KWARTALNIK PEDAGOGICZNY
2015 NUMER 4(238)

Aleksandra Jasińska-Maciążek
Grzegorz Humenny
Instytut Badań Edukacyjnych w Warszawie

Wykorzystanie egzaminów o niskiej doniosłości do oceny efektywności nauczania – możliwości i ograniczenia

Summary

LOW-STAKES TESTS AS A BASIS FOR THE ASSESSMENT OF TEACHING EFFECTIVENESS

The article discusses the possibility of using low-stakes tests to assess teaching effectiveness measures (educational value-added, EVA) for primary schools in Poland. It was shown that EVA, calculated using the results of the OBUT test for diagnosis of skills of third grade pupils, which is voluntary for schools, and the *Sprawdzian* – external examination conducted in the sixth grade, can be a valuable tool for evaluation, provided that the third-grade test is conducted and assessed in accordance with the procedures. The authors identified problems in this field. Disturbing characteristics of the OBUT test may suggest that these procedures have not been complied with in every school. For this reason, the authors recommend changes in the implementation of such diagnoses of school achievements.

Key words: low-stakes tests, assessment of teaching effectiveness, educational value-added, primary school.

red. Paulina Marchlik

Testy osiągnięć szkolnych możemy podzielić, ze względu na konsekwencje, jakie wiążą się z wykorzystaniem ich wyników, na testy o doniosłym znaczeniu (wysokiej stawki – *high stakes*) i testy o niskiej doniosłości (powszednie, niskiej stawki – *low stakes*) (Niemierko 2009). W podziale tym bierze się pod uwagę konsekwencje dla uczniów. Egzaminy doniosłe decydują o ich losach na kolejnych etapach kształcenia. Testy o niskiej doniosłości nie wiążą się z poważnymi konsekwencjami (są to np.: szkolne klasówki, testy diagnostyczne). Jednak test

o niskiej doniosłości dla ucznia, może być testem doniosłym z punktu widzenia szkoły lub nauczycieli, jeśli jest wykorzystywany do oceny ich pracy.

Obecnie coraz większe znaczenie przypisuje się testom osiągnięć szkolnych jako narzędziu ewaluacji placówek edukacyjnych oraz monitorowania systemu oświaty, służącemu poprawie efektywności nauczania (EACEA 2010; OECD 2013). W niektórych krajach wykorzystuje się do tego celu egzaminy doniosłe (np. Szkocja, Irlandia, Łotwa), w innych testy o niskiej doniosłości (np. Australia, Węgry, Anglia). Poniżej przybliżamy założenia systemów opartych na testach niskiej stawki zaznaczając, że są to przykłady, a nie wyczerpująca lista takich rozwiązań.

Programy ewaluacyjne w Australii, Anglii i na Węgrzech

Australijski system ogólnokrajowych egzaminów *National Assessment Program – Literacy and Numeracy* (NAPLAN) istnieje od 2008 roku. Do testów przystępują wszyscy uczniowie kończący naukę w klasach 3, 5, 7 i 9 (Freeman 2009). Główną ich funkcją jest monitorowanie postępów uczniów, w celu zapewnienia im możliwości jak najlepszego wykorzystania potencjału. Na Węgrzech od 2001 roku istnieje ogólnokrajowa ocena podstawowych kompetencji (*National Assessment of Basic Competencies* – Sinka 2010). Od 2008 roku pomiar obejmuje wszystkich uczniów kończących 6, 8 i 10 klasę (wcześniej badania były realizowane na próbach), a uczniowie dodatkowo wypełniają ankietę dotyczącą ich pochodzenia społecznego, co pozwala na kontekstową ocenę wyników szkół. Ponadto w czwartej klasie (10 lat) realizowane jest badanie diagnostyczne (*Diagnostic skills and ability assessment*), które ma na celu ocenę poziomu umiejętności uczniów i dostarczenie nauczycielom informacji potrzebnych do oceny i planowania własnej pracy z uczniami (Hungarian Eurydice Unit, 2009). Program węgierski został wprowadzony w celu rozwoju systemu ewaluacji jednostek edukacyjnych, dostarczenia szkołom danych i narzędzi umożliwiających profesjonalną ocenę własnej pracy oraz porównanie swoich wyników z innymi szkołami w kraju (Balázsi 2006). Podobnie w angielskim systemie egzaminów krajowych (*National Curriculum Assessment*) od 2008 roku testuje się wszystkich uczniów w 2, 6 i 11 (lub 10) roku nauki. Dodatkowo istnieje dobrowolny pomiar w dziewiątym roku nauki. System ten stworzono w przekonaniu, że monitorowanie rozwoju poznawczego uczniów jest dobrą praktyką, a ocena poziomu osiągnięć może wspierać proces nauczania i uczenia się (National Foundation for

Educational Research 2009). Żaden z wymienionych pomiarów umiejętności nie ma charakteru doniosłego dla uczniów.

Testy te, w zależności od systemu i etapu edukacji, przeprowadzane i oceniane są w różny sposób. W Anglii pomiar po drugim roku nauki oceniany i przeprowadzany jest przez nauczycieli badanych klas. Podobnie jak egzamin w czwartej klasie na Węgrzech (za wyjątkiem arkuszy wybranych do ogólnokrajowej próby, ocenianych zewnętrznie). Pozostałe testy w tych krajach, oceniane są przez zewnętrznych egzaminatorów. W systemach australijskim i angielskim wyniki kolejnych pomiarów są przedstawiane na wspólnej skali osiągnięć szkolnych odpowiadającej założonym poziomom realizacji krajowego programu nauczania. Pozwala to na ocenę postępów ucznia oraz sprawdzenie, czy jego wyniki są zgodne z oczekiwanymi (EACEA 2010; ACARA 2014). Na Węgrzech testy sprawdzają nie tyle stopień opanowania programu nauczania, co umiejętności wykorzystania posiadanej wiedzy do rozwiązywania problemów praktycznych (Hungarian Eurydice Unit 2009).

