniku sumarycznym – normalizacji ekwikwantylowej i modelu Rascha – w odniesieniu do polskich egzaminów zastosowanie znajduje tylko pierwsza. Model Rascha jest najbardziej restrykcyjną formą modeli IRT – zakłada się w nim, że wyniki każdego zadania w teście są tak samo silnie powiązane z poziomem mierzonej cechy - a egzaminy przygotowywane w naszym kraju, niestety, nie spełniają tego założenia. W związku z tym, do przekształcania wyników egzaminów na potrzeby wyliczania jednorocznych wskaźników EWD wykorzystywana jest od 2012 r. metoda normalizacji ekwikwantylowej. Wcześniej, w latach 2009-2011, była ona stosowana także do przekształcania wyników egzaminów na potrzeby wyliczania trzyletnich wskaźników EWD gimnazjów.

[1] Inne różnice pomiędzy jednorocznymi a trzyletnimi modelami EWD opisane zostały w podrozdziałach 3.2, 3.3 i 3.4.

2.1. Normalizacja ekwikwantylowa[2]

Procedura normalizacji ekwikwantylowej ma na celu takie przekształcenie wartości zmiennej, aby jej rozkład miał własności możliwie zbliżone do założonego rozkładu – w przypadku procedury stosowanej w procesie wyliczania wskaźników EWD rozkładu normalnego standaryzowanego. Jednocześnie następuje standaryzacja wyników, w rezultacie której punktem odniesienia dla skali wyników staje się średni wynik w ramach grupy, dla której przeprowadzana jest normalizacja, a jednostką skali odchylenie standardowe wyników w ramach tej grupy. Następnie poprzez przekształcenie liniowe możliwe jest dowolne ustalenie średniej i odchylenia standardowego skali. Przeliczenie wyników polskich egzaminów na skalę znormalizowaną o średniej 100 i odchyleniu standardowym 15 dokonywane jest tzw. metodą Hazena (Barnett, 1975) na podstawie wzoru:

$$U(X = x_i) = 100 + 15\Phi^{-1}\left(\frac{N(X \leq x_i) - \frac{N(X = x_i)}{2} - 0{,}5}{n}\right) \quad (1)$$

gdzie:

$U\,(X = x_i)$	wynik znormalizowany dla wyniku surowego równego x_i;
Φ^{-1}	funkcja odwrotna do dystrybuanty rozkładu normalnego standaryzowanego;
$N\,(X \leq x_i)$	liczba zdających z wynikiem surowym nie wyższym niż x_i;
$N\,(X = x_i)$	liczba zdających z wynikiem surowym równym x_i;
n	liczba wszystkich zdających.

Warto zaznaczyć, że w przypadku nowej, wprowadzonej w 2012 r. formuły egzaminu gimnazjalnego, znormalizowane wyniki dla dwóch części egzaminu gimnazjalnego (humanistycznej i matematyczno-przyrodniczej) wyliczane są na podstawie rozkładu sumy wyników surowych z odpowiednich dwóch testów, a nie z przekształcenia ich wyników znormalizowanych. W szczególności znormalizowany wynik danej części egzaminu gimnazjalnego nie jest średnią z wyników znormalizowanych dwóch tworzących daną część testów. Zależność pomiędzy znormalizowanymi wynikami testów tworzących daną część i znormalizowanym wynikiem dla tej części jako całości wynika z łącznego rozkładu wyników tych dwóch testów, jest złożona i nie daje się łatwo opisać.

Ponieważ wyniki uzyskiwane zarówno na maturze jak i na egzaminie gimnazjalnym przez uczniów liceów ogólnokształcących są zdecydowanie wyższe, niż wyniki uzyskiwane przez uczniów techników, zdecydowano się oddzielnie wyliczać wskaźniki dla obu typów szkół. Uznano bowiem, że w takiej sytuacji licea nie mogą być dobrym punktem odniesienia dla techników i odwrotnie (oczywiście istotne są tu też różnice w celach kształcenia tych dwóch typów szkół). Aby ułatwić interpretację wyników dla każdego z tych typów szkół, na potrzeby maturalnych modeli EWD znormalizowane wyniki przekształcane są na skalę o średniej 100 i odchyleniu standardowym 15 oddzielnie w ramach grupy uczniów liceów ogólnokształcących i w ramach grupy uczniów techników. Dotyczy to zarówno wyników matury (obecnie tylko z matematyki na poziomie podstawowym), jak i wyników egzaminu gimnazjalnego (przy czym normalizacja wyników egzaminu gimnazjalnego dokonywana jest wcześniej w oparciu o tabele przeliczeniowe wyliczone w przedstawiony powyżej sposób na podstawie danych wykorzystywanych w modelach EWD gimnazjów).

---

[2] Obszerny fragment tego podrozdziału został zaczerpnięty z: Pokropek, A. i Żółtak, T. (2012). Nowe modele jednorocznej EWD. [W:] B. Niemierko i M. K. Szmigel (Red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy*. Kraków: Grupa Tomami.

2.2. Skalowanie z użyciem modeli IRT

W przypadku trzyletnich wskaźników EWD, które wyliczane są centralnie, a następnie publikowane za pośrednictwem strony internetowej możliwe jest posłużenie się bardziej wyrafinowanymi statystycznie metodami przekształcania wyników egzaminacyjnych. W przypadku wskaźników maturalnych jest to wręcz konieczne – wyłącznie wykorzystanie modeli IRT pozwala dobrze poradzić sobie z wyznaczeniem oszacowań poziomu umiejętności „na wyjściu" w sytuacji, gdy różni uczniowie rozwiązywali częściowo inny zestaw zadań, choć fakt nielosowego wybierania przez uczniów zdawanych przedmiotów wymaga wprowadzenia pewnych dodatkowych modyfikacji modeli.

W porównaniu z normalizacją ekwikwantylową zastosowanie modeli IRT ma dwie zalety. Po pierwsze odwołanie do modelu opisującego związek pomiędzy mierzoną cechą a wynikami poszczególnych zadań pozwala na ocenę jakości egzaminu jako całości i poszczególnych zadań. Co prawda w momencie, gdy egzamin został już przeprowadzony możliwości wykorzystania tak zdobytych informacji ograniczają się do usunięcia z modelu skalowania zadań o szczególnie słabych własnościach pomiarowych. Po drugie, oszacowania poziomu umiejętności uzyskiwane z takich modeli mają lepsze własności statystyczne, nawet w sytuacji, gdy rozkłady wyników surowych są bardzo nietypowe (Jasińska i Żółtak, 2013). W szczególności zależności pomiędzy wynikami egzaminów wyskalowanych przy pomocy modeli IRT mają nieco gładszy, bliższy liniowemu przebieg, co jest rzeczą bardzo istotną przy ustalaniu postaci modeli regresji, używanych do wyliczania wskaźników EWD.

Podstawowym zagadnieniem, jakie należy rozpatrzyć, jest wybór formy stosowanego modelu. Jak już wcześniej wspomniano, polskie egzaminy nie spełniają w zadowalający sposób założeń modelu Rascha, więc nie może on zostać wykorzystany. W związku z tym w grupie modeli typowo stosowanych do testów umiejętności należy rozważyć po pierwsze wybór pomiędzy modelem dwuparametrycznym a trzyparametrycznym dla zadań ocenianych binarnie, po drugie zaś wybór pomiędzy modelem *graded response* a modelem *partial credit* dla zadań o kilku możliwych poziomach wykonania (Kondratek i Pokropek, 2013; Linden i Hambelton, 1997).

W trzyparametrycznym logistycznym (3PL) modelu IRT prawdopodobieństwo udzielenia poprawnej odpowiedzi na *k*-te zadanie testowe w zależności od natężenia mierzonej cechy ukrytej θ opisywane jest wzorem:

$$P(X_k = 1) = 1 - \frac{1 - c_k}{1 + exp\left(a_k(\theta - b_k)\right)} \qquad (2)$$

gdzie poszczególne parametry modelu przyjęło się określać jako:
a_k – dyskryminacja – wskazuje ona na siłę związku pomiędzy mierzoną cechą a wynikiem zadania;
b_k – trudność – co nie wymaga komentarza;
c_k – zgadywanie – wyznacza granicę, do której zbiega prawdopodobieństwo udzielenia prawidłowej odpowiedzi gdy θ zbiega do minus nieskończoności.

W dwuparametrycznym modelu logistycznym (2PL) przyjmuje się, że wartość parametru c jest równa zero i nie uwzględnia się go w modelowaniu:

$$P(X_k = 1) = 1 - \frac{1}{1 + exp\left(a_k(\theta - b_k)\right)} \qquad (3)$$

Łatwo przy tym zauważyć, że jest to równoznaczne z przyjęcie założenia, że zależność pomiędzy mierzoną cechą ukrytą a wynikiem odpowiedzi na zadanie testowe opisywana jest modelem regresji logistycznej:

$$\log\left(\frac{P(X_k = 1)}{1 - P(X_k = 1)}\right) = (-a_k b_k) + a_k \theta \qquad (4)$$

Dla zadań ocenianych binarnie model trzyparametryczny zapewnia lepsze dopasowanie do danych. Poza tym wydaje się, że przy dużej uwadze, jaką w Polsce (i nie tylko) przywiązuje się do problemu wpływu ewentualnego zgadywania na wyniki egzaminów (Twardowska i in. 2011), atrakcyjne jest dodanie do modelu parametru pozwalającego uwzględnić występowanie tego zjawiska. Jednak w praktyce wiarygodne szacowanie parametrów zgadywania jest zadaniem dosyć trudnym, szczególnie dla łatwych i bardzo łatwych zadań, których w polskich egzaminach (zwłaszcza sprawdzianie) nie brakuje. Problem polega na tym, że informacji użytecznych dla określenia wartości parametru zgadywania zadania dostarczają uczniowie, których poziom umiejętności jest zdecydowanie niższy od poziomu trudności zadania. Dla zadań łatwych grupa ta jest niewielka, a więc oszacowania tego parametru obarczone są znaczną niepewnością, co nie pozostaje bez wpływu również na dokładność szacowania pozostałych parametrów zadania.

Warto też zauważyć, że same oszacowania punktowe poziomu umiejętności uzyskiwane w wyniku zastosowania do tych samych danych modelu dwuparametrycznego i trzyparametrycznego są do siebie bardzo zbliżone, nawet w sytuacji, gdy istotnie występuje zjawisko zgadywania. Co prawda oba modele mogą dawać bardzo różniące się od siebie oszacowania błędów standardowych dla oszacowań poziomu umiejętności, jednakże z punktu widzenia procedury wyliczania wskaźników EWD fakt ten nie ma większego znaczenia. Dzieje się tak, gdyż w modelach regresji używanych do wyliczania wskaźników EWD informacja o tych błędach standardowych nie jest w żaden sposób wykorzystywana. Można więc stwierdzić, że w ostatecznym rozrachunku – z punktu widzenia skalowania wyników egzaminów na potrzeby wyliczania wskaźników EWD – model trzyparametryczny nie ma wyraźnych przewag nad modelem dwuparametrycznym.

Jeśli chodzi o modele dla zadań o kilku możliwych poziomach wykonania, to dla *k*-tego zadania testowego o liczbie różnych możliwych poziomów wykonania równej m_k (aby być w zgodzie z konwencją punktowania zadań na egzaminach przyjmijmy, że poziomy numerowane są począwszy od zera: $0, 1, \ldots, m_k - 1$) prawdopodobieństwo, że uczeń osiągnie co najmniej *g*-ty poziom wykonania jest w modelu *graded response* opisywane wzorem:

$$P(X_k \geq g) = 1 - \frac{1}{1 + exp\left(a_k\left(\theta - b_{kg}\right)\right)} \qquad (5)$$

Prawdopodobieństwo osiągnięcia dokładnie *g*-tego poziomu wykonania można zaś obliczyć jako:

$$P(X_k = g) = P(X_k \geq g) - P\left(X_k \geq (g + 1)\right) \qquad (6)$$

przy czym przyjmuje się, że:

$$P(X_k \geq 0) = 1 \qquad (7)$$
$$P(X_k \geq m_k) = 0 \qquad (8)$$

Podobnie jak w przypadku dwuparametrycznego model logistycznego jest to równoważne przyjęciu założenia, że zależność pomiędzy mierzoną cechą ukrytą a wynikiem odpowiedzi na zadanie testowe opisywana jest modelem regresji logistycznej, z tym że w tym przypadku jest to jej odmiana - wielowartościowa regresja logistyczna dla zmiennej zależnej mierzonej na skali porządkowej:

$$\log\left(\frac{P(X_k \geq g)}{1 - P(X_k \geq g)}\right) = \left(-a_k b_{kg}\right) + a_k\theta \qquad (9)$$

W modelu *partial credit* prawdopodobieństwo osiągnięcia dokładnie *g*-tego poziomu wykonania opisywane jest z kolei wzorem:

dla $g > 0$:

$$P(X_k = g) = \frac{exp\left(\sum_{l=1}^{g} a_k(\theta - b_{kl})\right)}{1 + \sum_{l=1}^{m_k-1} exp \sum_{j=1}^{l}\left[a_i(\theta - b_{kj})\right]} \qquad (10)$$

dla $g = 0$:

$$P(X_k = 0) = \frac{1}{1 + \sum_{l=1}^{m_k-1} exp \sum_{j=1}^{l}\left[a_i(\theta - b_{kj})\right]} = 1 - \sum_{g=1}^{m_k-1} P(X_k = g) \qquad (11)$$

Współczynniki b_{kg} z modlu *graded response* i modelu *partial credit* posiadają analogiczną interpretację i wskazują wartości *θ*, w których prawdopodobieństwo uzyskania *g*-tego poziomu wykonania zrównuje się z prawdopodobieństwem uzyskania poziomu (*g*+1)-tego. Jednocześnie w modelu *partial credit* możliwe jest, że $b_{k(g+1)} < b_{kg}$. Jeśli taka sytuacja ma miejsce, określa sią ją mianem zaburzeniem kolejności poziomów wykonania ze względu na trudność. Oznacza to też, że nie istnieje taki przedział wartości *θ*, dla którego uzyskanie (*g*+1)-tego poziom wykonania byłoby bardziej prawdopodobne niż każdego innego poziomu wykonania. Jest to sytuacja niepożądana przy konstruowaniu, wskazująca, że należałoby skorygować schemat oceniania zadania tak, by był on bardziej efektywny. W modelu *graded response* nie dopuszcza się występowania takiej sytuacji. Ze sposobu parametryzacji modelu wynika, że zawsze $b_{k(g+1)} \geq b_{kg}$. W związku z tą własnością wykorzystywanie modelu *partial credit* jest wskazane na etapie tworzenia i pilotażu arkuszy testowych, jednak w odniesieniu do dopracowanych testów egzaminacyjnych nie ma przewagi nad modelem *graded resonse*. Istotną cechą tego drugiego jest z kolei ścisłe pokrewieństwo formalne z modelami strukturalnymi dla porządkowych zmiennych mierzalnych (CSEM), estymowanymi na podstawie macierzy korelacji polihorycznych. Choć należy zaznaczyć, że aby były to podejścia formalnie ekwiwaletne, konieczne byłoby zastąpienie w równaniach (9) i (5) (analogicznie dla modelu 2PL w równaniach (4) i (3)) logistycznej funkcji łączącej funkcją probitową (Bartholomew 1987; Takane, de Leeuw 1987).

Ostatecznie do skalowania wyników egzaminów na potrzeby wyliczania wskaźników EWD zdecydowano się wykorzystać modele dwuparametryczny dla zadań binarnych i model *graded response* dla zadań o kilku możliwych poziomach wykonania. Decyzja ta częściowo podyktowana została faktem dostępności oprogramowania w ramach projektu – do skalowania wykorzystano bowiem program Mplus, dający możliwość estymacji modeli strukturalnych, na potrzeby innych prac badawczych prowadzonych w ramach projektu, w szczególności związanych z badaniami podłużnymi. Jednocześnie program ten nie daje możliwości estymacji modelu trzyparametrycznego, ani modelu *partial credit*. W świetle przedstawionej powyżej argumentacji nie powinno to być jednak uznawane za poważny problem.

Estymacja modeli dokonywana jest metodą *Marginal Maximum Likelihood*, to znaczy przy założeniu, że rozkład mierzonej cechy w badanej populacji jest normalny. W przypadku egzaminu gimnazjalnego i sprawdzianu, gdy wszyscy zdający rozwiązują ten sam zestaw zadań, możliwe jest zastosowanie zarówno estymacji opartej na analizie macierzy korelacji tetra/polihorycznych jak też estymacji

odwołującej się do optymalizacji funkcji wiarygodności bezpośrednio w odniesieniu do pełnej macierzy danych. Pierwsze z nich wywodzi się z tradycji analizy modeli strukturalnych (SEM/CSEM), drugie zaś jest bliżej związane z tradycją modelowania IRT. Zaletą pierwszego podejścia jest mniejsza złożoność obliczeniowa procedur estymacji oraz istnienie wielu indeksów pozwalających ocenić stopień dopasowania modelu do danych. Nie znajduje ono jednak zastosowania w sytuacji, gdy niektórzy zdający rozwiązywali inny zestaw zadań niż inni, z czym mamy do czynienia w przypadku egzaminu maturalnego. Ze względu na chęć stosowania tego samego podejścia przy estymacji modeli dla wszystkich egzaminów oraz przywiązanie do tradycji IRT zdecydowano się na zastosowanie drugiego podejścia – optymalizacji na podstawie pełnej macierzy danych.

Oszacowania umiejętności uczniów uzyskiwane są metodą EAP (Expected A'Posteriori), a następnie przekształcane liniowo tak, aby ich średnia w grupie wszystkich zdających wynosiła 100, a odchylenie standardowe 15.

Pewien problem przy skalowaniu wyników polskich egzaminów zewnętrznych z użyciem modeli IRT sprawia określenie, czy poszczególne części wyróżnione w arkuszu egzaminacyjnym i reprezentowane w zbiorze danych przez oddzielne zmienne stanowią odrębne zadania, czy też należałoby je traktować jako powiązane ze sobą części tego samego zadania. W tym drugim przypadku punktacja za takie składowe powinna zostać zsumowana, a otrzymana suma powinna być traktowana jako wskaźnik poziomu wykonania całego zadania. W szczególności jako jedno zadanie powinny być traktowane takie następujące po sobie (choć oddzielnie oceniane) polecenia, w których możliwość poprawnego wykonania następnego uwarunkowana jest uzyskaniem prawidłowego wyniku w poprzednim kroku. Mówiąc formalnie, prawdopodobieństwo poprawnej odpowiedzi na zadanie (uzyskania raczej wyższego niż niższego poziomu wykonania zadania) przy kontroli poziomu umiejętności ucznia nie powinno zależeć od udzielenia poprawnej odpowiedzi na inne zadania. Warunek ten określany jest mianem lokalnej niezależności.