W każdym z tych systemów najważniejszym odbiorcą wyników są szkoły i nauczyciele, a podstawowym celem pomiarów jest poprawa efektywności nauczania. Wyniki szkół z większości pomiarów obowiązkowych są powszechnie dostępne[1]. Są one prezentowane nie tylko na tle średniej w kraju i w mniejszych jednostkach terytorialnych, ale także jako przetworzone wskaźniki uwzględniające kontekst pracy szkoły lub uprzednie osiągnięcia szkolne uczniów. Wyniki te wykorzystywane są nie tylko przez szkoły. Stanowią też źródło informacji dla rodziców zainteresowanych wyborem szkoły dla dziecka. Służą także władzom i innym organom nadzoru do ewaluacji zewnętrznej (Balázsi 2006; EACEA 2010). To sprawia, że stają się one dla szkół testami wysokiej stawki.

Egzaminy w Polsce

W Polsce powszechny pomiar osiągnięć odbywa się od 2002 roku w ramach systemu egzaminów zewnętrznych. Egzamin gimnazjalny oraz matura pełnią m.in. funkcję selekcyjną, ewaluacyjną i monitorującą (Dolata, Szaleniec 2012). Są wykorzystywane przez szkoły do oceny i planowania własnej pracy (Matuszczak, Wasilewska 2015; Wasilewska i in., 2014) oraz przez organy prowadzące

[1] Portal australijski: www.myschool.edu.au i angielski: http://www.education.gov.uk/schools/performance.

i inne podmioty w ewaluacji zewnętrznej (Herczyński i in. 2012; Bąbiak i in. 2013; SEO 2013).

Obowiązkowy egzamin na zakończenie szkoły podstawowej – sprawdzian – miał pełnić funkcję diagnostyczną, jednak jest testem krótkim, mało zróżnicowanym pod względem treści zadań, nie posiadającym norm dla poszczególnych wymagań programowych (Dolata, Szaleniec 2012), co utrudnia wykorzystanie jego wyników do oceny realizacji wymagań programowych. Dla uczniów, co do zasady, jest to test niskiej stawki. W praktyce niektóre gimnazja wykorzystują jednak jego wyniki jako jedno z kryteriów selekcji do szkoły uczniów spoza rejonu[2]. Sprawdzian wypełnia także funkcję ewaluacyjną – pomaga szkołom w ocenie efektów nauczania i znajdowaniu obszarów do poprawy (Matuszczak, Wasilewska 2015).

Egzaminy zewnętrzne nie obejmują wcześniejszych etapów edukacji. Do 2010 roku szkoły mogły korzystać jedynie z narzędzi przygotowanych samodzielnie albo z oferty wydawnictw proponujących własne testy. Wadą tych rozwiązań był brak możliwości zobaczenia rezultatów na tle wyników populacyjnych. Od 2011 do 2014 roku organizowane było Ogólnopolskie Badanie Umiejętności Trzecioklasistów (OBUT)[3] – dobrowolna diagnoza umiejętności z języka polskiego i matematyki. W 2011 roku wzięło w niej udział ponad 75% szkół podstawowych, co dało możliwość analizowania wyników szkół na tle wyników krajowych (Pregler, Wiatrak 2011).

Wykorzystanie wyników testów do oceny efektywności nauczania

Jednym z negatywnych skutków egzaminów zewnętrznych jest traktowanie ich wyników jako głównego lub jedynego wskaźnika oceny pracy szkoły i nauczycieli (Szaleniec 2004). Średnie wyniki wykorzystywane są np. do tworzenia prasowych rankingów szkół. Zestawienia te nie biorą pod uwagę faktu, że szkoły nie mają takich samych szans na uzyskanie porównywalnych rezultatów. Wyniki zależą od szeregu czynników, na które szkoła nie ma wpływu, takich jak inteligencja, status społeczno-ekonomiczny rodziny, wykształcenie i aspiracje rodziców (Konarzewski 2012; Dolata i in. 2014; Jasińska-Maciążek, Modzelewski 2014).

[2] Np. Gimnazjum nr 46 w Bydgoszczy: http://www.2lo.bydgoszcz.pl

[3] W latach 2006–2011 badanie było realizowane na reprezentatywnych próbach ogólnopolskich pod kierownictwem M. Dąbrowskiego i M. Żytko.

Szkoły, wykorzystując wyniki testów do oceny własnej pracy, przenoszą uwagę z ucznia na proces nauczania. Może przełożyć się to na skuteczniejsze nauczanie. Aby szkoła mogła podejmować trafne decyzje, potrzebuje jednak dobrych wskaźników procesów w niej zachodzących (MacBeath i in. 2005). Wskaźniki te powinny być niezależne od tego, na co szkoła nie ma wpływu. Wyniki testowe szkół powinny być odnoszone do wyników innych placówek pracujących z podobnymi uczniami. Z takiego myślenia wywodzi się idea edukacyjnej wartości dodanej.

Edukacyjna wartość dodana (EWD) jako miara efektywności nauczania

Wskaźniki EWD są miarą względnego przyrostu osiągnięć szkolnych uczniów na danym etapie kształcenia, a do ich obliczenia potrzebne są co najmniej dwa pomiary osiągnięć: na początku i na zakończenie tego etapu. Pokazują one, czy uczniowie w danej szkole osiągnęli średnio wyniki wyższe, niższe czy porównywalne do wyników uczniów podobnych pod względem uprzednich osiągnięć. Z tego względu uważane są za lepszą miarę jakości nauczania niż np. średnie wyniki z testu (OECD 2008).

Edukacyjna wartość dodana traktowana jest jako miara wkładu szkoły lub nauczycieli w wyniki egzaminacyjne – wskaźnik efektywności nauczania (McCaffrey i in. 2003; OECD 2008). Najprostsze modele uwzględniają jedynie informację o poziomie uprzednich osiągnięć uczniów, bowiem nie tylko bezpośrednio determinuje on wyniki na kolejnych etapach nauczania, ale jest także nośnikiem informacji o czynnikach niezależnych od szkoły (Dolata i in. 2013). W modelach EWD można również uwzględniać charakterystyki uczniów (np. inteligencja, status społeczny rodziny) i otoczenia szkoły (np. wielkość miejscowości, zamożność mieszkańców, poziom bezrobocia), by w ten sposób oczyścić wskaźniki z ich wpływu. To, jakie zmienne zostaną uwzględnione powinno zależeć od celów, jakim mają służyć obliczone miary (Raudenbush, Willms 1995). Niezależnie od liczby zmiennych kontekstowych nie należy jednak traktować wskaźników EWD jako bezwzględnego dowodu na efektywność działań podejmowanych przez szkoły i nauczycieli (McCaffrey i in. 2003; Raudenbush 2004). Dobre wykorzystanie miary EWD w ewaluacji wymaga sięgania po dodatkowe informacje dostępne nauczycielom pracującym w szkole. Analiza wskaźników

EWD razem z kontekstem pracy szkoły może pomóc w ocenie skuteczności podejmowanych działań i bardziej świadomym planowaniu pracy dydaktycznej (Stożek 2012).

W Polsce metoda EWD dopiero się rozwija. Gimnazja mogą analizować wskaźniki EWD od 2006 roku, a licea i technika od 2010 roku. Dla szkół podstawowych do 2014 roku nie było możliwości obliczania wskaźników EWD, ponieważ jedynym powszechnym pomiarem osiągnięć szkolnych był sprawdzian w klasie szóstej. Badanie OBUT umożliwiło myślenie o modelach EWD dla drugiego etapu edukacyjnego (klasy IV–VI). Uczniowie biorący w nim udział w 2011 roku trzy lata później przystąpili do sprawdzianu. Był to pierwszy rocznik uczniów, dla którego można było obliczyć wskaźniki EWD dla szkół podstawowych.

Niebezpieczeństwa wynikające z wykorzystania egzaminów o niskiej doniosłości do oceny nauczania

Wyniki testów o niskiej doniosłości, jako że nie wiążą się z poważnymi konsekwencjami dla ucznia, mogą być szczególnie podatne na zniekształcenia związane z wpływem motywacji testowej. Determinuje ona podjęcie przez ucznia próby rozwiązania zadań, wysiłek poznawczy i wytrwałość w ich rozwiązaniu (Dolata, Pokropek 2010; Skórska i in. 2014). Wyniki takich testów prawdopodobnie zaniżają poziom umiejętności uczniów (Wolf, Smith 1995; Wise, DeMars 2005), przekładając się na niedoszacowanie efektów kształcenia.

Fakt, że test nie ma znaczących konsekwencji dla uczniów nie oznacza, że jest on testem niskiej stawki dla nauczycieli czy szkół (Corcoran i in. 2011). Jeśli wyniki wykorzystywane są do oceniania pracy nauczycieli albo podejmowania istotnych dla nich decyzji, mogą oni próbować wpłynąć na poziom motywacji testowej uczniów. W efekcie uzyskane wyniki mogą być inne niż w sytuacji, gdy pomiar jest testem o niskiej doniosłości dla wszystkich zainteresowanych (Papay 2011; Jager i in. 2012). W kontekście porównywania wyników szkół bardziej problematyczny jest zróżnicowany stosunek placówek do testów. W przypadku badania OBUT jest prawdopodobne, że taka sytuacja miała miejsce. Autorom artykułu znane są przypadki, w których wyniki testu OBUT niosły za sobą ważne konsekwencje dla nauczycieli, np. były komunikowane rodzicom,

którzy mieli wybrać nauczyciela dla swojego dziecka. Brakuje jednak danych o skali tego zjawiska.

Nie tylko konsekwencje formułowane wprost mogą sprawić, że nauczyciele będą postrzegać dany test jako doniosły. Ważna może być obawa przed negatywną oceną koleżeńską lub dyrektorską. Im wyższa postrzegana lub rzeczywista stawka, o której decyduje wynik, tym większe prawdopodobieństwo pojawienia się niezamierzonych, negatywnych konsekwencji. W literaturze można znaleźć różne przykłady oszustw i nadużyć związanych z sytuacją testową. W Holandii, niektóre szkoły podstawowe, nie obejmowały pomiarem słabszych uczniów, próbując w ten sposób podnieść średni wynik (EACEA 2010). W USA wskazuje się na takie praktyki jak zawężanie programów nauczania do treści egzaminu, przekazywanie nauczycielom testów z wyprzedzeniem, by przygotowali uczniów do egzaminu, podpowiadane uczniom podczas pisania testu, podawanie im wcześniej prawidłowych odpowiedzi, „naciąganie" wyników podczas sprawdzania prac, czy zmiana uczniowskich odpowiedzi na kartach egzaminacyjnych (Jacob, Levitt 2003; Nichols, Berliner 2007).

Procedury przeprowadzania i oceniania testów o niskiej doniosłości często są mniej rygorystyczne, niż egzaminów doniosłych. Badanie OBUT było na przykład prowadzone przez pracowników szkoły (zwykle wychowawców klas), bez zewnętrznych obserwatorów – jak w przypadku egzaminów zewnętrznych. Dodatkowo, testy z badania OBUT były sprawdzane przez nauczycieli badanych klas lub innych pracowników szkoły według reguł opracowanych przez zespół badawczy (CKE 2011). Jest to zrozumiałe z punktu widzenia kosztów testowania, jednak sytuacja taka może nasilać problemy związane z oszustwami egzaminacyjnymi, a także skutkować zniekształceniem wyników wynikającym z różnej (niewystandaryzowanej) interpretacji procedur testowania czy reguł oceniania prac.

Warunki obliczenia wartościowych wskaźników efektywności nauczania dla drugiego etapu edukacyjnego

W literaturze poświęconej metodzie EWD zwraca się uwagę na kilka kluczowych dla obliczenia wskaźników aspektów (OECD 2008; Chudowsky i in. 2010). Przede wszystkim potrzebne są dwa dobre pomiary osiągnięć szkolnych: na początku i na zakończenie etapu nauczania, który chcemy oceniać. Dobry

w tym kontekście test, to taki, który trafnie mierzy poziom realizacji celów kształcenia. Pomiary muszą być rzetelne oraz nieobciążone działaniem czynników niezwiązanych z poziomem osiągnięć (np. motywacją testową, czy niedotrzymaniem procedur realizacji lub oceniania testów). Dodatkowo skala, na której prezentowane są wyniki, powinna być interwałowa (Ballou 2009; Reardon, Raudenbush 2009) i mieć wystarczająco dużo wartości, by mogła być traktowana w analizach jako zmienna ciągła.