W praktyce stwierdzenie, czy zadania nie łamią warunku lokalnej niezależności nie zawsze jest łatwe i przede wszystkim wymaga uważnej analizy treści zadań. Niestety ani struktura arkuszy (numeracja zadań w ramach arkusza), ani udostępniana przez Centralną Komisję Egzaminacyjną dokumentacja dotycząca egzaminów nie pozwalają w łatwy sposób stwierdzić, które z oddzielnie punktowanych działań zdającego należy traktować jako niezależne zadania, a które jako części składowe jednego zadania. Wydaje się przy tym, że to właśnie autorzy testu byliby najwłaściwszymi osobami do dokonania takich rozstrzygnięć. Z racji ograniczonych zasobów ludzkich również zespół EWD nie jest niestety w stanie samodzielnie prowadzić analizy treści arkuszy egzaminacyjnych w celu dokonania takiej klasyfikacji. W związku z tym w przypadku sprawdzianu i egzaminu gimnazjalnego przyjęto podejście, że każda z oddzielnie punktowanych czynności ucznia traktowana jest w modelu jako niezależne zadanie. Niestety, jeżeli w wyniku złamane będzie założenie o lokalnej niezależności zadań, wpłynie to niekorzystnie na jakość uzyskiwanych oszacowań. W takiej sytuacji można spodziewać się przypisania zbyt wysokich wartości dyskryminacji zadaniom, które łamią założenie lokalnej niezależności i w efekcie zawyżenie wkładu tychże zadań w uzyskiwane z modelu oszacowania poziomu umiejętności.

Przy skalowaniu wyników egzaminów modelami IRT przyjęto podejście, że każda z części egzaminu gimnazjalnego skalowana będzie oddzielnie. W odniesieniu do nowej formuły egzaminu gimnazjalnego przyjęto, że oddzielnymi modelami skalowany będzie każdy z czterech testów, a do tego niezależnie przeprowadzona zostanie estymacja modeli traktujących jako jeden konstrukt oba testy odpowiednio części humanistycznej i części matematyczno-przyrodniczej. Decyzję o niestosowaniu wielowymiarowych modeli IRT podjęto z dwóch powodów. Po pierwsze, znacznie

większego stopnia komplikacji i złożoności obliczeniowej modeli wielowymiarowych. Po drugie, ze względu na fakt nieuchronnie zawyżanej korelacji (w stosunku do korelacji latentnych) pomiędzy oszacowaniami poziomów umiejętności na różnych wymiarach, uzyskiwanych z takich modeli. W efekcie oszacowania poziomu umiejętności z różnych dziedzin (części egzaminu) uzyskiwane z modeli wielowymiarowych bardzo niewiele się od siebie różnią, co stawiałoby pod znakiem zapytania zasadność wyliczania różnych wskaźników EWD, opartych na poszczególnych częściach lub testach wchodzących w skład egzaminu gimnazjalnego.

Przy skalowaniu wyników egzaminów przyjęto zasadę, że zadania o dyskryminacji mniejszej niż 0,2 są usuwane. W pojedynczym kroku usuwane jest tylko jedno zadanie o najmniejszej dyskryminacji, a następnie model estymowany jest ponownie. Procedura kontynuowana jest do momentu, aż wszystkie zadania uwzględnione w modelu będą mieć dyskryminację powyżej przyjętego progu. Zadania, które zostały usunięte z modeli skalowania w wyniku zastosowania tej procedury zestawione zostały w Aneksie C.

2.3. Skalowanie wyników matury

Skalowanie wyników egzaminu gimnazjalnego wiąże się ze specyficznymi problemami. Pierwszym z nich jest fakt, że wyjąwszy dwa przedmioty obowiązkowo zdawane na poziomie podstawowym – język polski i matematykę – uczniowie mają dalece posuniętą swobodę wyboru zdawanych przedmiotów. W związku z tym w ramach poszczególnych szkół może występować bardzo niewielka liczba uczniów zdających konkretny przedmiot. Ponieważ wskaźniki EWD stanowią źródło użytecznych informacji tylko w sytuacji, gdy są wyliczane na podstawie wyników większej grupy uczniów, pożądane, a wręcz konieczne staje się wyliczenie wskaźników pozwalających w sposób bardziej ogólny określić poziom umiejętności uczniów, w ramach szerszych dziedzin wiedzy.

Takie wskaźniki ogólnych umiejętności muszą uwzględniać wiele różnych przedmiotów i to w ten sposób, aby poziom umiejętności zdających można było wyrazić na jednej skali bez względu na zdawany przez nich zestaw przedmiotów. Jest to możliwe przy użyciu modeli IRT, przy czym duże znaczenie ma tu fakt, że istnieją dwa przedmioty obowiązkowo zdawane przez wszystkich. Pozwala to zapewnić odpowiedni poziom powiązania (*linkage*) danych, niezbędny aby móc traktować wyniki uzyskane z różnych przedmiotów (poziomów) jako wskaźniki jednego, wspólnego konstruktu. Po drugie, nawet w przypadku przedmiotów zdawanych obowiązkowo na poziomie podstawowym – języka polskiego i matematyki – które mogą być z powodzeniem wykorzystane samodzielnie do wyliczania wskaźników EWD, wartościowe byłoby jednoczesne uwzględnienie wyników uzyskanych przez zdających na poziomie rozszerzonym (oczywiście o ile deklarowali jego zdawanie).

W związku z tym zdecydowano się na wyliczanie na podstawie wyników matury czterech różnych wskaźników poziomu umiejętności zdających:

1. W zakresie języka polskiego - na podstawie wyników z języka polskiego na poziomie podstawowym i rozszerzonym.
2. W zakresie matematyki - na podstawie wyników z matematyki na poziomie podstawowym i rozszerzonym.
3. W zakresie przedmiotów humanistycznych - na podstawie wyników z języka polskiego, historii i wiedzy o społeczeństwie na poziomie podstawowym i rozszerzonym.
4. W zakresie przedmiotów matematyczno-przyrodniczych - na podstawie wyników z matematyki, biologii, chemii, fizyki, geografii i informatyki na poziomie podstawowym i rozszerzonym.

Pominięto przy tym rzadziej wybierane przedmioty, gdyż w ich przypadku występowałyby trudności z wyestymowaniem parametrów zadań.

Warto przy tym zauważyć, że modele, w których estymowane są ogólne umiejętności – w zakresie przedmiotów humanistycznych lub matematyczno-przyrodniczych – są już modelami bardzo złożonymi, zawierającymi odpowiednio po sto kilkadziesiąt i ponad dwieście różnych zmiennych. Aby nieco uprościć ich strukturę w przypadku arkuszy maturalnych przyjęte zostało inne niż w przypadku sprawdzianu i egzaminu gimnazjalnego podejście do określania, co jest oddzielnymi zadaniami. Zastosowano tu prostą regułę, w myśl której zadanie definiowane jest przez swój numer. Wszelkie „podpunkty", wyróżnione w arkuszu czy to przez dodanie drugiego stopnia numeracji, czy przez dodanie do numeru zadania oznaczenia literowego, traktowane są jako składowe tego samego zadania, a punktacja za nie jest sumowana. Poszczególne kryteria oceny wypracowania z języka polskiego traktowane są przy tym jako oddzielne zadania. Reguła ta ma niestety często niewiele wspólnego z występowaniem rzeczywistych zależności pomiędzy treścią łączonych kryteriów oceny, jednak ma przynajmniej tę zaletę praktyczną, że pozwala nieco zmniejszyć liczbę zmiennych w modelach skalowania. Oczywiście również w przypadku arkuszy maturalnych najlepiej byłoby, gdyby informacja o tym, które kryteria oceny należy łączyć w zadania na potrzeby skalowania wyników modelami IRT udostępniana była prze Centralną Komisję Egzaminacyjną.

Dodatkowo w celu uproszczenia modelu stosowana jest procedura skracania skal punktacji poszczególnych zadań. Przyjęto, że dla każdego zadania najrzadziej występujący poziom wykonania łączony jest z rzadziej występującym spośród sąsiadujących z nim poziomów tak długo, aż liczba różnych możliwych poziomów wykonania zadania nie zostanie ograniczona do pięciu. W szczególności oznacza to znaczne skrócenie skal w przypadku kryteriów oceny wypracowania z języka polskiego.

Jak już wcześniej wspomniano, specyfika egzaminu maturalnego polega na tym, że uczniowie sami decydują, jakie przedmioty i na jakim poziomie będą zdawać. Oczywiście wyborów tych nie dokonują oni losowo, lecz w sposób strategiczny, związany z poziomem swoich umiejętności (w szczególności należy się spodziewać, że przedmioty na poziomie rozszerzonym wybierają uczniowie o wyższym poziomie umiejętności). Aby uzyskać dobre oszacowania parametrów modelu, a potem także dobre oszacowania poziomu umiejętności zdających, konieczne jest więc uwzględnienie w modelu jakichś parametrów pozwalających opisać takie procesy autoselekcji.

W sytuacji, gdyby możliwych do zdawania części egzaminu było niewiele, dobrym rozwiązaniem byłoby założenie, że każda grupa zdających wyróżniona ze względu na zestaw zdawanych części egzaminu charakteryzuje się innym średnim poziomem i zróżnicowaniem mierzonej umiejętności. Niestety w sytuacji, gdy liczba możliwych do wyboru części egzaminu sięga jedenastu – jak to się dzieje w przypadku modelu dla umiejętności matematyczno-przyrodniczych – maksymalna liczba grup możliwych do wyróżnienia ze względu na zestaw zdawanych części egzaminu przekracza jedenaście tysięcy. Choć w praktyce liczba rzeczywiście występujących w danych zestawów jest z pewnością dużo mniejsza, i tak jest ich zbyt wiele, aby móc rozsądnie myśleć o estymowaniu średniej i odchylenia standardowego mierzonej umiejętności oddzielnie dla każdego z nich.

W związku z tym przyjęto nieco inne podejście, wzorowane na rozwiązaniu zaproponowanym w publikacji Korobko, Glasa, Boskera i Luytena (2008). Polega ono na włączeniu do modelu zmiennych (po jednej dla każdej części egzaminu) traktowanych tak, jak zadania, opisujących, czy uczeń zdawał daną część egzaminu, czy też nie. Wartości parametrów dyskryminacji takich „zadań" –

określane mianem parametrów selekcji – pozwalają potem określić, w jakim stopniu wybór poszczególnych części egzaminu powiązany jest z wysokim poziomem mierzonych umiejętności.

Trzeba jednak zaznaczyć, że aby przyjęte rozwiązanie dawało dobre rezultaty konieczne jest, aby uczniowie (a przynajmniej zdecydowana większość z nich) przystępowali do różnych części egzaminu maturalnego z podobnym zaangażowaniem (poziomem motywacji). Jest to tym bardziej istotne, że zadania rozwiązywane przez wszystkich, zapewniające możliwość przedstawienia wyników na jednej skali, obejmują tylko jedną dziedzinę (język polski lub matematykę na poziomie podstawowym). W związku z tym różnice średnich oszacowań umiejętności pomiędzy różnymi grupami uczniów wynikają przede wszystkim z tego, o ile lepiej lub gorzej poradzili sobie uczniowie należący do tych grup z rozwiązaniem zadań „wspólnych". Jeśli pewna grupa uczniów – przykładowo przyjmijmy, że są to uczniowie klas biologiczno-chemicznych – mniej „przykłada się do rozwiązania egzaminu z matematyki na poziomie podstawowym" (niezależnie od rzeczywistego poziomu swoich umiejętności) niż inna grupa uczniów – powiedzmy uczniowie klas matematyczno-fizycznych – to oszacowania umiejętności uczniów z tej pierwszej grupy będą zaniżone w stosunku do uczniów drugiej grupy. Niestety nie mamy obecnie możliwości wiarygodnej oceny, jaka może być skala występowania tego rodzaju problemu. Wydaje się, że na motywację uczniów podczas pisania poszczególnych części egzaminu maturalnego może wpływać przede wszystkim użyteczność danego przedmiotu w procesie rekrutacji na wybrany kierunek studiów. Można przy tym zauważyć, że wyniki uzyskane z przedmiotów obowiązkowych w przypadku większości kierunków odgrywają drugorzędną rolę. Jednocześnie egzaminy z języka polskiego i matematyki na poziomie podstawowym są raczej łatwe i wydaje się,że poświęcenie większej uwagi nauce innych przedmiotów nie stanowi przeszkody dla uzyskania z nich dobrego wyniku. Pozwala to mieć nadzieję, że uczniowie wybierające różne przedmioty nieobowiązkowe nie różnią się znacząco jeśli chodzi o przeciętny poziom motywacji podczas rozwiązywania przedmiotów obowiązkowych.

Problemem bardzo zbliżonym do wyboru zdawanych przedmiotów jest wybór tematu wypracowania z języka polskiego, a także wypowiedzi pisemnych wchodzących w skład arkuszy z WOSu i historii na poziomie rozszerzonym. Może on zostać rozwiązany w analogiczny sposób, tj. poprzez potraktowanie tych samych kryteriów oceny wypracowania jako różnych zadań – w zależności od wybranego tematu – i dodanie do modelu dodatkowego „zadania" opisującego dokonany wybór tematu (Pokropek, 2011: 441-447). Jednocześnie występują tu dwie dodatkowe komplikacje. Po pierwsze, przed 2013 r. nie zbierano w ramach danych egzaminacyjnych scalanych na potrzeby wyliczania wskaźników EWD informacji o temacie wybranym przez ucznia. Po drugie, występuje problem z uwzględnieniem w skalowaniu laureatów, jako że ich wyników nie da się automatycznie przypisać do żadnego tematu. W praktyce najlepszym rozwiązaniem wydaje się wyłączenie ich z estymacji, podczas której wyliczane są parametry modelu. Następnie należy przypisać laureatom w danych wybór tego tematu, który wybierany był przez uczniów o wyższym poziomie mierzonej cechy i na podstawie wyliczonych wcześniej wartości parametrów modelu wyestymować oszacowania poziomu umiejętności laureatów. Rozwiązania te zostaną zapewne wdrożone w przyszłości, jednak do chwili obecnej problem wyboru tematu wypracowania nie był odwzorowywany w modelach wykorzystywanych do skalowania wyników matury.

Przy skalowaniu matury przyjęto, że zarówno wyniki uczniów liceów ogólnokształcących jak i wyniki uczniów techników skalowane są w ramach jednego modelu, a uczniów tych dwóch typów szkół traktuje się jako tworzących jedną, homogeniczną populację, w ramach której badane umiejętności mają rozkład normalny. Założenie to upraszcza estymację modelu, choć w oczywisty sposób stanowi idealizację. Wyniki uzyskiwane na maturze przez uczniów liceów ogólnokształcących są zdecydowanie wyższe niż uczniów techników. W związku z tym, aby zachować intuicyjność

interpretacji skal, na jakich prezentowane są wyniki (wskaźniki prezentowane są oddzielnie dla liceów ogólnokształcących i oddzielnie dla techników), wyskalowane wyniki matury standaryzowane są oddzielnie dla uczniów szkół obu typów, a następnie oddzielnie przekształcane liniowo do skali o średniej 100 i odchyleniu standardowym 15.

W przyszłości wskazane byłoby zbadanie własności modeli, w których uczniów liceów ogólnokształcących i uczniów techników traktowano by jako pochodzących z dwóch odrębnych populacji, w ramach których badane umiejętności mają rozkłady normalne, jednak rozkłady te mogą różnić się co do średniej i wariancji.

3. Modele EWD

W tej części raportu omówione zostaną kwestie związane z formą i metodami estymacji modeli EWD oraz wyliczania wartości wskaźników EWD. Kolejno poruszone zostaną problemy wyboru metody najlepszego odwzorowania nieliniowej, w polskich warunkach, relacji pomiędzy wynikami egzaminu „na wejściu" i „na wyjściu" ze szkoły, włączania do modelu dodatkowych zmiennych kontrolnych, wyboru metody estymacji modelu i metody wyliczania samych wskaźników. Na zakończenie przedstawione zostaną pierwsze doświadczenia z wyliczaniem wskaźników EWD przy pomocy modeli strukturalnych (SEM).

3.1. Modelowanie relacji pomiędzy wynikami „na wejściu", a wynikami „na wyjściu"

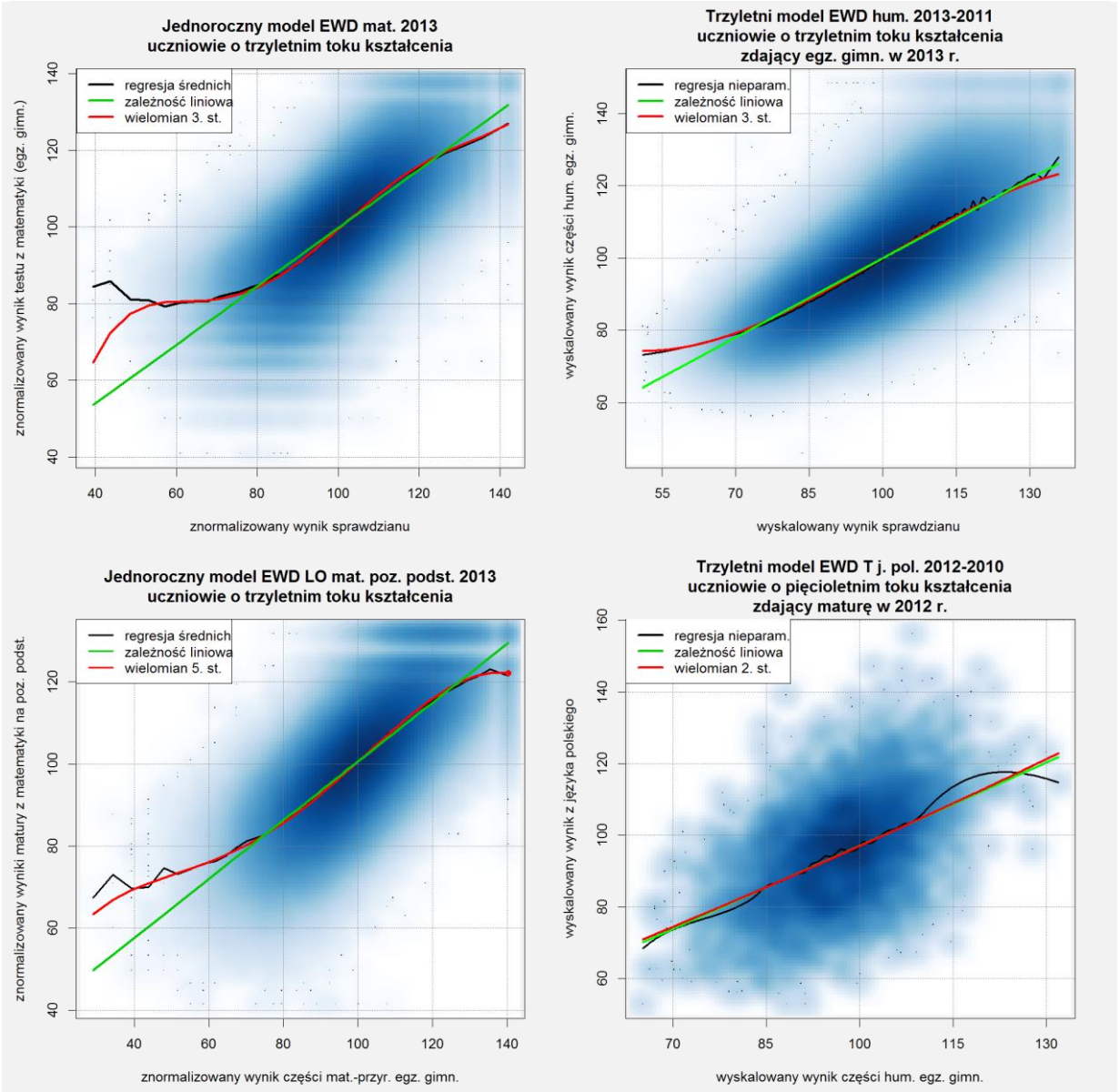
Wyniki egzaminu „na wejściu" są najważniejszą zmienną, wykorzystywaną w modelach EWD do określenia przewidywanego wyniku egzaminu „na wyjściu". Zarówno z powodów praktycznych (prostoty i łatwości interpretacji modelu) jak i teoretycznych (założenie o jedności konstruktu mierzonego przez oba egzaminy) najbardziej pożądaną sytuacją byłoby, gdyby zależność pomiędzy tymi dwoma zmiennymi dawała się dobrze opisać prostą funkcją liniową. Niestety, w przypadku polskich egzaminów mamy do czynienia z bardziej złożoną sytuacją, w której zależność ta ma wyraźnie krzywoliniowy charakter. Zmusza to do rozważenia kwestii, w jaki sposób należy zależność tą modelować (jak ją sparametryzować) i jakie problemy mogą się z tym wiązać.