Do obliczenia wskaźników EWD można wykorzystać dane populacyjne, jak i dane z próby szkół. Musi to być jednak próba reprezentatywna. Niezbędne jest także, aby dla każdego ucznia wyniki obu pomiarów były poprawnie połączone.

Kolejna kwestia wiąże się z właściwym statystycznym wymodelowaniem względnych postępów uczniów. Modele EWD opierają się na zdroworozsądkowym założeniu, że funkcja opisująca zależność wyników egzaminu na zakończenie danego etapu edukacji od wyników egzaminu na jego początku powinna być niemalejąca (Żółtak 2013). Musi być ona dobrze dopasowana do danych, co w przypadku związku nieliniowego wymaga zastosowania odpowiednich metod.

Istotne jest również to, na ile skutecznie w modelu EWD udaje się kontrolować znaczenie czynników wpływających na postępy uczniów, które są niezależne od szkoły. Dostępne dane pozwalają jedynie na sprawdzenie, czy obliczone wskaźniki są niezależne od średniego w szkole poziomu osiągnięć „na wejściu" (czyli po III klasie). Brak takiej zależności na poziomie uczniów wynika bezpośrednio z modelu. Związek na poziomie szkół wskazywałby na problemy w tym zakresie.

Zdajemy sobie sprawę z tego, że przytoczona powyżej lista warunków nie jest wyczerpująca i dodatkowe badania są potrzebne, by z pełnym zaufaniem wykorzystywać wskaźniki EWD. Nie uda nam się też wystarczająco wnikliwie przeanalizować każdego z przytoczonych problemów. Celem artykułu nie jest jednak sformułowanie jednoznacznej i wielowymiarowej oceny jakości wskaźników dla drugiego etapu kształcenia. Został on zarysowany skromniej, jako zbadanie możliwości wykorzystania wyników pomiaru takiego, jak badanie OBUT i wyników sprawdzianu do szacowania wskaźników EWD oraz naświetlenie możliwych problemów, jakie możemy napotkać wykorzystując do tego testy o niskiej doniosłości. Jest to więc dopiero początek drogi do zbudowania wartościowych miar efektywności nauczania dla szkół podstawowych.

Metoda

Wykorzystane dane

W analizach wykorzystano wyniki sprawdzianu w szóstej klasie – obowiązkowego dla uczniów egzaminu zewnętrznego realizowanego 1.04.2014 (dane z arkuszy standardowych) oraz dane z badania OBUT realizowanego 17.05.2011.

Skalowanie wyników przeprowadzono na danych populacyjnych. Z uwagi na fakt, że wyniki piszących test OBUT zostały zanonimizowane, niemożliwe było połączenie ich z wynikami sprawdzianu na poziomie centralnym. Dlatego modele EWD obliczono na danych z losowej ogólnopolskiej próby szkół, dla których pozyskano połączone dla uczniów wyniki testów. Szkoły przekazały sumy punktów z testów OBUT i sprawdzianu, informację o płci ucznia oraz posiadaniu przez niego zaświadczenia o potrzebie dostosowania warunków na sprawdzianie (tzw. dysleksji rozwojowej).

Dobór próby szkół odbył się przez systematyczne losowanie placówek, które przystąpiły do testu OBUT w 2011 roku, w warstwach wydzielonych ze względu na województwa oraz średni wynik testu OBUT w szkołach. W warstwach szkoły uporządkowane zostały ze względu na liczbę uczniów przystępujących do tego pomiaru. Liczba szkół losowana w każdej warstwie była proporcjonalna do liczby uczniów. Próba założona liczyła 200 szkół. W ustalonym terminie udało się pozyskać dane od 181 szkół.

Charakterystyka narzędzi i skalowanie wyników testów

Test OBUT składał się z dwóch części (języka polskiego i matematyki). Łącznie test zawierał 20 zadań, za które można było uzyskać 41 punktów. Test był relatywnie łatwy (lewoskośny rozkład wyników, współczynnik skośności = –0,70). Sprawdzian zawierał 26 zadań, za które można było uzyskać 40 punktów. Był to również test łatwy (współczynnik skośności = –0,47).

Wyniki obu pomiarów zostały wyskalowane z zastosowaniem modelu Rascha dla zadań ocenianych binarnie oraz modelu *partial credit* dla zadań punktowanych na dłuższych skalach (modele funkcjonujące w ramach teorii IRT – *item response theory*) (De Ayala 2009; Szaleniec 2009). Zastosowano metodę estymacji *marginal maximum likelihood*. Za wskaźnik poziomu osiągnięć przyjęto oszacowania *expected a posteriori* (EAP). Wyniki badania OBUT przedstawiono na jednej

jednowymiarowej skali opisującej osiągnięcia szkolne po III klasie (w modelu uwzględniono łącznie zadania z testu z języka polskiego i matematyki). Wyniki sprawdzianu również przedstawiono na jednowymiarowej skali.

Model Rascha wybrano z dwóch powodów. Po pierwsze, uważa się, że tylko modele jednoparametryczne (a takim jest model Rascha, w którym jedynym szacowanym parametrem jest trudność zadania) mają teoretyczne podstawy co do tego, by skale utworzone z ich zastosowaniem można było traktować jako interwałowe (Ballou 2009). Czyli, by można było przyjąć, że wzrost o jedną jednostkę na skali oznacza równoważny przyrost osiągnięć szkolnych w każdym jej zakresie. Przy czym można tak stwierdzić tylko wtedy, gdy model jest dobrze dopasowany. Dopasowanie to było przedmiotem analiz i zostało omówione w części wynikowej. Założenie to nie jest natomiast spełnione dla skali nieprzetworzonej sumy punktów.

Ponadto spośród różnych modeli IRT tylko w modelu Rascha suma punktów zdobyta w teście jest statystyką dostateczną dla oszacowania poziomu osiągnięć (wyniku wyskalowanego) (De Ayala 2009). Miało to bardzo duże znaczenie praktyczne. Wyniki testu OBUT są przechowywane jedynie w szkołach, zwykle w postaci sumy punktów uzyskanych w teście. Aby z wyników analiz mogło korzystać jak najwięcej szkół, należało wybrać taki model, który wymaga znajomości jedynie sumy punktów. Z tego względu nie można też było podczas skalowania usunąć słabszych pod względem psychometrycznym zadań, ani skracać skal dla zadań punktowanych na dłuższych skalach.