Ogólnie rzecz biorąc – choć od reguły tej zdarzają się wyjątki - zależność pomiędzy wynikami egzaminu „na wejściu" a wynikami egzaminu „na wyjściu" ma „esowaty" kształt, a dokładniej przebiega z dobrym przybliżeniem liniowo w dosyć szerokim spektrum wyników średnich, jednak w obszarze wyników wysokich i niskich ulega stopniowemu wypłaszczeniu[3]. Dodatkowo w zakresie skrajnie wysokich i skrajnie niskich wyników egzaminu „na wejściu" mogą występować dodatkowe nieregularności. Generalnie odchylenia od liniowego przebiegu są większe, gdy wykorzystywane są zrównane ekwikwantylowo wyniki egzaminów, a mniejsze, gdy wykorzystywane są wyniki egzaminów wyskalowane przy pomocy modeli IRT. Ponadto problem nieliniowości występuje dużo wyraźniej w modelach EWD dla gimnazjów niż w modelach EWD dla szkół kończących się maturą.

Źródeł nieliniowego przebiegu zależności można upatrywać przede wszystkim w niedoskonałych własnościach psychometrycznych wykorzystywanych egzaminów. Należy zaznaczyć, że polskie egzaminy zewnętrzne na różnych poziomach kształcenia nie są konstruowane z założeniem pomiaru tych samych konstruktów. Oczywiście empirycznie możemy stwierdzić, że mierzą one w gruncie rzeczy zbliżone, dające się ze sobą rozsądnie porównywać dyspozycje uczniów (Jasińska i Żółtak, 2013; Pokropek, 2013), jednak różnice w zakresie treściowym mogą przekładać się na nieregularności w przebiegu zależności. Z pewnością rzeczą bardzo pożądaną, z punktu widzenia wzmocnienia ewaluacyjnej funkcji egzaminów zewnętrznych, byłoby uwzględnienie w założeniach konstruowania polskich egzaminów, że mają one służyć nie tylko ocenie tego, w jakim stopniu

[3] W sytuacji, gdy posługujemy się surowymi wynikami egzaminów, lub wynikami zrównanymi ekwikwantylowo w celu śledzenia przebiegu zależności można wykorzystać regresję średnich. Gdy wykorzystywane są wyniki wyskalowane przy pomocy modeli IRT dobrą techniką będzie z kolei regresja nieparametryczna.

uczniowie opanowali treści nauczania kończonego właśnie etapu kształcenia, ale również, że ich wyniki mają w przyszłości stanowić punkt odniesienia do oceny, jakie (relatywne) postępy poczynili oni na następnym etapie edukacji.



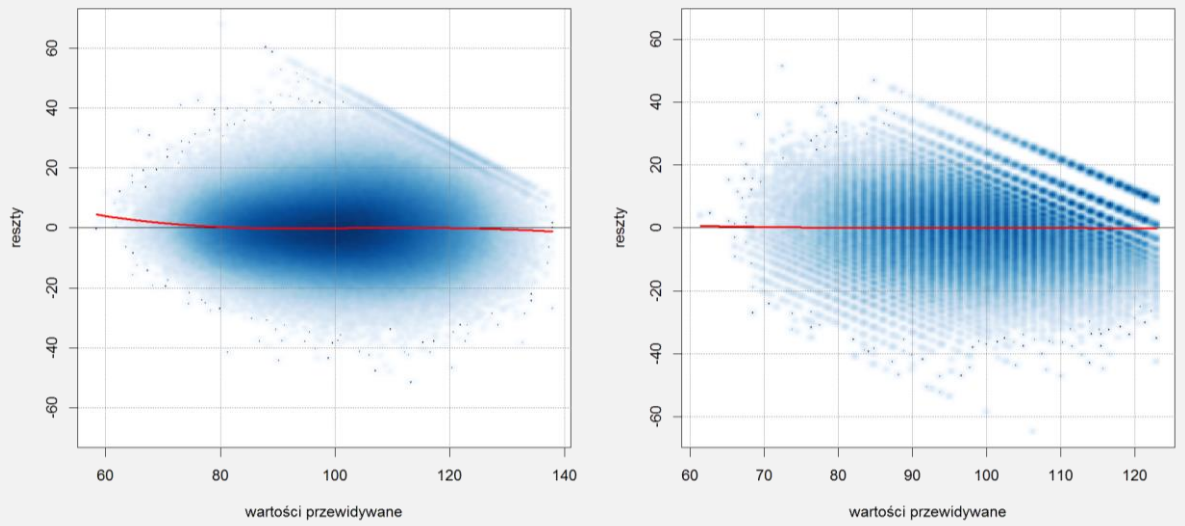
Rysunek 1. Przykłady przebiegu zależności pomiędzy wynikami egzaminu „na wejściu” a wynikami egzaminu „na wyjściu”.

Czynnikiem, który wpływa na wypłaszczanie się zależności między wynikami egzaminu „na wejściu” a wynikami egzaminu „na wyjściu” w zakresie bardzo wysokich wyników tego pierwszego, jest występowanie efektu sufitowego na egzaminie „na wyjściu”. Graficznie efekt ten można zaobserwować zarówno na wykresach rozrzutu obrazujących zależność pomiędzy wynikami obu egzaminów (por. Rysunek 1.), jak też na wykresie diagnostycznym modelu EWD, obrazującym relację pomiędzy wynikami przewidywanymi a resztami regresji (por. Rysunek 2.). Na tych pierwszych objawia się on w postaci przebiegającej wzdłuż prawej części górnej krawędzi wykresu linii – są to osoby o najwyższym możliwym do uzyskania wyniku. Na drugim typie wykresów obecność efektu sufitowego zaznacza się jeszcze wyraźniej, w formie skośnego „ucięcia” wykresu rozrzutu w jego

prawej górnej części. Występowanie takich wzorów na wykresach rozrzutu sugeruje, że osoby zgrupowane „na krawędzi" w rzeczywistości różnią się między sobą co do poziomu umiejętności „na wyjściu" ze szkoły i części z nich należałoby przypisać wyższy poziom umiejętności – aby w efekcie „zapełniła się" ucięta część wykresów. Niestety, wykorzystywane przez nas do pomiaru umiejętności narzędzie nie pozwala zróżnicować tej grupy uczniów. Mechanizm wpływu efektu sufitowego na wypłaszczanie zależności jest więc taki, że gdyby efekt ten nie występował (egzamin lepiej różnicował uczniów o najwyższych wynikach), części uczniów o wysokich wynikach egzaminu „na wejściu" zostałyby przypisane wyższe wyniki egzaminu „na wyjściu", co „ciągnęłoby w górę" również wartości przewidywane.

Należy przy tym zaznaczyć, że skala efektu sufitowego jest bardzo zróżnicowana w zależności od egzaminu, a nawet roku, w którym był on przeprowadzony. Efekt sufitowy najsilniej zaznacza się w jednorocznych modelach EWD liceów ogólnokształcących w zakresie matematyki, w których jako miara poziomu umiejętności „na wyjściu" brany jest wynik matury z matematyki na poziomie podstawowym. W przypadku techników problem jest zdecydowanie mniejszy, z racji ogólnie niższego poziomu umiejętności matematycznych uczniów. Również w przypadku egzaminu gimnazjalnego efekt sufitowy nie jest duży (zwykle nieco większy dla części humanistycznej, niż dla części matematyczno-przyrodniczej egzaminu), choć pozostaje zauważalny.

Dodatkowym źródłem nieregularności w prawej części linii przewidywanego wyniku jest występowanie laureatów konkursów przedmiotowych – zarówno w grupie osób z maksymalnym wynikiem egzaminu „na wejściu", jak i z maksymalnym wynikiem egzaminu „na wyjściu". Na szczęście nie jest to duża liczebnie grupa osób, ale i tak wywiera ona znaczny wpływ na przebieg linii regresji w jej skrajnie prawej części.



Rysunek 2. Wykres reszt regresji w funkcji wartości przewidywanych. Lewy panel: trzyletni model EWD gimnazjów, w zakresie przedmiotów humanistycznych 2013-2011.Prawy panel: jednoroczny model EWD liceów ogólnokształcących w zakresie matematyki 2013. Czerwoną linią zaznaczono linię przewidywania z regresji nieparametrycznej.

Niestety trudno stwierdzić, jakie czynniki wpływają na wypłaszczanie się zależności w zakresie niskich wyników „na wejściu". Z pewnością nie jest to sytuacja analogiczna do omówionej przed chwilą – w wynikach wykorzystywanych egzaminów nie obserwujemy efektu podłogowego. Być może problem

polega na wspomnianej już wcześniej kwestii niedoskonałej zgodności mierzonych konstruktów. Treści będące bardzo proste już w momencie dokonywania pomiaru „na wejściu" okazują się mieć niewiele wspólnego z konstruktem mierzonym trzy lata później.

Do modelowania nieliniowej zależności pomiędzy wynikami egzaminu „na wejściu" a wynikami egzaminu „na wyjściu" wykorzystywane były w modelach EWD dwie różne metody. Na początku prac nad jednorocznymi modelami EWD gimnazjów przyjęto rozwiązanie wykorzystujące regresję dwoma kawałkami liniową (Jakubowski, 2007), estymowaną metodą największej wiarygodności. Była ona stosowana w modelach dla roczników egzaminu 2005-2007. W późniejszym okresie została ona zastąpiona regresją MNK wykorzystującą wielomianowe przekształcenie wyników egzaminu „na wyjściu". Zaletą drugiego podejścia jest możliwość wykorzystania prostszych, bardziej standardowych technik estymacji modelu oraz – w przypadku wykorzystania wielomianów wyższych stopni – lepsze dopasowanie do nieregularnego przebiegu zależności.

W podejściu wykorzystującym wielomian do modelowania nieliniowej zależności pomiędzy wynikami egzaminu „na wejściu" a wynikami egzaminu „na wyjściu" konieczne jest rozstrzygnięcie kwestii, jaki będzie stopień wykorzystywanego w modelach wielomianu. W latach 2008-2010 nie wdrożono w tym aspekcie żadnej sformalizowanej procedury. W przypadku każdego modelu przygotowujący go analityk dokonywał subiektywnej oceny na podstawie porównania dopasowania wynikającego z zastosowania wielomianów kolejnych stopni z regresją średnich. W 2011 r. wdrożona została sformalizowana procedura wyboru stopnia wielomianu, opierająca się na dwóch kryteriach:

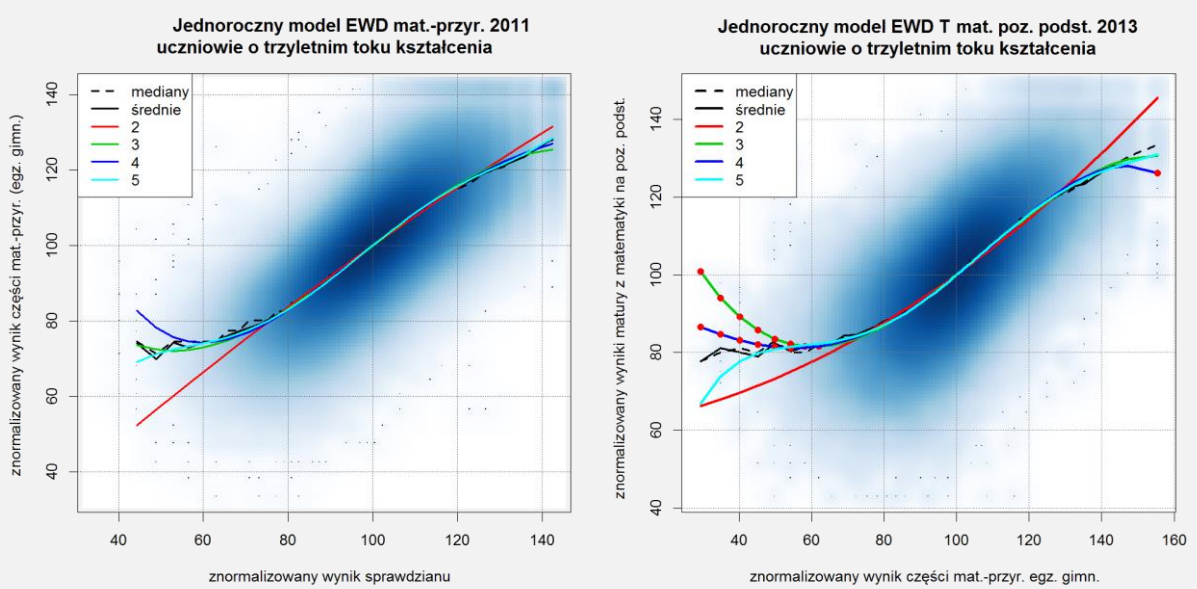
1) monotonicznego przebiegu zależności w zakresie występujących w danych wyników,
2) ocenie stopnia niedopasowania modelu do danych z wykorzystaniem zmodyfikowanego testu „linktest".

Pierwsze kryterium zakłada, że w ramach modelu, który zostanie wybrany, zależność pomiędzy wynikami egzaminu „na wejściu" i „na wyjściu" musi być rosnąca na przedziale od minimum do maksimum wyników egzaminu „na wejściu", jakie występują w danych. Kryterium to jest motywowane teoretycznie. Jeśli chcemy myśleć o wynikach egzaminów „na wejściu" i „na wyjściu" jako o pomiarach tego samego konstruktu, to powinniśmy oczekiwać, że związek pomiędzy nimi będzie pozytywny. Drugie kryterium zakłada, że „linktest" przeprowadzony na modelu musi okazać się nieistotny statystycznie (przyjęto poziom istotności równy 0,01), co wskazuje na bezcelowość dalszego zwiększania stopnia wielomianu. Opis stosowanej modyfikacji „linktestu" zawarty został w Aneksie D. Wybierany jest najniższy stopień wielomianu, który jednocześnie spełnia oba kryteria. Dodatkowo założono, że stopień wielomianu nie może być wyższy niż pięć. W przypadku modeli uwzględniających uczniów o wydłużonym toku kształcenia i/lub kilka kohort absolwentów oba kryteria sprawdzane są niezależnie w ramach każdej spośród grup uczniów wyróżnionych na podstawie pary: (rok zdawania egzaminu „na wyjściu", rok zdawania egzaminu „na wejściu"). W miarę możliwości wybierany jest najniższy stopień wielomianu, który w każdej z tak wyróżnionych grup spełnia oba wymienione kryteria.

Warto zaznaczyć, że wybór konkretnego stopnia wielomianu ma znikomy wpływ na przebieg krzywej przewidywanego wyniku w zakresie wyników średnich[4], może jednak w znacznym stopniu zmieniać jej

[4] Uczniów o skrajnych wynikach jest po prostu dalece zbyt mało, by mogli oni wywierać taki wpływ. W związku z tym, patrząc z punktu widzenia diagnostyki całego modelu nie można ich uznać za jednostki wpływowe

położenie w zakresie skrajnych wyników egzaminu „na wejściu", zwłaszcza wyników niskich, co obrazuje Rysunek 3. Jednocześnie można zauważyć, że przebieg regresji średnich, czy dopasowania nieparametrycznego niejednokrotnie jest niemonotoniczny, co sprawia, że postulat lepszego dopasowania modelu parametrycznego do danych może wchodzić w sprzeczność z postulatem monotoniczności przewidywania. W praktyce jednoczesne spełnienie obu przedstawionych wyżej kryteriów bywa niemożliwe i konieczne jest niewielkie rozluźnienie kryteriów, aby móc wybrać stopień wielomianu, jaki zostanie użyty w modelu. Niestety tego typu problemy związane z nieregularnym przebiegiem zależności pomiędzy wynikami egzaminu „na wejściu" a wynikami egzaminu „na wyjściu" mogą zostać efektywnie rozwiązane tylko na etapie przygotowywania egzaminów. W Aneksie E zestawione zostały informacje dotyczące wyboru stopnia wielomianu dla modeli EWD wykorzystanych do wyliczenia aktualnie prezentowanych wskaźników.



Rysunek 3. Przykłady przebiegu krzywej przewidywanego wyniku w zależności od stopnia wielomianu. Lewy panel – jednoroczny model EWD gimnazjów w zakresie przedmiotów mat.-przyr. 2011 (uczniowie o trzyletnim toku kształcenia). Prawy panel – jednoroczny model EWD techników w zakresie matematyki na poziomie podstawowym 2010 (uczniowie o czteroletnim toku kształcenia).

3.2. Dodatkowe zmienne w modelu

Kwestia uwzględniania w modelach EWD dodatkowych zmiennych, pozwalających kontrolować wpływ na czynników niezależnych od szkoły (innych niż tylko wyniki egzaminu „na wejściu") na wyniki osiągane przez uczniów pod koniec danego etapu edukacji była już niejednokrotnie poruszana w polskich publikacjach dotyczących wskaźników EWD (Dolata, 2007b; Żółtak ,2011). Omówione zostały w nich również problemy teoretyczne i interpretacyjne związane z włączeniem do modeli szerokiego zakresu zmiennych kontrolnych, a w szczególności zmiennych zagregowanych na poziomie szkoły (Żółtak, 2013a, 2013b). Wiele ważnych informacji na temat wpływu różnych czynników na osiągnięcia szkolne w polskich szkołach przynoszą opublikowane ostatnio wyniki badań podłużnych (Dolata i in., 2013; Karwowski, 2013). Pozwalają one wnioskować, jakiego rodzaju i jak silne obciążenie może wprowadzać do wskaźników EWD nie uwzględnienie w modelach EWD danych o statusie społeczno-ekonomicznym uczniów, czy pozaszkolnym wsparciu edukacyjnym.

W praktyce zakres informacji, jakie mogą być uwzględnione w polskich modelach EWD jest jednak bardzo wąski. Problemem jest tu słabość polskiego sytemu informacji oświatowej. Niestety w skali ogólnopolskiej niemożliwe jest dołączenie do wyników egzaminacyjnych żadnych dodatkowych informacji o uczniach ponad te, które przechowywane są w bazach danych okręgowych komisji egzaminacyjnych, a które obejmują płeć, wiek (datę urodzenia), oraz informacje o tym, czy uczeń na egzaminie posiadał zaświadczenie o dysleksji oraz czy był laureatem. Zakres zmiennych uwzględnianych w modelach jest przy tym jeszcze nieco węższy, gdyż obejmuje tylko płeć i informacje o dysleksji (podczas egzaminu „na wejściu", podczas egzaminu „na wyjściu" oraz interakcję tych dwóch zmiennych).