Zastosowanie modeli IRT wymaga spełnienia założenia o lokalnej niezależność pozycji. W celu wykrycia grup zadań, które mogą łamać to założenie wykonano analizę skupień zmiennych wokół cechy latentnej (CLV – Vigneau, Qannari 2003). W efekcie po kilka pozycji zarówno w teście OBUT, jak i ze sprawdzianu zostało połączonych, co pozwoliło spełnić założenia modelu.

Wyskalowanie wyniki testów przedstawiono na skali o średniej 100 i odchyleniu standardowym 15 w populacji uczniów.

Schemat analizy

Ocena jakości skal osiągnięć szkolnych

Jakość dopasowania danych do modelu Rascha została zweryfikowana za pomocą miar *outfit* i *infit*. Miary te liczone są dla zadań. *Outfit* jest równie wrażliwa na odstępstwa od modelu dla całego zakresu skali, natomiast *infit* dla tych

przypadków, które znajdują się w okolicy parametru trudności zadania. Jeśli miary te przyjmują wartość 1, uznaje się, że zadanie jest dobrze dopasowane do modelu. Dopuszcza się jednak pewne rozchwianie ich wartości (De Ayala 2009: 55–57), przyjmując przedział wartości od 0,8 do 1,2 jako świadczący o dobrym dopasowaniu.

Jako miary rzetelności wykorzystano alfę Cronbacha oraz współczynnik rzetelności EAP/PV (Adams 2005; Jasińska, Modzelewski 2014). EAP/PV jest jedną z miar rzetelności wykorzystywaną w modelach IRT. Jest to stosunek wariancji oszacowania punktowego (EAP) do całkowitej wariancji zmiennej ukrytej (obliczanej jako wariancja PV – *plausible values*). Obie miary mogą przyjmować wartości od 0 do 1.

W celu przeprowadzenia analizy ewentualnego obciążenia wyników testowych błędem nielosowym związanym z oddziaływaniem różnych czynników na poziomie szkoły (np. różnej motywacji testowej, niedotrzymywaniem reguł realizacji badania lub kodowania prac) zestawiono międzyszkolną wariancję wyników testu OBUT z danymi pokazującymi zróżnicowanie międzyszkolne po III klasie w innych badaniach. Dekompozycji wariancji dokonano wykorzystując dwupoziomowe modele mieszanych efektów, a wariancję międzyszkolną obliczono jako stosunek wariancji efektów losowych związanych z podziałem na szkoły do wariancji całkowitej (Domański, Pokropek 2011). Współczynnik ten mówi nam o tym, jaką część zmienności wyników możemy przypisać podziałowi na szkoły. Większe niż oczekiwane międzyszkolne zróżnicowanie wyników testu OBUT mogłoby świadczyć o istnieniu w szkołach procesów zniekształcających wyniki testowe.

Analiza reprezentatywności próby

Analizę reprezentatywności próby przeprowadzono porównując średnią arytmetyczną oraz odchylenie standardowe wyników uczniów z próby z wynikami populacyjnymi. Dla wyników OBUT wykorzystano dane ze wszystkich szkół, które przystąpiły do badania, a dla sprawdzianu wyniki populacyjne.

Modelowanie EWD

Poziom uprzednich osiągnięć szkolnych jest najważniejszą zmienną niezależną w modelach EWD. Dlatego kluczowe jest dobre dopasowanie funkcji

łączącej wynik „na wyjściu" z wynikiem „na wejściu". W praktyce funkcja ta przeważnie jest krzywoliniowa i z tego powodu modeluje się ją wielomianem odpowiedniego stopnia (Żółtak 2013). W prezentowanych analizach przyjęto analogiczne podejście. Poza wynikiem z testu OBUT w modelu uwzględniono informację o płci i dysleksji. Ponadto przyjęto, że model EWD powinien spełniać dwa kryteria. (1) Funkcja łącząca wyniki sprawdzianu z wynikami testu OBUT powinna być niemalejąca. (2) Wszystkie współczynniki regresji (w tym kolejne potęgi wyników testu OBUT) powinny być istotne statystycznie; przy czym im niższy stopień wielomianu, tym lepiej. Modelowanie przeprowadzone zostało na połączonych danych z próby szkół.

Procedura wyboru najlepiej dopasowanego modelu polegała na przeprowadzeniu szeregu regresji estymowanych metodą najmniejszych kwadratów (MNK), do których wprowadzane były: informacja o płci, dysleksji, wyniku testu OBUT oraz krokowo kolejne jego potęgi. Wyniki każdego modelu były weryfikowane pod względem spełniania przyjętych kryteriów. Regresję MNK wybrano, gdyż daje ona możliwość obliczania wskaźników EWD dla dowolnie zdefiniowanych grup uczniów (np. uczniów jakiejś szkoły) jako średniej z reszt. Dzięki temu użytkownicy metody mogą prowadzić analizy w podziale na różne podgrupy (np. na klasy lub płeć).

Ocena właściwości wskaźników EWD

Weryfikację, czy wskaźniki EWD są niezależne od tego, z jakimi uczniami (o jakich cechach) szkoła pracuje, przeprowadzono licząc współczynnik korelacji między średnim wynikiem testu OBUT w szkole a wartością wskaźnika EWD dla szkoły.