Włączenie do modeli płci oznacza, że jeśli w skali ogólnokrajowej występują różnice w średnich postępach dziewcząt i chłopców, to nie będą one mieć wpływu na wskaźniki EWD szkół - punktem odniesienia dla dziewcząt będą inne dziewczęta, a dla chłopców inni chłopcy. Analogiczna sytuacja występuje w przypadku informacji o posiadaniu zaświadczenia o dysleksji, z tym że inny punkt odniesienia jest tu ustalany w ramach każdej z czterech grup: uczniów, którzy nie posiadali takiego zaświadczenia na żadnym z egzaminów, uczniów, którzy posiadali je tylko na egzaminie „na wejściu", uczniów, którzy posiadali je tylko na egzaminie „na wyjściu" i wreszcie uczniów, którzy posiadali je na obu egzaminach. Istnieją powody by przypuszczać, że posiadanie zaświadczenia o dysleksji nie jest zbyt dobrym wskaźnikiem występowania specyficznych trudności w nauce czytania i pisania (Dolata, 2007b), jest za to jasne, że wyniki egzaminów takich osób nie powinny być bezpośrednio porównywane z wynikami innych zdających, gdyż uczniowie posiadający zaświadczenie o dysleksji mają wydłużony czas pisania egzaminu i stosuje się wobec nich łagodniejsze kryteria oceny, np. odnośnie błędów ortograficznych.

Odsetek osób posiadających zaświadczenia o dysleksji wynosi 7%-9% na sprawdzianie, 9%-11% na egzaminie gimnazjalnym i ~7% na maturze. Odsetek uczniów, którzy posiadali zaświadczenie zarówno na sprawdzianie jak i na egzaminie gimnazjalnym to 5%-8%, a zarówno na egzaminie gimnazjalnym jak i na maturze ~6% (względem wszystkich zdających).

Laureaci konkursów przedmiotowych również stanowią specyficzną grupę uczniów. Nie podchodzą oni do egzaminu (w przypadku egzaminu gimnazjalnego - do jednej z jego części, odpowiadającej tematyce konkursu[5], którego są laureatami), lecz mają przypisywane maksymalne możliwe do uzyskania wyniki. Jednocześnie laureaci stanowią grupę niezwykle zróżnicowaną wewnętrznie. Liczba konkursów dających możliwość zwolnienia z egzaminu jest znaczna, a do tego różna na terenie różnych województw, ponieważ organizatorami konkursów przedmiotowych są kuratoria oświaty. Oczywiście brak też standaryzacji co do trudności różnych konkursów. Sprawia to, że sensowność porównywania laureatów do innych laureatów jest mocno problematyczna. W związku z tym zdecydowano się nie uwzględniać w modelach informacji o byciu laureatem, uczniowie tacy traktowani są tak samo jak inni, którzy otrzymali z egzaminu maksymalną liczbę punktów.

Warto przy tym zaznaczyć, że liczba laureatów jest niewielka. Na sprawdzienie i egzaminie gimnazjalnym w zależności od egzaminu (części egzaminu) i roku jest ich od około tysiąca do blisko trzech tysięcy, co w odniesieniu do ogólnej liczby zdających daje 2‰-7‰. Na maturze na kazdy

[5] Przy czym po rozbicu egzaminu gimnazjalnego na 4 części w 2012 roku bycie laureatem konkursu zwalnia zależnie od tematyki konkursu, z dwóch części egzaminu (j.polski oraz historia i WOS / matematyka oraz przedmioty przyrodnicze)

przedmiot przypada od około czterdziestu do około stu laureatów, z wyjątkiem języka polskiego, gdzie jest ich około dwustu – łącznie jest to około siedmiuset różnych osób rocznie (biorąc pod uwagę tylko przedmioty uwzględniane przy wyliczaniu wskaźników EWD), to jest ~2‰ zdajacych.

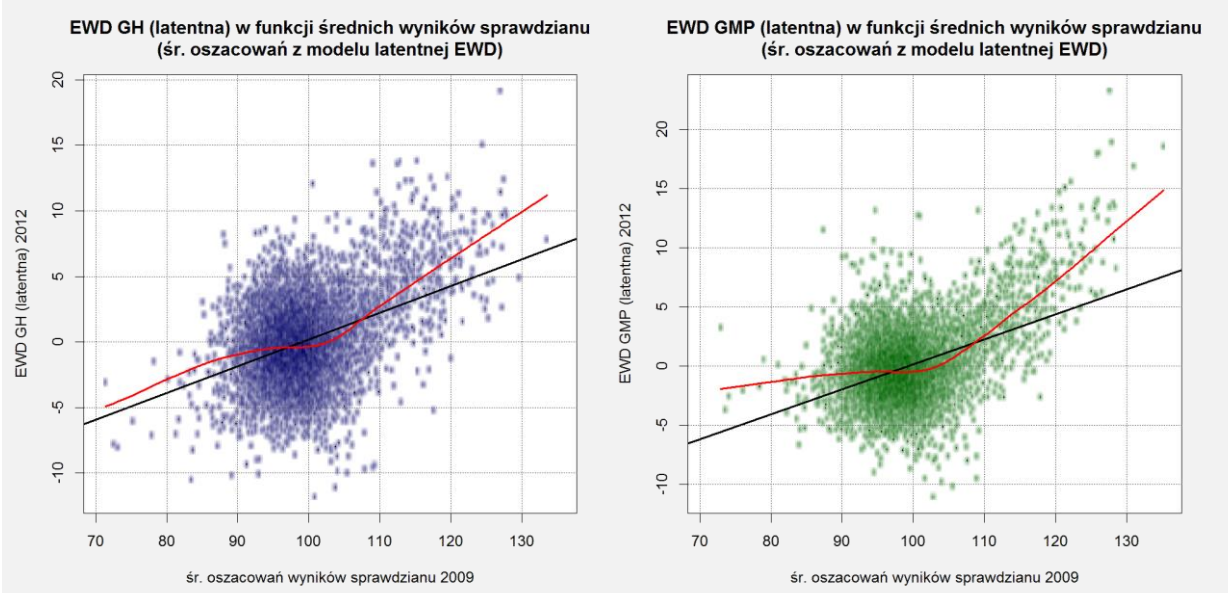
Z punktu widzenia EWD najlepszym rozwiązaniem problemu związanego z laureatami byłoby, gdyby pisali oni egzaminy tak, jak wszyscy uczniowie. Niestety można się spodziewać, że jeśli status laureata dawałby jak obecnie uprzywilejowaną pozycję przy rekrutacji do szkół na następnym etapie kształcenia, to motywacja laureatów podczas pisania egzaminów byłaby wyraźnie niższa, niż pozostałych uczniów. Co za tym idzie, relatywnie zaniżone (względem pozostałych zdających, o wyższym poziomie motywacji) byłyby zapewne ich wyniki egzaminu „na wyjściu". Z drugiej strony, w porównaniu do sytuacji panującej obecnie, wydaje się, że bardziej uprawione byłoby wtedy traktowanie laureatów jako grupy homogenicznej ze względu na obciążenie wyników egzaminacyjnych. Dawałoby to szanse uwzględnienia wpływu bycia laureatem na wyniki w modelach EWD.

Wiek uczniów nie jest uwzględniany w modelach EWD z powodów, które można by określić jako historyczne. W pierwszym okresie prac nad rozwojem wskaźników informacje o dacie urodzenia uczniów (na podstawie której można wyliczyć ich wiek) nie była zbierana wraz z wynikami egzaminów, co oczywiście uniemożliwiało uwzględnienie tego czynnika. W późniejszym okresie modyfikacje metodologii skupiły się z kolei na skalowaniu wyników egzaminacyjnych i uwzględnieniu uczniów o wydłużonym toku kształcenia. W świetle ostatnich wyników badań (Dolata i Pokropek, 2012; Hawrot i Jasińska, 2013) kwestia uwzględnienia wieku w gimnazjalnych modelach EWD wydaje się ważna i z pewnością będzie rozważana. Warto przy tym pamiętać, że sparametryzowanie wpływu wieku na wyniki egzaminacyjne jest zadaniem dosyć skomplikowanym, ze względu na występowanie w danym roczniku absolwentów uczniów spoza głównej kohorty wiekowej tj. uczniów, którzy rozpoczęli naukę wcześniej, uczniów, którzy rozpoczęli naukę z rocznym opóźnieniem, uczniów o wydłużonym toku kształcenia. Należy spodziewać się, że wpływ wieku zostanie opisany w modelu zależnością przedziałami liniową, na przedziałach wyznaczonych przez kolejne roczniki uczniów.

Nieco odrębny problem stanowi ewentualne uwzględnienie w modelach EWD danych zagregowanych na poziomie szkół. Najczęściej wykorzystywana jest tu średnia wyników egzaminu „na wejściu" wśród uczniów danej szkoły, traktowana jako wskaźniki wpływu rówieśników (Evans, 2008; Raudenbush i Willms, 1995). Niezaprzeczalną zaletą uwzględnienia średnich wyników „na wejściu" w modelu EWD, w sytuacji, gdy wskaźniki EWD mają być wykorzystywane do ewaluacji pracy szkół, jest zapewnienie niewystępowania związku pomiędzy wskaźnikami EWD a właśnie średnią osiągnięć uczniów „na wejściu" do szkoły. Inaczej mówiąc, w takiej sytuacji każda szkoła ma szanse na wysokie lub niskie EWD, bez względu na to, jacy uczniowie do niej przyszli. W sytuacji, gdy nie uwzględniać w modelu średnich wyników egzaminu „na wejściu" nie ma gwarancji, że taka niezależność będzie zachodzić. W praktyce, w przypadku polskich wskaźników, okazuje się ona wyraźna (Pokropek i Żółtak, 2012: 186; Żółtak, 2011) i to bez względu na różnice w metodzie estymacji modelu EWD i wyliczania wartości samych wskaźników.

Mimo to nie doszło do włączenia informacji o średnich wynikach egzaminu „na wejściu" do polskich modeli EWD, gdyż głębsze analizy problemu związku tej zmiennej z osiągnięciami uczniów pod koniec etapu kształcenia wykazały, że mamy tu do czynienia z bardziej złożonymi mechanizmami. Okazało się, że analizowany związek przebiega w inny sposób w grupach szkół wiejskich, szkół z mniejszych miast i szkół wielkomiejskich (Żółtak, 2011). Podczas gdy w tej ostatniej grupie związek był bardzo wyraźny, w małych miastach wyraźnie słabł, a w grupie szkół wiejskich właściwie nie występował. Późniejsze analizy prowadzone przy okazji badania własności „latentnych" wskaźników

EWD (zostały one opisane w końcowej części tego rozdziału) wskazują z kolei, że związek pomiędzy średnimi wyniki egzaminu „na wejściu" a wynikami egzaminu „na wyjściu" pojawia się w grupie szkół przyjmujących uczniów o ponadprzeciętnych wynikach „na wejściu", a nie występuje w grupie szkół przyjmujących uczniów o średnich wynikach egzaminów, co pokazuje **Błąd! Nie można odnaleźć ródła odwołania.**Rysunek 4. Wynika z tego, że proste wprowadzenie do polskich modeli EWD średnich wyników egzaminu „na wejściu" tylko pozornie stanowiłoby rozwiązanie problemu i taki model nie opisywał by dobrze rzeczywiście zachodzących zależności.



Rysunek 4. Zależność pomiędzy średnimi wynikami egzaminów „na wejściu" a wartościami wskaźników EWD gimnazjów 2012 z modeli latentnych (uczniowie o trzyletnim toku kształcenia). Czarną linią oznaczono przewidywanie liniowe, czerwoną linią przewidywanie nieparametryczne.

Uzyskane wyniki sugerują, że analizowany problem może być powiązany z selektywnością szkół, a więc z występowaniem segregacji ze względu na uprzednie osiągnięcia w ramach lokalnych systemów szkolnych. Weryfikacja tej hipotezy wymagałaby jednak dalszych analiz. Należy też zaznaczyć, że wszystkie opisane wyżej wyniki uzyskane zostały w odniesieniu do gimnazjów. Jak do tej pory brak jest systematycznych analiz omawianej problematyki w odniesieniu do szkół kończących się maturą, choć wiadomo, że również w ich przypadku występuje wyraźna korelacja pomiędzy średnimi wynikami egzaminu „na wejściu" a wartościami wskaźników EWD i że zależność ta jest silniejsza, niż w przypadku gimnazjów.

W kontekście rozpatrywania zakresu zmiennych kontrolnych w modelach EWD i procesów selekcji warto wspomnieć o jeszcze jednym problemie. Występowanie selekcji przy przyjmowaniu uczniów do szkół oznacza zwykle, że uczniowie o wyższych wynikach egzaminów trafiają do lepszych, efektywniej uczących szkół. Niestety dostępne metody statystyczne nie pozwalają dobrze kontrolować wpływu takiej selekcji - część wpływu, jaki na wyniki egzaminu „na wyjściu" wywierają działania szkół może być w takiej sytuacji przypisywana (w procesie estymacji modelu) oddziaływaniu zmiennych kontrolnych. Poszerzanie zakresu czynników kontrolowanych w modelu z jednej strony chroni więc przed „niesprawiedliwym" ocenianiem szkół, jednak z drugiej może utrudniać rozróżnianie od siebie szkół pracujących efektywnie i nieefektywnie (Ballou, Sanders i Wright, 2004; McCaffrey i in., 2003 s. 68-75; McCaffrey i in., 2004; OECD, 2008, s. 125-139; Raudenbush i Willms, 1995).

3.3. Metody estymacji modeli EWD i metody wyliczania oszacowań punktowych wskaźników EWD

Skupimy się tu na porównaniu ze sobą dwóch metod estymacji modeli EWD, oraz powiązanych z nimi metod estymacji wartości samych wskaźników EWD, które stosowane są w Polsce. Wspomniana wcześniej metoda regresji kawałkami liniowej, wykorzystywana przez pewien czas do estymacji jednorocznych modeli EWD gimnazjów, nie będzie rozpatrywana oddzielnie, gdyż z punktu widzenia interesujących nas własności nie różni się ona od zastosowania regresji MNK. Na świecie wykorzystywane są również inne metody, które pokrótce zostały już opisane w polskojęzycznych publikacjach na temat metody EWD (raporty), jednak nie będziemy się tu nimi zajmować, gdyż nigdy nie zostały one zastosowane w polskich warunkach.

Do wyliczania polskich wskaźników EWD używana jest jedna z dwóch metod:

1. Estymacje modelu regresji MNK i wyliczanie wskaźników EWD jako średniej z reszt tej regresji w ramach grup będących przedmiotem zainteresowania. Model w ogólności ma postać:

$$y_i = w^k(x_i, rok\_we_i, rok\_wy_i) + bz_i + r_i \qquad (12)$$

gdzie:

y_i wynik egzaminu „na wyjściu" *i*-tego ucznia;

x_i wynik egzaminu „na wejściu" *i*-tego ucznia;

rok_we_i rok zdawania egzaminu „na wejściu" przez *i*-tego ucznia;

rok_wy_i rok zdawania egzaminu „na wyjściu" przez *i*-tego ucznia;

z_i wektor wartości zmiennych kontrolnych charakteryzujących *i*-tego ucznia;

r_i reszta regresji (realizacja błędu losowego) *i*-tego ucznia;

$w^k(x_i, rok\_we_i, rok\_wy_i)$ wielomian *k*-tego stopnia opisujący zależność pomiędzy wynikami egzaminu „na wejściu" przeprowadzonego w roku rok_we_i a wynikami egzaminu „na wyjściu" przeprowadzonego w roku rok_wy_i;

b wektor parametrów związanych ze zmiennymi kontrolnymi charakteryzującymi uczniów.

Zakłada się przy tym, że błąd losowy ma rozkład normalny o wartości oczekiwanej równej zero i odchyleniu standardowym będącym parametrem modelu oraz że jego wartość oczekiwana i wariancja są nieskorelowane z wartościami przewidywanymi na podstawie modelu:

$$[\, r \mid w^k(x, rok\_we\,,\ rok\_wy) + bz] \sim \mathrm{N}(0, \sigma) \qquad (13)$$

Szacowanie parametrów funkcji $w^k(x_{ij}\ ,\ rok\_we_{ij}\ ,\ \ rok\_wy_{ij})$ następuje za pośrednictwem wprowadzenia do modelu efektów pierwszego rzędu związanych z kolejnymi potęgami (od zerowej do *k*-tej) zmiennej X, ich interakcji ze zmiennymi zero-jedynkowymi opisującymi przynależność obserwacji do grup wyróżnionych ze względu na kombinację wartości zmiennych rok_we i rok_wy oraz efekty pierwszego rzędu związane z tymi zmiennymi zero-jedynkowymi.

W praktyce, jako że metodą tą estymowane są wyłącznie modele jednoroczne, a uczniowie o toku kształcenia wydłużonym o więcej niż jeden rok są wykluczani z modelowania, parametry wielomianu $w^k(x_{ij}\ ,\ rok\_we_{ij}\ ,\ \ rok\_wy_{ij})$ szacowane są w konkretnym modelu tylko dla dwóch grup: uczniów o standardowej długości toku kształcenia (trzy lata w gimnazjach i liceach ogólnokształcących, cztery lata w technikach) i uczniów o toku kształcenia wydłużonym o jeden rok.

Wartość wskaźnika EWD dla dowolnie wybranej grupy uczniów może być następnie obliczona jako:

$$EWD_I = \sum_{i \epsilon I} r_i \qquad (14)$$

2. Estymacja modelu mieszanych efektów (modelu wielopoziomowego) z efektem losowym dla stałej regresji, związanym z przydziałem uczniów do szkół, a następnie wyliczanie wskaźników EWD jako bayesowskich predykcji a'posteriori (określanych też mianem Best Linear Uniased Predictors) realizacji tego efektu losowego. Model w ogólności ma postać:

$$y_{ij} = w^k\left(x_{ij}, rok\_we_{ij}, rok\_wy_{ij}\right) + \gamma z_{ij} + u_j + r_{ij} \qquad (15)$$

gdzie:

y_{ij} wynik egzaminu „na wyjściu" *i*-tego ucznia w *j*-tej szkole;

x_{ij} wynik egzaminu „na wejściu" *i*-tego ucznia w *j*-tej szkole;

rok_we_{ij} rok zdawania egzaminu „na wejściu" przez *i*-tego ucznia w *j*-tej szkole;

rok_wy_{ij} rok zdawania egzaminu „na wyjściu" przez *i*-tego ucznia w *j*-tej szkole;

z_{ij} wektor wartości zmiennych kontrolnych charakteryzujących *i*-tego ucznia w *j*-tej szkole;

u_j realizacja efektu losowego dla *j*-tej szkoły;

r_{ij} reszta indywidualna (realizacja błędu losowego z poziomu indywidualnego) *i*-tego ucznia w *j*-tej szkole;

$w^k(x_{ij}, rok\_we_{ij}, rok\_wy_{ij})$ wielomian *k*-tego stopnia opisujący zależność pomiędzy wynikami egzaminu „na wejściu" przeprowadzonego w roku rok_we_i a wynikami egzaminu „na wyjściu" przeprowadzonego w roku rok_wy_i;

γ wektor parametrów (efektów stałych) związanych ze zmiennymi kontrolnymi charakteryzującymi uczniów.