Wyniki

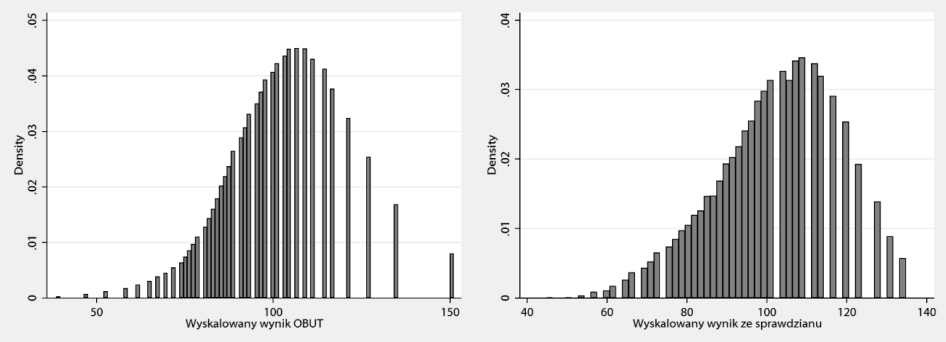
Właściwości wykorzystanych testów

Analiza miar *infit* i *outfit* dla zadań z testu OBUT i sprawdzianu pokazała zadowalające dopasowanie danych do modelu Rascha (statystyki opisowe zamieszczono w Tabeli 1). W przypadku testu OBUT tylko dla dwóch zadań wartości

miar dopasowania wykroczyły poza przyjęty zakres (0,8–1,2), w tym tylko *outfit* dla jednego zadania przyjął wartość odbiegającą od tego zakresu w stopniu znaczącym (1,77). W przypadku sprawdzianu miary dla dwóch zadań w sposób nieznaczny odbiegały od przyjętego zakresu świadczącego o dobrym dopasowaniu. Wyniki te pokazały, że ze względu na kryterium dopasowania skale zbudowane z wykorzystaniem modelu Rascha można traktować jak interwałowe.

Tabela 1. Statystyki opisowe miar dopasowania zadań do modelu

Statystyki	OBUT		Sprawdzian	
	outfit	*infit*	*outfit*	*infit*
Minimum	0,80	0,87	0,77	0,88
1 kwartyl	0,88	0,93	0,93	0,96
3 kwartyl	1,09	1,04	1,04	1,04
Maksimum	1,77	1,39	1,05	1,04



Rysunek 1. Rozkłady wyskalowanych wyników testu OBUT i sprawdzianu

Bardziej problematyczne okazało się założenie o ciągłości skal z uwagi na małą „gęstość" wyników w zakresie wysokiego poziomu osiągnięć. Rysunek 1 przedstawia rozkłady wyskalowanych wyników testu OBUT i sprawdzianu. Duża łatwość testu OBUT spowodowała, że nie różnicuje on dobrze poziomu osiągnięć uczniów zdolnych. Ich wyniki zawierają się w kilku wartościach punktowych, znacznie od siebie oddalonych (zbliżenie kształtu rozkładu wyników do rozkładu normalnego jest pożądaną konsekwencją zastosowanego modelu skalowania). W zakresie wyników wysokich skala ma więc charakter skokowy. W przypadku sprawdzianu analogiczny problem jest widoczny, choć nie jest tak silny. Wyniki

te pokazują, że trzeba zachować ostrożność w interpretowaniu wskaźników obliczonych na podstawie tych danych (zarówno średnich wyników, jak i miar EWD) w szczególności dla szkół pracujących z uczniami o wyższych osiągnięciach szkolnych.

Rzetelność obu miar wykorzystywanych w modelach EWD jest na akceptowalnym poziomie dla potrzeb budowania wskaźników zagregowanych. Alfa Cronbacha dla testu OBUT wyniosła 0,72, a dla sprawdzianu 0,86. Współczynnik rzetelności EAP/PV dla testu OBUT wyniósł 0,85, a dla sprawdzianu 0,86.

Analiza zróżnicowania międzyszkolnego wyników testu OBUT doprowadziła do niepokojących wniosków. Wariancja międzyszkolna tego pomiaru jest znacząco wyższa niż obserwowana w innych badaniach dla pomiarów po III klasie szkoły podstawowej. W przypadku wyników wszystkich uczniów piszących test OBUT, podział na szkoły wyjaśnia prawie 18% zróżnicowania wyników. Na danych z próby wykorzystywanej do modelowania EWD, współczynnik ten wyniósł niemal 20%. Wartości te są znacznie wyższe niż uzyskane w innych badaniach, w których wariancja międzyszkolna osiągnięć na koniec I etapu edukacyjnego wahała się od 8,6% do 13,7% (Dolata i in. 2014; Jasińska, Modzelewski 2014 – por. Tabela 2). Może to przemawiać za hipotezą, że przynajmniej w części szkół uzyskane wyniki są obciążone nielosowym błędem związanym z różną motywacją testową, niedotrzymaniem reguł przeprowadzania lub oceniania testu OBUT w szkołach.

Tabela 2. Wariancja międzyszkolna osiągnięć uczniów na koniec I etapu edukacyjnego

Dane		Odsetek wariancji międzyszkolnej (w %)
OBUT 2011	populacja	17,7
	próba	19,8
Badanie podłużne EWD (2012)	czytanie	8,6[a]
	świadomość językowa	11,8[a]
	matematyka	10,5[a]
PIRLS 2011	czytanie	12,2[b]
TIMSS 2011	matematyka	13,7[b]

Na podstawie: [a] Dolata i in. (2013); [b] Jasińska i Modzelewski (2013).

Reprezentatywność próby

Średnie wyniki uczniów z próby w obu pomiarach okazały się nieco lepsze niż w populacji, a odchylenia standardowe zbliżone (patrz Tabela 3). Można zatem przyjąć, że próba dobrze odzwierciedla wyniki populacyjne.

Tabela 3. Porównanie danych z próby i populacji

Pomiar	Grupa	Średnia	Odchylenie st.
OBUT 2011	populacja	100,0	15,0
	próba	101,7	15,6
Sprawdzian 2014	populacja	100,0	15,0
	próba	101,4	13,7

Model EWD dla II etapu edukacyjnego

Na tak zgromadzonych danych testowano różne modele EWD dla II etapu edukacyjnego. Najlepszym okazał się model, w którym wyniki testu OBUT zostały uwzględnione jako wielomian 5-go stopnia.