Zakłada się przy tym, że błąd losowy z poziomu indywidualnego r oraz efekt losowy dla stałej regresji u mają rozkłady normalne o wartości oczekiwanej równej zero i odchyleniach standardowych będących parametrami modelu, że obie te zmienne losowe (r i u) są nieskorelowane ze sobą nawzajem oraz ich wartości oczekiwane i wariancje są nieskorelowane z wartościami przewidywanymi na podstawie części stałej modelu:

$$[\, u \mid w^k(x, rok\_we, rok\_wy) + \gamma z \,] \sim \mathrm{N}(0, \sqrt{\tau}) \qquad (16)$$

$$[\, r \mid u, w^k(x, rok\_we, rok\_wy) + \gamma z \,] \sim \mathrm{N}(0, \sigma) \qquad (17)$$

Szacowanie parametrów funkcji $w^k(x_{ij}, rok\_we_{ij}, rok\_wy_{ij})$ następuje tak samo, jak w omówionym powyżej modelu MNK. W praktyce, jako że metoda ta stosowana jest do estymacji trzyletnich modeli EWD, parametry wielomianu szacowane są oddzielnie dla sześciu różnych grup (trzy roczniki egzaminu na wyjściu przecięte z podziałem na uczniów o standardowej długości toku kształcenia i uczniów o toku kształcenia wydłużonym o jeden rok).

Różnica w stosunku do modelu estymowanego MNK polega na włączeniu efektu losowego dla stałej regresji, związanego z przydziałem uczniów do szkół. Wartość realizacji tego efektu losowego dla danej szkoły u_j interpretujemy tu jako EWD tej szkoły. Jednocześnie należy pamiętać, że u_j nie są parametrami modelu i ich wartości nie są bezpośrednio określane

w procesie estymacji, w którym optymalizowana jest jedynie wartość wariancji efektu losowego (*τ*).

Dla tego typu prostego modelu mieszanych efektów bayesowskie przewidywanie realizacji efektu losowego u dla *j*-tej szkoły można przedstawić jako „ściągniętą" w kierunku zera (tj. przemnożoną przez współczynnik mniejszy od jedności) średnią różnic pomiędzy wartościami zmiennej zależnej a wartościami przewidywania wynikającego z części stałej modelu:

$$EWD_j = \hat{u}_j = \left(\frac{\tau}{\tau + \frac{\sigma^2}{n_j}} \right) \frac{1}{n_j} \sum_{i=1}^{n_j} \left(y_{ij} - \hat{y}_{ij}^f \right) \qquad (18)$$

gdzie:

$$\hat{y}_{ij}^f = w^k \left(x_{ij}, rok\_we_{ij}, rok\_wy_{ij} \right) + \gamma z_{ij} \qquad (19)$$

Zauważmy przy tym, że estymator największej wiarygodności dla realizacji efektu losowego w *j*-tej grupie ma postać:

$$\hat{u}_j^{ML} = \frac{1}{n_j} \sum_{i=1}^{n_j} \left(y_{ij} - \hat{y}_{ij}^f \right) \qquad (20)$$

Łatwo stwierdzić, że w obu metodach ogólna postać modelu jest bardzo podobna, różni je jednak to, że w pierwszej metodzie na etapie estymacji modelu regresji MNK w żaden sposób nie jest uwzględniane pogrupowanie uczniów. W efekcie nie otrzymujemy też z modelu regresji żadnych informacji na temat tego, jaką część wariancji zmiennej zależnej niewyjaśnianej przez efekty stałe (wyniki egzaminu „na wejściu", inne kontrolowane cechy ucznia) możemy przypisać różnicom między szkołami, a jaką zróżnicowaniu wewnątrz szkół. W efekcie z modelu regresji nie wynika, jaki powinien być „odpowiedni" sposób szacowania błędów standardowych wskaźników EWD (przyjęte rozwiązanie tego problemu przedstawione zostało w następnym podrozdziale).

Warto też zauważyć, że fakt pominięcia w modelu informacji o pogrupowaniu nie oznacza, że zakładamy brak wpływ pogrupowania (w szczególności związanego z przydziałem do szkół), niezależnie od uwzględnionych w modelu charakterystyk uczniów, na wyniki uzyskiwane przez nich na egzaminie „na wyjściu" (założenie, że taki wpływ istnieje leży przecież u podstaw zainteresowania wskaźnikami EWD). Jednak w związku z tym estymatory parametrów szacowanych w modelu są nieobciążone tylko wtedy, jeśli nie występuje związek pomiędzy liczbą obserwacji w grupie, a wpływem grupy na wyniki egzaminu „na wyjściu" (nie jest tak, że małe szkoły efektywniej, lub mniej efektywnie, uczą) oraz nie występują związki pomiędzy wpływem grupy na wyniki egzaminu a składem grupy ze względu na zmienne uwzględnione w modelu.

W przypadku drugiej metody już na etapie specyfikowania modelu regresji konieczne jest określenie, wpływ jakiego pogrupowania chcemy rozpatrywać. W wyniku estymacji uzyskamy informacje o tym, jaka część wariancji przypisanej losowej części modelu przypada na poziom międzygrupowy, a jaka na poziom wewnątrzgrupowy. Informacje te mogą następnie zostać wykorzystane do zmniejszenia błędu przewidywania realizacji efektów losowych szkół (Biecek, 2011; Robinson, 1999). Łatwozauważyć, że składnik σ^2/n_j opisuje wariancję estymatora największej wiarygodności $\hat{u}_j^{ML}$ na podstawie n_j obserwacji (a więc dokonania n_j losowań z rozkładu $r \sim \mathrm{N}(0, \sigma)$). Jeśli wnioskowanie o u_j oparte jest na dużej liczbie obserwacji, efekt „ściągnięcia" będzie minimalny, co odzwierciedla duże zaufanie, jakim możemy w takiej sytuacji obdarzyć estymator największej wiarygodności. Jeśli jednak

estymator największej wiarygodności obarczony jest dużą niepewnością, w przewidywaniu będziemy w większym stopniu posiłkować się informacją o wartości oczekiwanej u (pamiętając, że $u \sim \mathrm{N}(0, \sqrt{\tau})$) co oznacza przypisanie mniejszej wagi do $\hat{u}_j^{ML}$ i w efekcie ściągnięcie przewidywania w kierunku zera. Jednocześnie generalnie tym bardziej możemy ufać estymatorowi największej wiarygodności, im mniejsza jest wariancja indywidualnego (wewnątrzgrupowego) składnika błędu w stosunku do wariancji efektu losowego na poziomie międzygrupowym.

W modelach wykorzystywanych do wyliczania trzyketnich wskaźników EWD udział wariancji międzyszkolnej τ w całkowitej wariancji efektów losowych (tj. $\tau + \sigma^2$) wynosi od 8% - 9% w modelach dla gimnazjów poprzez 9% - 16% w modelach dla techników do 12% - 20% w modelach dla liceów ogólnokształcących. Przy tym w przypadku szkół kończących się maturą zdecydowanie większy udział wariancji międzyszkolnej występuje w modelach uwzględniających przedmioty przyrodnicze i matematykę. Dekompozycje efektów losowych w wyliczonych modelach trzyletnich skaźników EWD zestawione zostały w Aneksie F.

Jeśli chodzi o własności estymatorów parametrów opisujących część stałą modelu mieszanych efektów, są one nieobciążone nawet w sytuacji, gdy zachodzi związek pomiędzy liczebnością grupy a wpływem danej grupy na wyniki egzaminu „na wyjściu". Warto przy tym zauważyć, że jeżeli taki związek zachodzi, to rozkład przewidywań $\hat{u}_j$ w grupie analizowanych szkół będzie skośny i może znacząco odbiegać od rozkładu normalnego. Jednocześnie w modelu obecne jest założenie, że EWD nie jest związana ze składem grup (szkół) ze względu na zmienne uwzględnione w części stałej modelu.

Porównując oba przedstawione powyżej podejścia należy stwierdzić, że zaletą wykorzystania modeli mieszanych efektów i bayesowskich predykcji a'posteriori są przede wszystkim dobre własności metody szacowania wskaźników EWD i, o czym w następnym podrozdziale, ich błędów standardowych na podstawie modelu. Dodatkowo zniesione zostaje tu założenie o niezależności EWD i wielkości grupy. Wadą tego podejścia jest brak elastyczności. Dla każdego pogrupowania, które może interesować osobę dokonującą potem analiz konieczne jest wyestymowanie oddzielnego modelu, uwzględniającego wpływ konkretnego pogrupowania, aby możliwe było później obliczenie w odpowiedni sposób wartości wskaźników EWD. Czyni to niemożliwym wykorzystanie takiego podejścia w sytuacji, gdy chcemy zapewnić użytkownikowi wskaźników EWD maksymalną swobodę w doborze analizowanych grup, jak to ma miejsce w przypadku ewaluacji wewnątrzszkolnej. Problem ten nie występuje w sytuacji, gdy EWD wyliczana jest jako średnia z reszt modelu regresji MNK. Jednocześnie pierwsze z opisywanych podejść ma również tę zaletę praktyczną, że jest znacznie bardziej przejrzyste i łatwe do wytłumaczenia użytkownikom wskaźników. Z tych powodów zdecydowano, że pierwsze podejście wykorzystywane będzie przy wyliczaniu jednorocznych wskaźników EWD, skierowanych do szkół i mających służyć ich autoewaluacji. Z kolei do wyliczania powszechnie dostępnych na stronie internetowej trzyletnich wskaźników EWD szkół wykorzystywane będzie drugie podejście, charakteryzujące się lepszymi własnościami statystycznymi wyliczanych wskaźników.

Na koniec warto zaznaczyć, że w praktyce różnice w wartościach wskaźników EWD wynikające z zastosowania jednego lub drugiego podejścia są ogólnie rzecz biorąc niewielkie (Jakubowski, 2007). Wynikają one głównie z zastosowania „ściągnięcia", podczas gdy różnice w oszacowaniach

parametrów części stałej modelu są bardzo małe[6]. W związku z tym różnice dotyczą przede wszystkim szkół małych i/lub takich, w których EWD wyliczane jako średnia reszt regresji jest szczególnie wysokie lub niskie.

3.4. Szacowanie błędu standardowego wskaźników EWD

Jak już zostało wspomniane, w sytuacji, gdy wskaźniki EWD wyliczane są jako średnia reszt regresji z modelu regresji MNK, występuje pewna trudność w określeniu odpowiedniej metody szacowania ich błędów standardowych. Wynika to z faktu, że błąd standardowy wskaźnika EWD powinien być związany z wewnątrzszkolną (wewnątrzgrupową) wariancją błędu – inaczej mówiąc, z wariancją błędu z poziomu indywidualnego. Rzecz w tym, że składnik błędu losowego z estymowanego w ramach takiego podejścia modelu zawiera w sobie zarówno interesujący nas błąd indywidualny, jak i zróżnicowania międzygrupowe, a więc jego odchylenie standardowe nie może zostać bezpośrednio wykorzystane do szacowania błędów standardowych wskaźników EWD.

W takiej sytuacji możliwe jest obranie dwóch ścieżek postępowania. Po pierwsze, można starać się przeprowadzić globalną dekompozycję wariancji błędu z wyestymowanego modelu regresji pomiędzy poziom wewnątrzgrupowy i międzygrupowy. Po drugie, można chcieć wykorzystać informację o zróżnicowaniu reszt w ramach konkretnej grupy na potrzeby wyliczenia błędu standardowego EWD w odniesieniu do tej grupy. Choć to drugie podejście nie jest zbyt eleganckie formalnie, jako że w praktyce prowadzi do łamania założenia o homoscedatyczności reszt indywidualnych, jednak jego istotną zaletą jest to, że może zostać łatwo zastosowane dla dowolnie wybranej grupy i to bez odwoływania się do informacji o tym, jakie jest zróżnicowanie reszt w innych grupach, wyróżnionych w populacji na podstawie takiego samego kryterium. Przykładowo chcąc oszacować błąd standardowy dla grupy dziewcząt uczęszczających do klasy A w szkole X nie musimy analizować, jaka jest wariancja wewnątrzgrupowa reszt w całej populacji, jeśli rozpatrywać podział uczniów na grupy ze względu na przecięcie ze sobą kryterium przynależności do klas i płci. Tej istotnej zalety praktycznej nie posiada oczywiście podejście, w którym chcielibyśmy odwoływać się do globalnej dekompozycji wariancji błędu z wyestymowanego modelu regresji pomiędzy poziom wewnątrzgrupowy i międzygrupowy.

W sytuacji, gdy podstawowym celem jest danie użytkownikom wskaźników możliwości wyliczania ich dla dowolnie zdefiniowanych grup i bez dostępu do danych ogólnopolskich – a z taką sytuacją mamy do czynienia w przypadku jednorocznych wskaźników EWD – wykorzystanie informacji o zróżnicowaniu reszt w ramach konkretnej grupy staje się jedyną dostępną metodą szacowania błędów standardowych. Są one szacowane w sposób analogiczny do błędu standardowego średniej, zakładając, że analizowani uczniowie stanowią rodzaj „próby” z hipotetycznej hiperpopulacji:

$$bs(EWD_I) = \frac{D(r_{i\epsilon I})}{\sqrt{n_I}} \qquad (21)$$

gdzie $D(r_i \epsilon I)$ oznacza odchylenie standardowe reszt w ramach grupy I.

[6] Choć należy przy tym mieć na uwadze, że w sytuacji, gdy relacja pomiędzy wynikami egzaminu „na wejściu” i wynikami egzaminu „na wyjściu” modelowana jest wielomianem trzeciego czy piątego stopnia, nawet niewielka zmiana wartości parametrów może skutkować znacznymi różnicami w przebiegu przewidywania dla skrajnych wartości egzaminu „na wejściu”.

W przypadku, gdy wskaźniki EWD wyliczane są jako bayesowskie predykcje a'posteriori na podstawie wyników estymacji modelu mieszanych efektów, ich błędy standardowe[7] wyliczane są na podstawie globalnej dekompozycji wariancji, uzyskanej w wyniku estymacji modelu. Analogicznie jak w przypadku wartości punktowej EWD, jej błąd standardowy można przedstawić jako „ściągnięty" błąd standardowy estymatora największej wiarygodności:

$$bs(EWD_j) = bs(\hat{u}_j) = bs(\hat{u}_j^{ML}) \sqrt{\frac{\tau}{\tau + \frac{\sigma^2}{n_j}}} = \frac{\sigma}{\sqrt{n_j}} \sqrt{\frac{\tau}{\tau + \frac{\sigma^2}{n_j}}} \qquad (22)$$

gdzie:

σ odchylenie standardowe błędu losowego z poziomu indywidualnego;

τ wariancja efektu losowego;

n_j liczba uczniów w *j*-tej grupie (szkole).

Różnice w oszacowaniach błędów standardowych pomiędzy dwoma przedstawionymi metodami wynikają więc, po pierwsze ze zróżnicowania empirycznie odnotowywanej wariancji reszt w ramach wyróżnionych grup, które w przypadku szacowania błędów EWD na podstawie globalnej dekompozycji wariancji jest ignorowane (wszystkim grupom o takiej samej liczbie obserwacji przypisywane jest takie samo oszacowanie błędu standardowego), jednak ma wpływ na „lokalnie" wyliczane błędy standardowe oszacowań wskaźników EWD wyliczanych jako średnie z reszt regresji. Drugim, w praktyce mniej istotnym źródłem różnic jest większa efektywność bayesowskich predykcji a'posteriori, objawiająca się we wzorze w postaci „ściągnięcia". Warto przy tym pamiętać, że ten wzrost efektywności jest możliwy dzięki wprowadzeniu do modelu założenia o tym, że efekt losowy ma rozkład normalny.

3.5. Wskaźniki średnich wyników egzaminu „na wyjściu" i ich związek z EWD

W przypadku polskich trzyletnich wskaźników EWD, udostępnianych powszechnie za pośrednictwem strony internetowej, zastosowana została nowatorska metoda prezentacji, w ramach której zawarta została zarówno informacja o EWD szkoły, jak i o średnich wynikach egzaminu „na wyjściu" jej uczniów, z uwzględnieniem niepewności szacowania obu tych wartości. Zakłada się przy tym, że estymator EWD i estymator średniego wyniku „na wyjściu" są skorelowanymi ze sobą zmiennymi losowymi o rozkładach normalnych. W takiej sytuacji możliwe jest wyznaczenie ich rozkładu łącznego prawdopodobieństwa, który będzie dwuwymiarowym rozkładem normalnym, i wyznaczenie na jego podstawie jednoczesnego obszaru ufności dla wartości obu parametrów, uwzględniającego ich wzajemny związek (taki obszar ufności będzie miał kształt elipsy). Analogiczne rozwiązanie stosowane jest np. przy estymacji przedziałowej parametrów regresji wielokrotnej, do wyznaczania jednoczesnych przedziałów ufności dla kilku parametrów (Biecek, 2011: 15-16; Faraway, 2004: 34-36).

---

[7] Formalnie odchylenia standardowe rozkładów a'posteriori, o których tu mowa nie są błędami standardowymi, gdyż określenie to pochodzi z innego porządku teoretycznego – statystyki klasycznej, a nie bayesowskiej. Niemniej ponieważ są one wykorzystywane w sposób zupełnie analogiczny do błędów standardowych w celu określania niepewności szacowania, będziemy je tu dla wygody językowej określać tym mianem.

By możliwe było zastosowanie takiego podejścia konieczne jest jednak dysponowanie oszacowaniami punktowymi EWD i średniego wyniku egzaminu „na wyjściu", oszacowaniami ich błędów standardowych, a także korelacji pomiędzy nimi. Sposób wyliczania oszacowania punktowego EWD i jego błędu standardowego omówiono w poprzednim podrozdziale. Tutaj zajmiemy się kwestią średniego wyniku egzaminu „na wyjściu" oraz korelacji.

W latach 2009-2012 w celu wyliczenia tych wartości średni wynik „na wyjściu" oraz jego błąd standardowy szacowane były przy pomocy wielopoziomowego modelu pustego, tj. modelu mieszanych efektów, w którym wyniki egzaminu „na wyjściu" uzyskane przez uczniów przewidywane były przy pomocy jedynie stałej regresji i efektu losowego dla tej stałej, związanego z przydziałem uczniów do szkół. Estymator średniego wyniku „na wyjściu" wyliczane były następnie, analogicznie jak w przypadku EWD, jako bayesowskie predykcje a'posteriori. Problem stanowiło w tym przypadku wyliczenie korelacji, jako że szacowane wartości wskaźników EWD i średniego wyniku „na wyjściu" nie łączyła dobrze określona, formalna zależność. W roli tej wykorzystywana były korelacje, wyliczane oddzielnie w ramach każdej szkoły, pomiędzy resztami regresji a zmienną zależną w modelu, na podstawie którego szacowane były wskaźniki EWD. Trzeba jednak zaznaczyć, że wartości te nie mogą być interpretowane jako korelacje pomiędzy wyliczonymi wcześniej estymatorami EWD i średniego wyniku egzaminu „na wyjściu".

W związku z niedoskonałością pierwotnie zastosowanego podejścia, w 2013 r. wprowadzona została nowa metoda wyliczania wskaźnika średniego wyniku egzaminu „na wyjściu", w ramach której jest on szacowany na podstawie tego samego modelu, który służy do wyliczenia wskaźników EWD. W efekcie oba estymatory określone są we wspólnej przestrzeni i możliwe jest łatwe wyliczenie korelacji pomiędzy nimi. Wynik *j*-tej szkoły szacowany jest w tym podejściu jako:

$$w_j = \hat{y}_j = \hat{y}_j^f + EWD_j \qquad (23)$$

gdzie:

w_j średnik wynik egzaminu „na wyjściu" uczniów *j*-tej szkoły;

$\hat{y}_j$ średni przewidywany wynik egzaminu „na wyjściu" uczniów *j*-tej szkoły;

$\hat{y}_j^f$ średnie przewidywanie wynikające z części stałej modelu dla uczniów *j*-tej szkoły;

EWD_j oszacowanie EWD *j*-tej szkoły.