Tabela 4. Współczynniki regresji w modelu EWD

Model	Współczynniki niestandaryzowane		Współczynniki standaryzowane	*t*	Istotność
	B	Bł. st.	Beta		
(Stała)	100,069	0,194		516,681	$p < 0,001$
Płeć	1,973	0,223	0,072	8,859	$p < 0,001$
Dysleksja	3,388	0,339	0,082	9,996	$p < 0,001$
OBUT	11,585	0,235	0,876	49,381	$p < 0,001$
$OBUT^2$	–1,130	0,138	–0,151	–8,158	$p < 0,001$
$OBUT^3$	–0,999	0,107	–0,428	–9,308	$p < 0,001$
$OBUT^4$	0,060	0,015	0,081	4,153	$p < 0,001$
$OBUT^5$	0,045	0,008	0,211	5,905	$p < 0,001$

W modelu tym funkcja łącząca wyniki na sprawdzianie z wynikami testu OBUT jest niemalejąca (w przeciwieństwie do modeli z niższą potęgą wielomianu) i wszystkie współczynniki regresji są istotne statystycznie (por. Tabela 4). Dopasowanie modelu do danych jest dobre – R-kwadrat wyniósł 0,45.

Podstawowe właściwości wskaźników EWD

Obliczone z opisanego powyżej modelu wskaźniki EWD skorelowano ze średnim wynikiem testu OBUT dla szkoły. Analiza ta miała pokazać, czy obliczone wskaźniki EWD są niezależne od tego, z jakimi uczniami pracuje szkoła. Niestety stwierdzono istnienie silnego negatywnego związku między tymi zmiennymi – korelacja Pearsona jest równa –0,37. Wynik ten okazał się zaskakujący, bowiem badania właściwości wskaźników EWD dla gimnazjów, liceów i techników wskazują na istnienie odwrotnej relacji (Żółtak 2013). Dla tych szkół obserwuje się słabszy, jednak pozytywny związek średniej egzaminu „na wejściu" ze wskaźnikami EWD (w szczególności w szkołach o wyższych wynikach), a zjawisko to tłumaczy się problemem selektywności szkół. W przypadku testu OBUT zaobserwowany związek może występować w sytuacji, w której w pewnych szkołach wyniki tego pomiaru są zawyżone. W efekcie w placówkach tych trudno o dodatnią EWD, gdyż obliczana jest ona na podstawie nierealistycznie wysokich uprzednich osiągnięć. W powiązaniu z informacją o większej niż oczekiwana wariancji międzyszkolnej wyników tego testu, hipoteza ta wydaje się całkiem prawdopodobna.

Podsumowanie i dyskusja

Głównym celem testów diagnostycznych o niskiej doniosłości jest przekazanie nauczycielom, rodzicom i uczniom profesjonalnej informacji o poziomie osiągnięć szkolnych uczniów. Ich wyniki mogą również być podstawą wartościowych danych o efektach i efektywności nauczania w szkołach, potrzebnych placówkom do świadomego planowania pracy i oceny skuteczności podejmowanych działań. Cele te mogą być jednak zrealizowane tylko wówczas, gdy wykorzystane narzędzia są odpowiedniej jakości a wyniki pomiarów są wiarygodne.

Przedstawione w artykule wyniki pokazały jednak, że wiarygodność danych z testu OBUT może być podważona, przynajmniej w części szkół. Wnioski te

są spójne ze spostrzeżeniami niektórych nauczycieli klas IV–VI szkół podstawowych, którzy: „od początku wyrażali wątpliwości, czy wyniki Ogólnopolskiego Badania Umiejętności Trzecioklasistów są na tyle wiarygodne, że mogą być podstawą do określania wyników przewidywanych na sprawdzianie po szóstej klasie. Małe zaufanie wynikało z przekonania, że w zbyt dużym stopniu na wyniki wpływał czynnik ludzki – przeprowadzanie badania w obecności nauczyciela prowadzącego, sprawdzanie przez nich prac, rywalizacyjna postawa wobec wyników badania, itp." (cytat z raportu z badania jakościowego z jednej ze szkół, za: Pfeiffer 2015: 22). Tak więc pomiar z założenia o niskiej doniosłości okazał się testem wysokiej stawki dla nauczycieli przynajmniej części szkół, co mogło doprowadzić do zniekształcenia danych. Wyniki te pokazały, że przynajmniej niektórzy nauczyciele nie postrzegają diagnozy osiągnięć szkolnych uczniów jako szansy na uzyskanie wartościowych danych o własnej pracy mogących im pomóc w rozwoju zawodowym. Nie czują, że współuczestniczą w procesie zbierania ważnych dla nich informacji. Jest to w ich oczach zewnętrzne, narzucone przez kogoś narzędzie kontroli czy opresji. Postawę taką może nasilać brak pewności własnych kompetencji dydaktycznych lub brak zaufania do przełożonych lub innych odbiorców wyników testowych (rodziców, innych nauczycieli, organu prowadzącego). W takiej kulturze testowania trudno o wiarygodne dane.

Zmiana procedur realizacji tego typu badania i kodowania testów mogłaby poprawić jakość danych. Niektóre szkoły same wyszły z inicjatywą rozwiązań korzystnie przekładających się na trafność pomiaru, zlecając przeprowadzenie badania i sprawdzanie prac innym nauczycielom ze szkoły (zamiast nauczającym dane oddziały) (Pfeiffer 2015). Wiarygodność wskaźników byłaby większa, gdyby zostały one obliczone na podstawie losowej próby szkół, w których zarówno realizacja, jak i kodowanie testów byłyby powierzone osobom z zewnątrz. Rozwiązanie to wprowadzono do organizowanego w 2015 roku przez Instytut Badań Edukacyjnych pomiaru kompetencji trzecioklasistów.

Wydaje się jednak, że dla jakości wyników kluczowe znaczenie ma nastawienie użytkowników testów. Jeśli nauczyciele będą postrzegać je jako narzędzie służące „rozliczalności", a nie jako źródło rzetelnych danych pozwalających na profesjonalne planowanie własnej pracy, nieuczciwości i wypaczenia będą się zdarzały. Warto podkreślić, że na stosunek nauczycieli wpływa w dużej mierze sposób wykorzystania wyników testów przez innych odbiorców. Ważne jest więc zarówno mądre konstruowanie wskaźników, jak i mądry sposób ich wykorzystania.