Warto zaznaczyć, że uzyskiwane w ten sposób oszacowania są niemal identyczne, jak w przypadku poprzedniej metody – wartość korelacji liniowej przekracza 0,99, a jeśli pominąć szkoły poniżej 30 uczniów, przekracza 0,999. Część różnic wynika przy tym z faktu, że podczas stosowaniu pierwotnej metody przy szacowaniu średnich wyników „na wyjściu" uwzględniani byli również uczniowie, dla których w bazach danych dysponowaliśmy informacją o wynikach egzaminu „na wyjściu", ale nie dysponowaliśmy informacją o ich wynikach egzaminu „na wejściu". W ramach zmienionej metody osoby takie oczywiście muszą zostać wykluczone.

Korzystając z założenia modelu mieszanych efektów o niezależności części stałej i części losowej modelu, wartość błędu standardowego dla wskaźnika średniego wyniku egzaminu „na wyjściu" można wyliczyć jako:

$$D^2(w_j) = D^2(\hat{y}_j^f) + D^2(EWD_j) \qquad (24)$$

Drugi składnik tego wyrażenia jest po prostu kwadratem błędu standardowego oszacowania EWD. Pozostaje rozpatrzeć pierwszy składnik – wariancję średniego przewidywania wynikającego z części

stałej modelu. Wariancja ta ma dwa źródła – niepewność co do oszacowania parametrów modelu regresji oraz, analogicznie jak w przypadku EWD, traktowanie uczniów, którzy znaleźli się w danej szkole jako próby spośród hipotetycznej hiperpopulacji uczniów, którzy mogli do niej trafić. W sytuacji, gdy liczba stopni swobody modelu wynosi od kilkuset tysięcy do ponad miliona niepewność związana z szacowaniem parametrów modelu jest na tyle mała, że może zostać pominięta. W efekcie możemy zapisać:

$$D^2(w_j) = \frac{D^2(\hat{y}_{ij})}{n_j} + D^2(EWD_j) \qquad (25)$$

a więc:

$$bs(w_j) = \frac{D(\hat{y}_{ij})}{\sqrt{n_j}} + bs(EWD_j) \qquad (26)$$

gdzie:

$D^2(\hat{y}_{ij})$ wariancja przewidywania w ramach *j*-tej szkoły;

n_j liczba uczniów w *j*-tej grupie (szkole).

Tak wyliczone błędy standardowe również nie odbiegają znacząco od błędów standardowych wyliczanych przy użyciu pierwotnej metody. Znaczne różnice występują za to, jeśli chodzi o wartości korelacji pomiędzy estymatorami EWD i średniego wyniku egzaminu „na wyjściu", które w ramach nowej metody można łatwo wyliczyć na podstawie udziału wariancji estymatora EWD w wariancji estymatora średniego wyniku egzaminu „na wyjściu":

$$cor(w_j, EWD_j) = \sqrt{\frac{D^2(EWD_j)}{D^2(w_j)}} \qquad (27)$$

3.6. Trendy w ramach okresów trzyletnich

W toku prac nad trzyletnimi wskaźnikami EWD pojawił się pomysł, aby zmodyfikować modele używane do ich wyliczania w ten sposób, aby dostarczały informacji również o tendencji zmiany efektywności nauczania w ramach okresu trzyletniego. Aby osiągnąć ten cel model uzupełniony został o dodatkowy efekt losowy, związany ze zmienną opisującą rok zdawania przez ucznia egzaminu „na wyjściu":

$$y_{ij} = w^k(x_{ij}, rok\_we_{ij}, rok\_wy_{ij}) + \gamma z_{ij} + u_{0j} + u_{tj}t_{ij} + r_{ij} \qquad (28)$$

gdzie:

u_{tj} realizacja efektu losowego dla trendu w *j*-tej szkole;

t_{ij} zmienna opisująca rok zdawania egzaminu „na wyjściu" przez *i*-tego ucznia w *j*-tej szkole;

Efekt losowy u_{tj} mówi w takim przypadku o liniowej dynamice zmian efektywności nauczania w *j*-tej szkole, w stosunku do uczniów zdających egzamin „na wyjściu" w kolejnych latach w ramach analizowanego trzyletniego okresu. Dla ułatwienia interpretacji wartości u_{tj} przyjęto następujące kodowanie zmiennej t_{ij}: -1 dla pierwszego rocznika uczniów; 0 dla drugiego rocznika uczniów; 1 dla trzeciego rocznika uczniów w ramach danego okresu trzyletniego. Z racji standaryzacji wyników egzaminu „na wyjściu" oddzielnie w ramach każdego rocznika, w modelu można założyć, że efekt stały związany ze zmienną t jest równy zero, tj. w skali populacyjnej nie występuje zmiana średnich wyników egzaminu.

W takim modelu EWD szkoły definiowane jest jako średnia z efektywności nauczania w odniesieniu do każdego z trzech kolejnych roczników absolwentów – przy czym każdy z roczników brany jest pod uwagę z taką samą wagą. Bayesowskie predykcje a'posteriori z takiego modelu muszą być tu wyliczone w sposób nieco bardziej skomplikowany, niż w przypadku prostego modelu z pojedynczym efektem losowym dla stałej regresji (Biecek, 2011: 157-159; Robinson, 1999), ale posiadają analogiczne właściwości. W szczególności wartości wskaźników EWD wyliczone na podstawie takiego modelu są niemal identyczne z wartościami wyliczanymi z modelu bez trendów. Zauważalne różnice mogą pojawiać się tu tylko w przypadku szkół, w których w poszczególnych latach występowały duże różnice w liczbie absolwentów. Informacja o trendzie prezentowana była dla szkoły w postaci stwierdzenia, że w ramach danego okresu trzyletniego występował istotny trend rosnący, istotny trend malejący, lub brak trendu istotnie statystycznie różniącego się od zera.

Choć modele z trendem stanowiły ciekawe rozwinięcie metodologii wyliczania trzyletnich wskaźników EWD, jednak ich praktyczna użyteczność została oceniona negatywnie. Stwierdzono bowiem, że niejednokrotnie informacja o występowaniu trendów wydaje się na pierwszy rzut oka niespójna ze zmianami położenia obszarów ufności w kolejnych latach – nawet gdy interpretacji dokonują eksperci zajmujący się na co dzień wskaźnikami EWD. Rozwikłanie takiej pozornej niespójności – rekonstrukcja prawdopodobnego poziomu efektywności nauczania poszczególnych roczników absolwentów – jest na podstawie tych informacji możliwe, jednak często stanowi rodzaj niełatwej matematyczno-logicznej łamigłówki. W związku z tym informacja o trendach nigdy nie została powszechnie udostępniona i ostatecznie zrezygnowano z wyliczania modeli EWD z estymacją trendu, na rzecz opisanych wcześniej, prostszych modeli.

3.7. „Latentne" wskaźniki EWD

Nierozwiązanym problemem polskich modeli EWD – jak z resztą również rozwiązań wdrożonych w innych krajach – jest nieuwzględnianie niepewności pomiarowej wykorzystywanych miar osiągnięć uczniów. Wyniki egzaminów traktowane są tu jako doskonale rzetelne, „bezpośrednie" miary poziomu umiejętności, choć w rzeczywistości są one obarczone, niekiedy znacznym, błędem losowym. Na dodatek modele psychometryczne, które wykorzystywane są do opisania sposobu, w jaki egzaminy mierzą umiejętności uczniów, a następnie do wyliczenia oszacowań poziomu tych umiejętności zakładają, że błędy oszacowań nie są takie same dla wszystkich uczniów, lecz zmieniają się wraz z poziomem umiejętności (egzaminy mierzą umiejętności w pewnych zakresach dokładniej, niż w innych zakresach). Nie odpowiada to dobrze założeniom modeli regresji, w których następnie wykorzystywane są uzyskane oszacowania poziomu umiejętności.

Przezwyciężenie tego problemu nie jest łatwe, wymaga bowiem wykorzystania do wyliczania wskaźników EWD modelu, który jednocześnie modelowałby sposób pomiaru umiejętności za pośrednictwem egzaminów i zależność pomiędzy tymi umiejętnościami, uwzględniając przy tym wpływ wybranego pogrupowania uczniów (w szczególności wpływ szkoły). Oznacza to konieczność odwołania się do modeli strukturalnych (SEM). W praktyce najprostszym modelem, jaki może być wykorzystany, jest szczególna postać dwupoziomowej regresji latentnej. Własności wskaźników EWD, które można uzyskać z takiego modelu zbadano na możliwie prostym przykładzie, jednorocznych modeli EWD gimnazjów 2012 w zakresie przedmiotów 1) humanistycznych i 2) matematyczno-przyrodniczych. W analizach uwzględniono tylko uczniów o trzyletnim toku kształcenia.

Dla tak zawężonej grupy analizowanych uczniów w każdym z dwóch modeli mamy po dwa konstrukty. Pierwszym są umiejętności ucznia „na wejściu" – mierzone sprawdzianem, a drugim umiejętności ucznia „na wyjściu" – mierzone odpowiednią częścią egzaminu gimnazjalnego. Jako model części

pomiarowej przyjęto dwuparametryczny model logistyczny (2PL), a dla zadań o kilku możliwych poziomach wykonania model Samejima Graded Response (Linden i Hambelton, 1997).

Ogólny model opisujący prawdopodobieństwo uzyskania przez ucznia co najmniej *g*-tego z możliwych poziomów rozwiązania *k*-tego zadania opisuje się w tym przypadku wzorem:

$$log\left(\frac{P(X_k \geq g)}{1 - P(X_k \geq g)}\right) = a_k^{ind}\theta^{ind} + a_k^{gr}\theta^{gr} - b_{kg} \qquad (29)$$

gdzie:

X_k poziom wykonania k-tego zadania;

θ^{gr} średni poziom umiejętności w grupie;

$\theta^{ind} = \theta - \theta^{gr}$ poziom umiejętności wycentrowany względem średniego poziomu umiejętności w grupie;

a_k^{gr} dyskryminacja *k*-tego zadania na poziomie międzygrupowym;

a_k^{ind} dyskryminacja *k*-tego zadania na poziomie wewnątrzgrupowym;

b_{kg} trudność *g*-tego poziomu wykonania *k*-tego zadania.

W celu adaptacji tej ogólnej postaci na potrzeby modelu EWD przyjęto dwa założenia. Po pierwsze, jako że w ramach modelowanego zjawiska rozpatrywane pogrupowanie nie ma wpływu na wyniki umiejętności „na wejściu", będą one mierzone wyłącznie na poziomie indywidualnym:

$$log\left(\frac{P(X_k \geq g)}{1 - P(X_k \geq g)}\right) = a_k^{ind}\theta_{we}^{ind} - b_{kg} \qquad (30)$$

gdzie indeksy „we" oznaczają, że chodzi o poziom umiejętności „na wejściu".

Po drugie, aby możliwe było traktowanie poziomu umiejętności „na wyjściu" z poziomu międzygrupowego jako EWD, konieczne było założenie, że dla każdego zadania mierzącego poziom umiejętności „na wyjściu" dyskryminacja zadania na poziomie międzygrupowym jest równa dyskryminacji zadania na poziomie indywidualnym:

$$log\left(\frac{P(X_k \geq g)}{1 - P(X_k \geq g)}\right) = a_k\left(\theta_{wy}^{ind} + \theta_{wy}^{gr}\right) - \tau_{kg} = a_k\theta_{wy} - b_{kg} \qquad (31)$$

gdzie indeksy „wy" oznaczają, że chodzi o poziom umiejętności „na wyjściu".

Część strukturalna modelu to latentna regresja na poziomie indywidualnym, w której umiejętności „na wejściu" przewidywane są ze względu na umiejętności „na wejściu" oraz – analogicznie do standardowo stosowanych modeli EWD – płeć i trzy zmienne opisujące posiadanie zaświadczenia o dysleksji:

$$\theta_{wy}^{ind} = b\theta_{we} + b_p płeć + b_{dg}dysl\_g + b_{ds}dysl\_s + b_{dgs}(dysl\_g \cdot dysl\_s) + \varepsilon \qquad (32)$$

Jako że mamy tu do czynienia z zależnością latentną, możemy oczekiwać, że związek pomiędzy umiejętnościami „na wejściu" i „na wyjściu" będzie liniowy.

Zauważmy przy tym, że jeśli do obu stron tego równania dodamy θ_{wy}^{gr}, to otrzymamy:

$$\theta_{wy} = b\theta_{we} + b_p płeć + b_{dg} dysl\_g + b_{ds} dysl\_s + b_{dgs}(dysl\_g \cdot dysl\_s) + \theta_{wy}^{gr} + \varepsilon \qquad (33)$$

co wyjaśnia, dlaczego średnie grupowe umiejętności „na wyjściu" θ_{wy}^{gr} możemy w takim modelu interpretować jako EWD.

Specyfikację modelu dopełniają założenia:

$\theta_{we} \sim N(0,1)$ (34) $\varepsilon \sim N(0,1)$ (36)

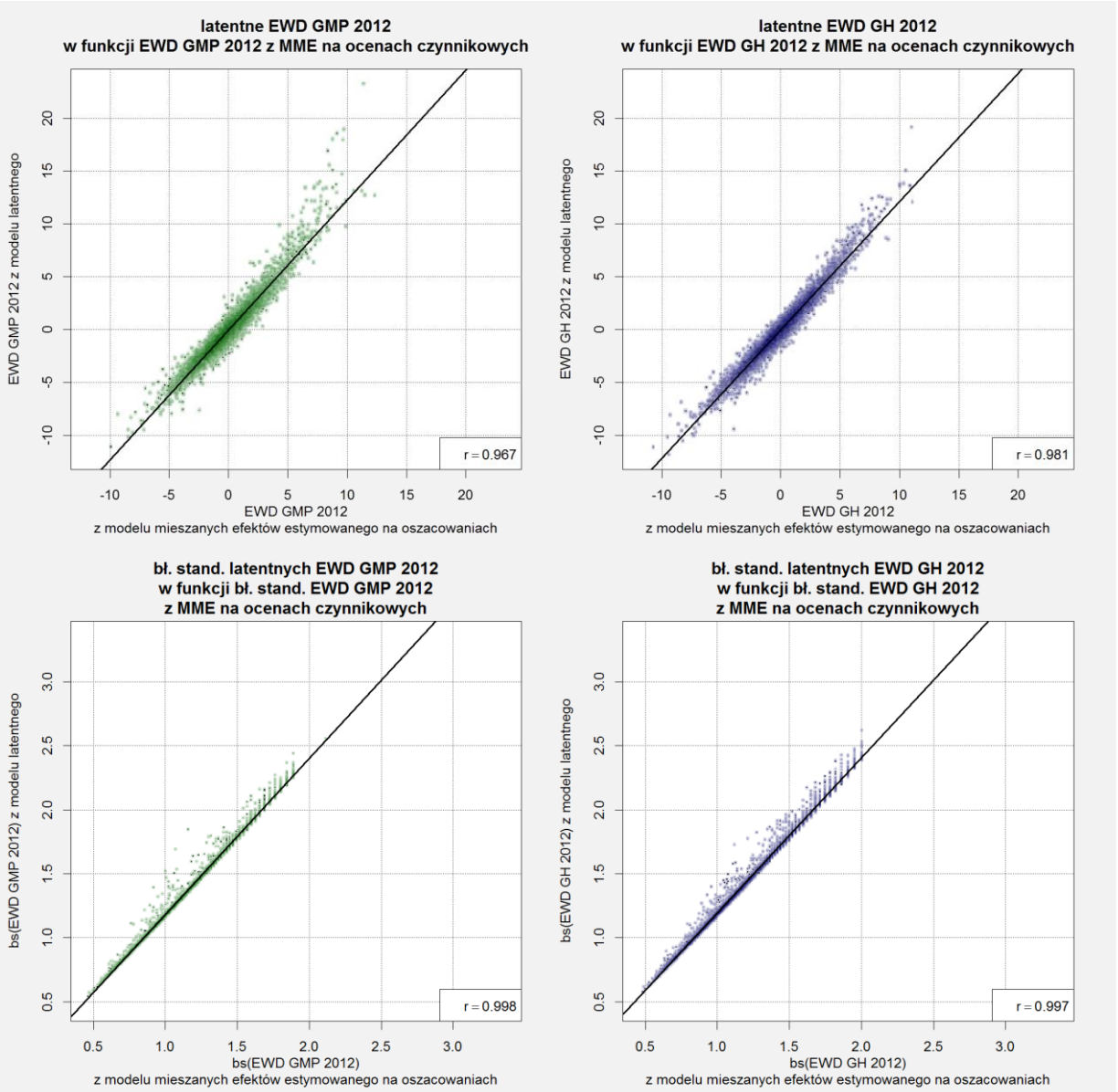
$\theta_{wy}^{ind} \sim N(0,\delta)$ (35) $\theta_{wy}^{gr} \sim N(0,\sqrt{\tau})$ (37)

Modele estymowane były w programie Mplus 7, metodą maksymalnej wiarygodności z optymalizacją funkcji wiarygodności w odniesieniu do pełnej macierzy danych (w ramach programu określaną skrótem MLR). Wskaźniki EWD uzyskano jako oceny czynnikowe dla zmiennej θ_{wy}^{gr}, wyliczone metodą EAP (Expected A'Posteriori) na podstawie wyestymowanych parametrów modelu.

Jakich własności można spodziewać się po takich „latentnych" oszacowaniach wskaźników EWD? Zastosowanie modelowania strukturalnego zamiast analiz prowadzonych na oszacowaniach pociąga zwykle za sobą dwie konsekwencje, z jednej strony wzrasta siła mierzonych związków (w odniesieniu do EWD oznacza to, że można spodziewać się wzrostu wariancji punktowych oszacowań EWD), jednak z drugiej strony, uwzględnienie w modelu niepewności związanej z częścią pomiarową modelu prowadzi też do zwiększenia się wartości błędów standardowych. Jako punkt odniesienia dla wyliczonych „latentnych" wskaźników EWD wykorzystane zostały oszacowania z wyestymowanych na tych samych danych modeli regresji mieszanych efektów, z tym że jako wyniki egzaminów „na wyjściu" i „na wejściu" wykorzystano w nich oszacowania EAP z wcześniej niezależnie od siebie wyestymowanych modeli IRT. Dodatkowo wyliczono także wskaźniki EWD jako średnie reszt z regresji MNK, nieuwzględniającej przydziału uczniów do szkół (również w tej regresji wykorzystano oszacowania EAP wyników egzaminów).

W wynikach przeprowadzonego porównania nieco zaskakuje bardzo duża zgodność oszacowań uzyskanych różnymi metodami. Korelacja liniowa pomiędzy wartościami wskaźników „latentnych" a wartościami wskaźników uzyskanych z regresji mieszanych efektów wynosi 0,967 w przypadku wskaźników matematyczno-przyrodniczych i aż 0,981 w przypadku wskaźników humanistycznych. Wyraźnie słabszy, ale wciąż bardzo silny jest również związek z wartościami wskaźników uzyskanych najprostszą metodą, jako średnie reszt regresji MNK. Wartości korelacji wynoszą tu odpowiednio 0,911 i 0,953. Dla pełnego obrazu, wartości korelacji pomiędzy wartościami wskaźników uzyskanych z modelu regresji mieszanych efektów i wartościami wskaźników wyliczonych jako średnie reszt regresji MNK to odpowiednio 0,944 (mat.-przyr.) i 0,965 (hum.).

Zgodnie z przewidywaniami, zróżnicowanie wartości wskaźników „latentnych" jest większe, niż wartości wskaźników uzyskanych z modelu regresji mieszanych efektów. Jednocześnie jednak, również zgodnie z oczekiwaniami, większe są błędy standardowe wskaźników „latentnych". Jeśli rozpatrzymy odsetek szkół, które można uznać za istotnie różne od zera, przy zastosowania dwustronnego testu na poziomie istotności 0,05, okazuje się, że wynosi on 0,372 dla wskaźników „latentnych" i 0,362 dla wskaźników wyliczonych na podstawie modelu regresji mieszanych efektów. Choć jest to bardzo zgrubne porównanie, można na jego podstawie powiedzieć, że z praktycznego punktu widzenia – oceny szkół i dokonywania porównań – wskaźniki „latentne" w bardzo niewielkim stopniu różnią się od wskaźników uzyskiwanych dotychczas stosowanymi metodami.



Rysunek 5. Związki pomiędzy „latentnymi" jednorocznymi wskaźnikami EWD gimnazjów 2012 (wyliczonymi przy pomocy modeli strukturalnych) i ich błędami standardowymi a jednorocznymi wskaźnikami EWD gimnazjów 2012 wyliczonymi na podstawie regresji mieszanych efektów, na oszacowaniach poziomu umiejętności z niezależnie estymowanych modeli IRT i błędami standardowymi tych wskaźników.

Należy przy tym stwierdzić, że estymacja modeli strukturalnych na danych tej wielkości, co wykorzystywane w modelach EWD (co najmniej kilkaset tysięcy jednostek obserwacji, kilkanaście do kilkudziesięciu zmiennych obserwowalnych powiązanych z każdym z konstruktów), jest bardzo wymagająca pod względem mocy obliczeniowej, a w przypadku stosowania metod wykorzystujących pełną macierz danych w całym procesie estymacji (a nie opierających się na dwustopniowej procedurze: 1) wyliczenie macierzy korelacji/kowariancji, 2) estymacja parametrów modelu w oparciu o macierz korelacji/kowariancji), również bardzo wymagająca pod względem dostępnej pamięci. Warto wspomnieć, że cały proces obejmujący wyestymowanie niezależnie od siebie trzech modeli IRT dla sprawdzianu 2009, części humanistycznej egzaminu gimnazjalnego 2012 i części matematyczno-przyrodniczej egzaminu gimnazjalnego 2012, a następnie wyestymowanie na podstawie uzyskanych

oszacowań umiejętności dwóch modeli regresji mieszanych efektów zajmuje około czterech godziny. Estymacja na tym samym komputerze latentnych modeli EWD zajęła odpowiednio nieco ponad dwa tygodnie w przypadku modelu matematyczno-przyrodniczego i ponad trzy tygodnie w przypadku modelu humanistycznego. Zauważmy przy tym, że ewentualne uwzględnienie w modelu uczniów o wydłużonym toku kształcenia, czy kilku roczników absolwentów prowadziłoby do dramatycznego wzrostu skomplikowania części pomiarowej modelu, stawiając pod znakiem zapytania możliwość jego estymacji.

Podsumowując, pomimo ogromnych zalet teoretycznych i elegancji formalnej, praktyczna użyteczność latentnych modeli EWD jest mocno problematyczna. Dużo większym kosztem – pod względem stopnia komplikacji wykorzystywanych modeli i czasu wyliczania wskaźników – otrzymujemy bowiem wskaźniki o własnościach bardzo zbliżonych do uzyskiwanych obecnie, znacznie prostszymi metodami.

4. Rekomendacje

W zakresie pożądanych kierunków zmian dotyczących systemu egzaminów zewnętrznych, sposobu konstruowania egzaminów oraz udostępniania wyników egzaminacyjnych i informacji o konstrukcji i własnościach arkuszy egzaminacyjnych, można na podstawie doświadczeń zebranych przy modelowaniu wskaźników EWD sformułować następujące rekomendacje:

- Zwiększenie poziomu trudności polskich egzaminów, w szczególności zaś sprawdzianu w klasie VI szkoły podstawowej, wpłynęłoby pozytywnie na ich własności psychometryczne i pozwoliło uniknąć występowania efektu sufitowego, ograniczającego trafność wskaźników EWD w odniesieniu do szkół uczących uczniów o bardzo wysokim poziomie umiejętności.

- Automatyczne przypisywanie laureatom konkursów przedmiotowych maksymalnych wyników egzaminu prowadzi do ograniczenia trafności wskaźników EWD w odniesieniu do szkół, które przyjęły dużą liczbę laureatów i/lub laureaci stanowią znaczną część w gronie ich absolwentów. Z punktu widzenia ewaluacyjnej roli egzaminów zewnętrznych laureaci powinni rozwiązywać egzaminy tak, jak wszyscy uczniowie. Niestety w takim przypadku, przy zachowaniu uprzywilejowanej pozycji laureatów w procesie rekrutacji do szkół na kolejnym etapie kształcenia, należy spodziewać się problemów z występowanie obniżonej motywacji laureatów podczas egzaminu. Niemniej z punktu widzenia modelowania EWD adekwatne uwzględnienie laureatów w modelach EWD byłoby w takiej sytuacji łatwiejsze niż obecnie.

- Z punktu widzenia wykorzystania wyników egzaminów do konstrukcji wskaźników EWD bardzo pożądane byłoby uwzględnienie w założeniach konstruowania polskich egzaminów, że mają one służyć nie tylko ocenie tego, w jakim stopniu uczeń opanował treści nauczania kończonego właśnie etapu kształcenia, ale również, że ich wyniki mają w przyszłości stanowić punkt odniesienia do oceny, jakie (relatywne) postępy poczyni on w na następnym etapie edukacji. Inaczej mówiąc, należy tak myśleć o konstruowaniu egzaminów przeprowadzanych na zakończenie kolejnych etapów kształcenia, by mierzyły one jak najbardziej zbliżone konstrukty. Jednocześnie z punktu widzenia modelowania EWD bynajmniej nie jest konieczne dążenie do pionowego zrównania egzaminów (wyrażenia wyników egzaminów po kolejnych etapach kształcenia na tej same skali).

- Publikowana przez Centralną Komisję Egzaminacyjną dokumentacja struktury arkuszy egzaminacyjnych powinna uwzględniać informacje pozwalające łatwo określić poprawną formę modeli psychometrycznych, które inne instytucje lub badacze, chcieliby wykorzystać do skalowania wyników egzaminacyjnych. W szczególności niezwykle istotne byłoby określenie, które kryteria oceny można traktować jako niezależne od siebie, a które nie mogą być traktowane w ten sposób.

- Należy wspierać wszelkie inicjatywy mające na celu powstawanie ogólnopolskich baz wyników egzaminacyjnych (łączących dane z poszczególnych OKE) oraz ogólniej integrację baz wyników egzaminacyjnych z Systemem Informacji Oświatowej w tym z Rejestrem Szkół i Placówek Oświatowych.

W zakresie kierunków dalszych prac nad rozwojem metodologii EWD w Polsce, można na podstawie dotychczas zebranych doświadczeń sformułować następujące rekomendacje:

- Należy prowadzić dalsze prace nad opracowaniem i wdrożeniem doskonalszego modelu skalowania wyników matury. W szczególności rozwiązana powinna zostać kwestia uwzględniania w modelu faktu wybierania przez zdających tematu wypracowania (wypowiedzi pisemnej) w arkuszach z języka polskiego, WOSu oraz historii.

- Należy kontynuować prace nad włączeniem wieku jako zmiennej kontrolnej do modeli EWD, zwłaszcza w odniesieniu do gimnazjów.

- Należy kontynuować analizy mające na celu określenie formy, w jakiej średnie wyniki egzaminu „na wejściu" uczniów szkoły wpływając na wyniki egzaminu „na wyjściu" oraz stwierdzenie, czy występowanie takiej zależności ma związek z selektywnością przy naborze do szkół.

Aneks A: Kalendarium rozwoju metod szacowania polskich wskaźników EWD

Wskaźniki EWD gimnazjów

lata	wskaźniki jednoroczne	wskaźniki trzyletnie
2005-2008	■ regresja (dwoma) kawałkami liniowa; ■ wyniki surowe egzaminów; ■ 2 wskaźniki: hum. i mat.-przyr.; ■ bez uczniów o wydł. toku kształcenia; ■ wyniki implementowane w *Kalkulatorze* EWD - arkuszu Excela (udostępnianie arkuszy kontynuowano do 2010 r.);	■ nie wyliczano;
2009-2011	■ regresja MNK z wielomianem; ■ wyniki implementowane w *Kalkulatorze EWD Plus* – zewnętrznej aplikacji do analiz;	■ regresja mieszanych efektów z wielomianem; ■ wyniki znormalizowane ekwikwantylowo (skala 100;15); ■ 2 wskaźniki: hum. i mat.-przyr.; ■ bez uczniów o wydł. toku kształcenia; ■ opublikowano też wskaźniki dla okresu 2008-2006; ■ powszechnie dostępne za pośrednictwem strony internetowej http://gimnazjum.ewd.edu.pl/
2012	■ wyniki znormalizowane ekwikwantylowo (skala 100;15); ■ 6 wskaźników (nowa formuła egz. gimn.): pol., hist.-WOS, mat., przyr., hum. i mat.-przyr.;	■ wyniki wyskalowane 2PL/SGR (skala 100;15); ■ z uczniami o toku kształcenia wydłużonym o 1 rok;
2013	■ wyniki implementowane w *Kalkulatorze EWD 100* – zewnętrznej aplikacji do analiz;	■ zmiana sposobu szacowania śr. wyniku szkoły na wyjściu;

Wskaźniki EWD szkół kończących się maturą

rok	wskaźniki jednoroczne	wskaźniki wieloletnie
2010	■ nie wyliczano;	■ wskaźniki jednoroczne; ■ regresja mieszanych efektów z wielomianem; ■ wyniki wyskalowane 2PL/SGR (skala 100;15); ■ 2 wskaźniki: hum. i mat.-przyr.; ■ bez uczniów o wydł. toku kształcenia; ■ powszechnie dostępne za pośrednictwem strony internetowej http://matura.ewd.edu.pl;
2011		■ wskaźniki dwuletnie; ■ 3 wskaźniki: hum., mat. i mat,-przyr.; ■ z uczniami o toku kształcenia wydłużonym o 1 rok;
2012		■ wskaźniki trzyletnie; ■ 4 wskaźniki: pol., hum., mat. i mat,-przyr.;
2013	■ regresja MNK z wielomianem; ■ wyniki znormalizowane ekwikwantylowo (skala 100;15); ■ wskaźnik mat. poz. podst.; ■ wyniki implementowane w *Kalkulatorze 100* – zewnętrznej aplikacji do analiz;	■ zmiana sposobu szacowania śr. wyniku szkoły na wyjściu;

Aneks B: Struktura polskich egzaminów zewnętrznych

Opisy zawierają fragmenty tekstu z informatorów Centralnej Komisji Egzaminacyjnej.

Sprawdzian:

Test wiedzy ogólnej, bez podziału na części. Składa się z 20 zadań zamkniętych punktowanych 0-1 i 5 zadań otwartych (części z nich przypisanych jest kilka różnych kryteriów oceny). Łącznie na egzaminie można uzyskać 40 punktów.

Egzamin gimnazjalny:

Egzamin ma formę pisemną. Przystąpienie do egzaminu jest warunkiem ukończenia gimnazjum, ale nie określa się minimalnego wyniku, jaki zdający powinien uzyskać, toteż egzaminu nie można nie zdać. Egzamin gimnazjalny składa się z trzech części: humanistycznej, matematyczno-przyrodniczej i części dotyczącej języka obcego nowożytnego, przy czym na potrzeby wyliczania wskaźników EWD brane są pod uwagę wyniki tylko z dwóch pierwszych części.

Część humanistyczna składa się z zadań z zakresu języka polskiego oraz zadań z zakresu historii i wiedzy o społeczeństwie. Zadania z języka polskiego mogą mieć formę zamkniętą lub otwartą. Wśród zadań otwartych z języka polskiego znajduje się dłuższa wypowiedź pisemna. Zadania z historii i wiedzy o społeczeństwie mają formę zamkniętą.

Część matematyczno-przyrodnicza składa się z zadań z zakresu matematyki oraz zadań z zakresu przedmiotów przyrodniczych: biologii, chemii, fizyki i geografii. Zadania z matematyki mają formę zamkniętą lub otwartą. Zadania z przedmiotów przyrodniczych mają formę zamkniętą.

Do 2011 r. za rozwiązanie każdej z części egzaminu gimnazjalnego można było uzyskać od 0 do 50 punktów. Od 2012 r. (tzw. *nowa formuła* egzaminu gimnazjalnego) wyniki podawane są zdającym w podziale na 4 zakresy (nie uwzględniając części językowej): język polski, historia i wiedza o społeczeństwie, matematyka, przedmioty przyrodnicze. Od 2012 r. suma punktów możliwych do uzyskania w ramach każdego zakresu nie jest sztywno określona i może być różna w poszczególnych latach (w latach 2012-2013 maksymalna możliwa do zdobycia w ramach poszczególnych zakresów liczba punktów wahała się od 28 do 33).

Matura (od 2010 r.):

Egzamin maturalny nie jest obowiązkowy, co oznacza, że każdy absolwent szkoły ponadgimnazjalnej samodzielnie podejmuje decyzję o przystąpieniu do niego. Egzamin maturalny jest przeprowadzany z przedmiotów obowiązkowych oraz przedmiotów dodatkowych i składa się z części ustnej oraz z części pisemnej. Aby zdać egzamin maturalny, absolwent musi otrzymać co najmniej 30% punktów możliwych do uzyskania z egzaminu z każdego przedmiotu obowiązkowego w części ustnej i w części pisemnej.

Od 2010 r. do przedmiotów obowiązkowych zaliczają się: język polski na poziomie podstawowym, matematyka na poziomie podstawowym oraz język nowożytny (wybrany przez ucznia) na poziomie podstawowym. Oprócz tego zdający może zdawać również wybrane spośród 24 przedmiotów dodatkowych na poziomie podstawowym albo rozszerzonym (z wyjątkiem polskiego i matematyki, które jako przedmioty dodatkowe można zdawać tylko na poziomie rozszerzonym). Wśród tych 24 przedmiotów 6 to języki obce (angielski, francuski, hiszpański, niemiecki, rosyjski, włoski), a 4 języki mniejszości narodowych lub etnicznych (białoruski, litewski, ukraiński, kaszubski). Pozostałe przedmioty dodatkowe to: biologia, chemia, filozofia, fizyka i astronomia, geografia, historia, historia muzyki, historia sztuki, informatyka, język łaciński i kultura antyczna, język polski, matematyka, wiedza o społeczeństwie, wiedza o tańcu.

Przy wyliczaniu wskaźników EWD brane są pod uwagę wyłącznie wyniki z części pisemnej egzaminu maturalnego i tylko z następujących przedmiotów: biologia, chemia, fizyka i astronomia, geografia, historia, informatyka, język polski, matematyka, wiedza o społeczeństwie.

Arkusze egzaminacyjne z różnych przedmiotów, jak również z poziomu podstawowego i rozszerzonego różnią się od siebie strukturą. Arkusze mogą zawierać zadania zamknięte i otwarte. Z reguły arkusze na poziomie rozszerzonym zawierają więcej zadań otwrtych. W trzech typach arkuszy: z języka polskiego na poziomie podstawowym i rozszerzonym oraz wiedzy o społeczeństwie na poziomie rozszerzonym występują zadania polegające na przygotowaniu wypowiedzi pisemnej na jeden z dwóch tematów (zdający sam wybiera jeden z tematów). Dokładne informacje o strukturze arkuszy można uzyskać z informatorów dostępnych na stronie internetowej Centralnej Komisji Egzaminacyjnej.

Aneks C: Zadania usunięte z modeli skalowania ze względu na niską dyskryminację

Sprawdzian (w latach 2006-2010):

- 2010: zad. 11;

Egzamin gimnazjalny (w latach 2005-2013):

- 2010: część mat.-przyr.: zad. 13, zad. 22;
- 2011: część hum.: zad. 14; część mat.-przyr.: zad. 33_1, zad. 33_2;
- 2012: przedmioty przyr.: zad. 1, zad. 14;
- 2013: język polski: zad. 4, zad. 16;

Matura (w latach 2010-2013):

- 2010: historia poz. podst.: zad. 24; WOS poz. podst.: zad. 6, zad. 11; chemia poz. podst.: zad. 10; geografia poz. podst.: zad. 10;
- 2011: WOS poz. podst.: zad. 8; WOS poz. rozsz.: zad. 6, zad. 14; fizyka poz. podst.: zad. 2, zad. 6;
- 2012: WOS poz. podst.: zad. 2, zad. 17; fizyka poz. podst. zad. 3, zad. 7;
- 2013: WOS poz. rozsz. zad. 21; fizyka poz. podst.: zad. 8;

Aneks D: „Linktest" (resztowy)

Klasyczny „linktest" (Cameron i Trivedi, 2010) jest testem diagnostycznym, pozwalającym na ocenę, czy model regresji może zostać udoskonalony poprzez włączenie do niego jako dodatkowego predyktora nieliniowego przekształcenia którejś z już wykorzystywanych zmiennych niezależnych, lub ich kombinacji liniowej (przy czym co do zasady stosuje się to pierwsze rozwiązanie).

Dla regresji postaci:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ \ldots \ + b_k x_{ki} + r_i \qquad \text{(A1)}$$

„linktest" przeprowadza się sprawdzając istotność współczynnika b_{l2} w regresji:

$$y_i = b_{l0} + b_{l1}\hat{y}_{Xi} + b_{l2}\hat{y}_{Xi}^2 + r_{li} \qquad \text{(A2)}$$

gdzie:

$$\hat{y}_{Xi} = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ \ldots \ + b_k x_{ki} \qquad \text{(A3)}$$

Nieistotność współczynnika sugeruje, że obecna forma modelu nie wymaga zmiany. Istotność, że wskazana jest zmiana postaci modelu poprzez rozszerzenie go o nieliniowe przekształcenie któregoś z predyktorów (lub ich kombinacji liniowej).

Możliwe jest też stosowanie nieco innych form testu, w których stosuje się odmienne nieliniowe przekształcenie $\hat{y}_{Xi}$ związane z parametrem b_{l2}, niemniej przekształcenie przez podniesienie do drugiej potęgi jest wykorzystywane najczęściej.

W przypadku modeli EWD jedyną zmienną, co do której zakładamy możliwość wprowadzania do modelu jej przekształconej nieliniowo postaci jest wynik ucznia „na wejściu". Rodzaj dozwolonych przekształceń zawężamy przy tym do zwiększenia o jeden stopnia związanego z tą zmienną wielomianu. W związku z tym rozsądne wydaje się zastosowanie takiej modyfikacji „linktestu", która pozwalałaby ocenić zasadność rozszerzania modelu o następny stopień wielomianu zmiennej dla wyniku sprawdzianu, nie biorąc jednocześnie pod uwagę skutków ewentualnego nieliniowego przekształcenia innych zmiennych niezależnych obecnych w modelu.

Dla modelu EWD estymowanego jako model mieszanych efektów[8] postaci:

$$y_{ij} = \gamma_{00} + \gamma_{11}x_{ij} + \ \ldots + \gamma_{1k}x_{ij}^{k} + \gamma z_{ij} + u_{oj} + r_{ij} \qquad \text{(A4)}$$

gdzie:

y_{ij} wynik egzaminu „na wyjściu" dla *i-tego* ucznia *j*-tej szkoły,
x_{ij} wynik egzaminu „na wejściu" dla *i-tego* ucznia *j*-tej szkoły,
z_{ij}- wektor wartości zmiennych wyjaśniających niezwiązanych z wynikiem egzaminu „na wejściu" (płeć, dysleksja) dla *i-tego* ucznia *j*-tej szkoły,
γ współczynniki związane z efektami stałymi,
u_{0j} wartość realizacji efektu losowego dla *j*-tej szkoły,
r_{ij} błąd losowy dla *i-tego* ucznia *j*-tej szkoły.

Jako podstawę do przeprowadzenia zmodyfikowanego „linktestu" należy wziąć resztę cząstkową (partial residual) przy przewidywaniu y_{ij} ze względu na z_{ij} i u_{0j}, ale z pominięciem efektów związanych z x_{ij} (i jego wyższymi potęgami):

$$r_{ijY|Z,U;X} = y_{ij} - \hat{y}_{ijZ,U;X} = y_{ij} - \gamma_{-s}z_{ij} - u_{oj} \qquad \text{(A5)}$$

Wartości takich reszt cząstkowych możemy przewidywać ze względu na sumę efektów związanych z x_{ij} (i jego wyższymi potęgami) z pierwotnego modelu oraz nieliniowe przekształcenie takiej sumy (zgodnie z najczęściej stosowanym rozwiązaniem, posługujemy się przekształceniem kwadratowym):

$$r_{ijY|Z,U;X} = \beta_{lr0} + \beta_{lr1}\hat{y}_{ijX;Z,U} + \beta_{lr2}\hat{y}^{2}_{ijX;Z,U} + r'_{ij} \qquad \text{(A6)}$$

gdzie:

$$\hat{y}_{ijX;Z,U} = \gamma_0 + \gamma_{11}x_{ij} + \ \ldots + \gamma_{1k}x_{ij}^{k} \qquad \text{(A7)}$$

Analogicznie do oryginalnej wersji „linktestu" decyzję podejmujemy na podstawie istotności współczynnika regresji związanego z nieliniowym przekształceniem przewidywania, tutaj β_{lr2}. Nieistotność współczynnika sugeruje, że obecny stopień wielomianu w wystarczająco dobry sposób opisuje zależność pomiędzy wynikami egzaminu „na wejściu" a wynikami egzaminu „na wyjściu". Istotność współczynnika sugeruje, że wprowadzenie do modelu wyższego stopnia wielomianu powinno wyraźnie polepszyć jakość przewidywania.

W przypadku modeli jednorocznych estymowanych MNK w równaniach (A4) i (A5) pomijane są składniki związane z przewidywaniem realizacji efektu losowego dla szkoły u_{0j}.

[8] W przypadku modelu regresji MNK opisaną procedurę stosuje się odpowiednio, z tym że w postępowaniu pomija się efekt losowy dla stałej regresji u, który w takim modelu nie występuje.

Aneks E: Wyniki procedury wyboru stopnia wielomianu

Jednoroczne modele EWD gimnazjów

rok	l. uczn. w modelu hum.	l. uczn. w modelu mat.-przyr.	stopień	część hum. (j. pol i hum.)		część mat-przyr.	
				monot.	linktest	monot.	linktest
2013	363 102	363 056	2	+	-	+	+
			3	-	-	-	+
			4	-	+	-	+
			5	+	+	+	+
2012	377 912	377 817	2	+	-	+	-
			3	+	+/-	-	+
			4	+/-	+	+/-	+
			5	+	+	+	+
2011	392 393	392 326	2	+	-	+	-
			3	+	+	-	-
			4	-	+	-	+
			5	+	+	+	+
2010	409 149	408 963	2	+	-	+	-
			3	+	+/-	-	+
			4	-	+	-	+
			5	+	+	+	+
2009	432 580	432 452	2	+	-	+	-
			3	+	+	-	+/-
			4	-	+	-	+
			5	+	+	+	+

Wyróżniono wybrany stopień wielomianu.

W kolumnach 'monot.':

'+' oznacza zależność monotoniczną we wszystkich analizowanych grupach;

'+/-' oznacza niewielkie odstępstwa od monotoniczności;

'-' oznacza pozostałe sytuacje.

W kolumnach 'linktest':

'+' oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,05 (we wszystkich analizowanych grupach);

'+/-' oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,01 (we wszystkich analizowanych grupach);

'-' oznacza pozostałe sytuacje.

Jednoroczne modele EWD gimnazjów - kontynuacja

rok	l. uczn. w modelu j.pol.	l. uczn. w modelu hum.	stopień	test z j. pol. monot.	test z j. pol. linktest	test hum. monot.	test hum. linktest
2013	363 105	363 107	2	+	-	+	+
			3	-	-	-	+
			4	-	+	-	+
			5	+	+	+	+
2012	377 912	377 915	2	+	-	+	-
			3	+	-	+/-	+/-
			4	+/-	+	-	+
			5	+	+	+	+

rok	l. uczn. w modelu mat.	l. uczn. w modelu przyr.	stopień	test mat. monot.	test mat. linktest	test przyr. monot.	test przyr. linktest
2013	363 057	363 059	2	+\-	-	+	+
			3	-	+	-	+
			4	-	+	-	+
			5	+	+	+	+
2012	377 817	377 822	2	+	-	+	-
			3	-	-	+/-	+
			4	+/-	+	+/-	+
			5	+	+	+	+

Wyróżniono wybrany stopień wielomianu.

W kolumnach 'monot.':

'+' oznacza zależność monotoniczną we wszystkich analizowanych grupach;

'+/-' oznacza niewielkie odstępstwa od monotoniczności;

'-' oznacza pozostałe sytuacje.

W kolumnach 'linktest':

'+' oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,05 (we wszystkich analizowanych grupach);

'+/-' oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,01 (we wszystkich analizowanych grupach);

'-' oznacza pozostałe sytuacje.

Trzyletnie modele EWD gimnazjów

okres	l. uczn. w modelu hum.	l. uczn. w modelu mat.-przyr.	stopień	część hum. (j. pol i hum.) monot.	część hum. (j. pol i hum.) linktest	część mat-przyr. monot.	część mat-przyr. linktest
2013-2011 (z wydł.)	1 124 690	1 124 409	2	+	-	+	-
			3	+	-	-	+/-
			4	-	+	-	+
			5	+/--	+	+/-	+
2012-2010 (z wydł.)	1 173 178	1 172 830	2	+	-	+	-
			3	+/--	+	-	-
			4	-	+	-	+
			5	-	+	+	+
2011-2009 (z wydł.)	1 230 650	1 230 267	2	+	-	+	-
			3	+/-	+/-	-	-
			4	-	+	-	+
			5	+/-	+	+/--	+
2010-2008	1 246 926	1 246 528	2	+	-	+	-
			3	+	+/-	-	+/-
			4	-	+	-	+
			5	+	+	+	+
2009-2007	1 313 671	1 313 350	2	+	-	+	-
			3	+	+/-	-	-
			4	-	+	-	+
			5	+	+	+	+
2009-2006	1 379 895	1 379 695	2	+	-	+	-
			3	+	+/-	-	-
			4	-	+	-	+
			5	+	+	+	+

Wyróżniono wybrany stopień wielomianu.

W kolumnach 'monot.':

'+' oznacza zależność monotoniczną we wszystkich analizowanych grupach;

'+/-' oznacza, że występuje niewielka niemonotoniczność, ale tylko w jednym roczniku i tylko w grupie uczniów o toku kształcenia wydłużonym o rok;

'+/--' oznacza, że monotoniczności są niewielkie, występują tylko w grupie dla uczniów o toku kształcenia wydłużonym o rok, ale w więcej niż jednym roczniku;

'-' oznacza pozostałe sytuacje.

W kolumnach 'linktest':

'+' oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,05 (we wszystkich analizowanych grupach);

'+/-' oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,01 (we wszystkich analizowanych grupach);

'-' oznacza pozostałe sytuacje.

Jednoroczne modele EWD szkół kończących się maturą

rok	l. uczn. w modelu LO	l. uczn. w modelu T	stopień	licea ogólnokształcące		technika	
				monot.	linktest	monot.	linktest
2013	182 102	96 273	2	+	-	+	-
			3	-	-	-	-
			4	-	+	-	+
			5	+/-	+	-	+
2012	190 851	102 507	2	+	-	+	-
			3	-	+	-	+
			4	-	+	-	+
			5	+	+	+	+
2011	197 032	101 559	2	+	-	+	-
			3	-	+	-	+
			4	-	+	-	+
			5	+	+	+	+
2010	205 009	95 829	2	+	-	+	-
			3	-	+	-	+
			4	+/-	+	+/-	+
			5	+	+	+	+

Wyróżniono wybrany stopień wielomianu.

W kolumnach ‘monot.’:

‘+’ oznacza zależność monotoniczną we wszystkich analizowanych grupach;

‘+/-’ oznacza niewielkie odstępstwa od monotoniczności;

‘-’ oznacza pozostałe sytuacje.

W kolumnach ‘linktest’:

‘+’ oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,05 (we wszystkich analizowanych grupach);

‘+/-‘ oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,01 (we wszystkich analizowanych grupach);

‘-‘ oznacza pozostałe sytuacje.

Trzyletnie modele EWD szkół kończących się maturą

okres	l. uczn. w modelu j.pol.	l. uczn. w modelu hum.	stopień	j. polski		przedmioty hum.	
				monot.	linktest	monot.	linktest
2013-2011 LO	563 351	563 351	1	+	-	+	-
			2	+	+	+	+
			3	+	+	+	+
			4	-	+	-	+
2013-2011 technika	294 688	294 688	1	+	-	+	-
			2	+	+	+	+
			3	+/-	+	+/-	+
			4	-	+	-	+
2012-2010 LO	592 816	592 847	1	+	-	+	-
			2	+	+	+	+
			3	-	+	+/-	+
			4	-	+	-	+
2012-2010 technika	300 064	300 077	1	+	-	+	-
			2	+	+	+	+
			3	+/-	+	+/-	+
			4	-	+	-	+

okres	l. uczn. w modelu mat.	l. uczn. w modelu przyr.	stopień	matematyka		przedmioty mat.-przyr.	
				monot.	linktest	monot.	linktest
2013-2010 LO	562 696	562 696	1	+	-	+	-
			2	+	+/-	+	+
			3	-	+	-	+
			4	-	+	-	+
2013-2010 technika	294 781	294 781	1	+	-	+	-
			2	+	+/-	+	+
			3	-	+	-	+
			4	-	+	-	+
2012-2010 LO	592 540	592 640	1	+	-	+	-
			2	+	-	+	-
			3	-	+	-	+
			4	-	+	-	+
2012-2010 technika	299 780	299 853	1	+	-	+	-
			2	+	-	+	-
			3	-	+	-	+
			4	-	+	-	+

Wyróżniono wybrany stopień wielomianu.

W kolumnach ‘monot.’:

‘+’ oznacza zależność monotoniczną we wszystkich analizowanych grupach;

‘+/-’ oznacza niewielkie odstępstwa od monotoniczności;

‘-’ oznacza pozostałe sytuacje.

W kolumnach ‘linktest’:

‘+’ oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,05 (we wszystkich analizowanych grupach);

‘+/-‘ oznacza współczynniki przy kwadracie przewidywania nieistotne na poziomie 0,01 (we wszystkich analizowanych grupach);

‘-‘ oznacza pozostałe sytuacje.

Aneks F: Dekompozycja wariancji efektów losowych w trzyletnich modelach EWD

Trzyletnie modele EWD gimnazjów

	część hum. (j. pol i hum.)			część mat-przyr.		
okres	τ	σ^2	$\tau / (\tau + \sigma^2)$	τ	σ^2	$\tau / (\tau + \sigma^2)$
2013-2011 (z wydł.)	8,60	86,11	9,1%	7,99	79,13	9,2%
2012-2010 (z wydł.)	7,73	89,05	8,0%	7,07	77,56	8,4%
2011-2009 (z wydł.)	7,69	86,84	8,1%	6,83	77,91	8,1%
2010-2008	7,49	88,93	7,8%	7,19	81,05	8,1%
2009-2007	7,85	86,19	8,3%	7,23	84,39	7,9%
2008-2006	8,69	91,28	8,7%	8,04	90,09	8,2%

Trzyletnie modele EWD szkół kończących się maturą

	j.polski			przedmioty hum.		
okres	τ	σ^2	$\tau / (\tau + \sigma^2)$	τ	σ^2	$\tau / (\tau + \sigma^2)$
2013-2011 LO	20,83	140,37	12,9%	21,99	136,74	13,9%
2012-2010 T	16,38	154,71	9,6%	15,21	152,36	9,1%
2012-2010 LO	20,28	140,53	12,6%	21,45	135,81	13,6%
2012-2010 T	17,07	156,01	9,9%	15,88	153,56	9,4%
	matematyka			przedmioty mat.-przyr.		
rok	τ	σ^2	$\tau / (\tau + \sigma^2)$	τ	σ^2	$\tau / (\tau + \sigma^2)$
2013-2011 LO	17,20	84,25	17,0%	18,50	75,39	19,7%
2012-2010 T	20,14	104,28	16,2%	19,20	102,00	15,8%
2012-2010 LO	16,54	85,67	16,2%	17,90	76,69	18,9%
2012-2010 T	19,68	104,62	15,8%	18,81	102,51	15,5%

Literatura cytowana

Ballou, D., Sanders, W. i Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics 29*, 37-65.

Barnett, V. (1975). Probability Plotting Methods and Order Statistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 24*(1), 95-108.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. New York: Oxford University Press.

Biecek, P. (2011). *Analiza danych z programem R*. Warszawa: Wydawnictwo Naukowe PWN.

Cameron, A. C. i P. K. Trivedi (2010), *Microeconometrics using Stat,* Michigan: StataPress.

Dolata, R. (Red.). (2007a). *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie egzaminów zewnętrznych*. Warszawa: Centralna Komisja Egzaminacyjna.

Dolata, R. (2007b). Krytyczna analiza metody edukacyjnej wartości dodanej. [W:] Dolata, R. (Red.), *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie egzaminów zewnętrznych*. Warszawa: Centralna Komisja Egzaminacyjna.

Dolata, R., Hawrot, A., Humenny, G., Jasińska, A., Koniewski, M., Majkut, P. i Żółtak,T. (2013). *Trafność metody edukacyjnej wartości dodanej dla gimnazjów*. Warszawa: IBE.

Dolata, R. i Pokropek, A. (2012). Czy warto urodzić się w styczniu? Wiek biologiczny a wyniki egzaminacyjne. [W:] B. Niemierko i M. K. Szmigel (Red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy*. Kraków: Grupa Tomami.

Evans, H. (2008). *Value-added in English schools*. Referat przedstawiony na National Conference on Value-Added Modeling, University of Wisconsin at Madison, 22-24 kwietnia. Pobrany 28 lutego 2013 z: http://www.wcer.wisc.edu/news/events/VAM%20Conference%20Final%20Papers/ VAMinEnglishSchools_HEvens.pdf.

Faraway, J. (2004). *Linear Models with R*. London: Chapman & Hall/CRC.

Grudniewska, M., Kondratek, B. (2012). Zróżnicowane funkcjonowanie zadań w egzaminach zewnętrznych w zależności od płci na przykładzie części matematyczno-przyrodniczej egzaminu gimnazjalnego. [W:] B. Niemierko i M. K. Szmigel (Red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy*. Kraków: Grupa Tomami.

Hawrot, A., Jasińska, A. (2013). Wiek i inteligencja a wyniki egzaminacyjne i EWD. [W:] Dolata, R., Hawrot, A., Humenny, G., Jasińska, A., Koniewski, M., Majkut, P. i Żółtak,T. , *Trafność metody edukacyjnej wartości dodanej dla gimnazjów*. Warszawa: IBE.

Jakubowski, M. (2007). Empiryczna analiza metod szacowania edukacyjnej wartości dodanej dla gimnazjum. [W:] Dolata,R. (Red.), *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie egzaminów zewnętrznych*. Warszawa: Centralna Komisja Egzaminacyjna.

Jasińska, A., Żółtak, T. (2013). Trafność egzaminów zewnętrznych z punktu widzenia wykorzystania ich do szacowania gimnazjalnych wskaźników EWD. [W:] Dolata, R., Hawrot, A., Humenny, G., Jasińska, A., Koniewski, M., Majkut, P. i Żółtak,T. , *Trafność metody edukacyjnej wartości dodanej dla gimnazjów*. Warszawa: IBE.

Karwowski, M. (Red.). (2013). *Scieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne.* Warszawa: Wydawnictwo Instytutu Filozofii i Socjologii PAN.

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

Korobko, O. B., Glas, C. A., Bosker, R. J., Luyten, J. W. (2008). Comparing the Difficulty of Examination Subjects with Item Response Theory. *Journal of Educational Measurement, 45*(2), 139-157.

Linden van der, W. J. i Hambelton, R. K. (Red.) (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Louis, T. A. i Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics 29*, 67-101.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M. i Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, California: RAND.

OECD. (2008). Measuring improvements in learning outcomes: best practices to assess the value-added of schools. Paryż: OECD.

Pokropek, A. (2011). Matura z języka polskiego. Wybrane problemy psychometryczne. [W:] B. Niemierko i M. K. Szmigel (Red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy*. Kraków: Grupa Tomami.

Pokropek, A. (2013). Trafność testów egzaminacyjnych. [W:] M. Karwowski (Red.), *Ścieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne*. Warszawa: Wydawnictwo Instytutu Filozofii i Socjologii PAN.

Pokropek, A. i Żółtak, T. (2012). Nowe modele jednorocznej EWD. [W:] B. Niemierko i M. K. Szmigel (Red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy*. Kraków: Grupa Tomami.

Raudenbush, S. W. i Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*, 307-335.

Robinson, G. (1999). That BLUP is a good thing: the estimation of random effects. *Statistical Science, 6*(1), 15-32.

Skórska, P., Koniewski, M., Majkut, P. (2013). Wpływ wersji arkusza egzaminacyjnego na zróżnicowane funkcjonowanie zadań na przykładzie egzaminu gimnazjalnego. [W:] B. Niemierko i M. K. Szmigel (Red.), *Polska edukacja w świetle diagnoz prowadzonych z różnych perspektyw badawczych*. Kraków: Grupa Tomami.

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F., Pokropek, A. (2012). Wyniki egzaminu gimnazjalnego 2002–2010 na wspólnej skali. *Edukacja, 3* (119), 9-30.

Takane, Y., de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discredited variables. *Psychometrika, 52*, 393-408.

Twardowska, A., Grajkowski, W., Chrzanowski, M., Ostrowska, B., Spalik, K. (2011) Dlaczego warto zamykać zadania? [W:] B. Niemierko i M. K. Szmigel (Red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy*. Kraków: Grupa Tomami.

Żółtak, T. (2011). Znaczenie informacji o średnim wyniku uczniów na wejściu dla własności jednorocznych wskaźników EWD gimnazjów. [W:] B. Niemierko i M. K. Szmigel (Red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy*. Kraków: Grupa Tomami.

Żółtak, T.(2013a). EWD jako sposób badania efektywności szkół. [W:] M. Karwowski (Red.), *Ścieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne*. Warszawa: Wydawnictwo Instytutu Filozofii i Socjologii PAN.

Żółtak, T.(2013b). Gimnazjalne wskaźniki EWD. [W:] Dolata, R., Hawrot, A., Humenny,G., Jasińska, A., Koniewski, M., Majkut, P. i Żółtak,T. , *Trafność metody edukacyjnej wartości dodanej dla gimnazjów*. Warszawa: IBE.