Wartościowe mogłyby być też badania nad możliwością wykorzystania metod wykrywania oszustw związanych z sytuacją testową (Jacob, Levitt 2003; Karabatsos 2003) do oczyszczenia danych z badania OBUT przed wykorzystaniem

ich do modelowania. Autorzy podjęli takie próby, jednak nie wpłynęły one na zmianę postaci modelu. Pogłębienie tych analiz jest wskazane.

Modele EWD dla drugiego etapu nauczania zostały obliczone z wykorzystaniem prezentowanego w artykule podejścia w ramach prac nad rozwojem metody w Polsce. Szkoły podstawowe mogą korzystać z przygotowanego dla nich narzędzia do analizy wyników[4]. Użytkownicy metody powinni jednak mieć na uwadze to, że daje ona o tyle wartościową informację, o ile pomiar osiągnięć testem OBUT został przeprowadzony w sposób rzetelny i uczciwy.

Bibliografia

ACARA. 2014. *NAPLAN Achievement in Reading, Persuasive Writing, Language Conventions and Numeracy: National Report for 2014*, Australian Curriculum, Assessment and Reporting Authority, Sydney.

Adams R.J. 2005. *Reliability as a measurement design effect*, „Studies in Educational Evaluation", nr 312–313, s. 162–172.

Balázsi I. 2006. *National Assessment of Basic Competencies in Hungary*, Center for Evaluation Studies, dostęp: http://www.iaea.info/documents/paper_1162a1d92d.pdf [08.10.2015].

Ballou D. 2009. *Test Scaling and Value-Added Measurement*, „Education Finance and Policy", nr 4 (4), s. 351–383.

Bąbiak I., Matuszczak K., Zielonka P. 2013. *Raport z ewaluacji wewnętrznej projektu EWD*, Instytut Badań Edukacyjnych, Warszawa.

Chudowsky N., Koenig J.A., Braun H.I. 2010. *Getting value out of value-added report of a workshop*, National Academies Press, Washington.

Centralna Komisja Egzaminacyjna 2011. *Ogólnopolskie badanie umiejętności trzecioklasistów OBUT 2011. Przewodnik przeprowadzającego badanie*, CKE, Warszawa.

Corcoran S.P., Jennings J. i L., Beveridge A.A. 2011. *Teacher Effectiveness on High- and Low-Stakes Tests*, Society for Research on Educational Effectiveness, New York.

De Ayala R.J. 2009. *The theory and practice of item response theory*, Guilford Press, New York.

Dolata R., Hawrot A., Humenny G., Jasińska A., Koniewski M., Majkut P., Żółtak T. 2013. *Trafność metody edukacyjnej wartości dodanej dla gimnazjów*, Instytut Badań Edukacyjnych, New York.

Dolata R., Hawrot A., Humenny G., Jasińska-Maciążek A., Koniewski M., Majkut P. 2014. *Kontekstowy model efektywności nauczania po pierwszym etapie edukacyjnym. Edukacyjna Wartość Dodana*, Instytut Badan Edukacyjnych, Warszawa.

Dolata R., Pokropek A. 2010. *Motywacja a wynik testu z nauk przyrodniczych. Studium na przykładzie PISA 2006*, [w:] *Teraźniejszość i przyszłość oceniania szkolnego: XVI Krajowa Konferencja Diagnostyki Edukacyjnej*, red. B. Niemierko, M.K. Szmigel, Grupa Tomami, Kraków.

[4] http://ewd.edu.pl/pobierz-2/

Dolata R., Szaleniec H. 2012. *Funkcje krajowych egzaminów w systemie edukacji*, „Polityka Społeczna", nr tematyczny 1, s. 37–41.

Domański H., Pokropek A. 2011. *Podziały terytorialne, globalizacja a nierówności społeczne: wprowadzenie do modeli wielopoziomowych*, Wydawnictwo IFiS PAN, Warszawa.

EACEA. 2010. *Ogólnokrajowe egzaminowanie uczniów w Europie: cele, organizacja i wykorzystanie wyników*, tłum. M. Kowalczyk, Fundacja Rozwoju Systemu Edukacji, Warszawa.

Freeman C. 2009. *First national literacy and numeracy tests introduced*, „Research Developments", nr 20 (20), s. 4.

Herczyński J., Borek A., Stożek E. 2012. *Ocena potrzeb szkoły na podstawie osiągnięć uczniów oraz wyników ewaluacji zewnętrznej*, [w:] *Finansowanie oświaty*, red. M. Herbst, t. 3, Biblioteczka Oświaty Samorządowej, Warszawa, s. 171–232.

Hungarian Eurydice Unit. 2009. *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results. Hungary*, Hungarian Eurydice Unit, Bruksela.

Jacob B.A., Levitt S.D. 2003. *Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating*, „The Quarterly Journal of Economics", nr 118 (3), s. 843–877.

Jager D.J., Merki K.M., Oerke B., Holmeier M. 2012. *Statewide Low-Stakes Tests and a Teaching to the Test Effect? An Analysis of Teacher Survey Data from Two German States*, „Assessment in Education: Principles, Policy Practice", nr 19 (4), s. 451–467.

Jasińska A., Modzelewski M. 2013. *Międzyszkolne zróżnicowanie wyników nauczania po pierwszym etapie kształcenia*, [w:] *Polska edukacja w świetle diagnoz prowadzonych z różnych perspektyw badawczych: XIX Krajowa Konferencja Diagnostyki Edukacyjnej*, red. B. Niemierko, M.K. Szmigel, Grupa Tomami, Kraków, s. 165–178.

Jasińska A., Modzelewski M. 2014. *Testy osiągnięć szkolnych TOS3: przykład narzędzia skonstruowanego z wykorzystaniem modelu Rascha*, „Edukacja", nr 2 (127), s. 85–107.

Jasińska-Maciążek A., Modzelewski M. 2014. *Modele analizy zróżnicowania wyników nauczania*, [w:] R. Dolata, *Czy szkoła ma znaczenie? Analiza zróżnicowania efektywności nauczania na pierwszym etapie edukacyjnym*, Instytut Badań Edukacyjnych, Warszawa.

Karabatsos G. 2003. *Comparing the aberrant response detection performance of thirty-six person-fit statistics*, „Applied Measurement in Education", nr 16 (4), s. 277–298.

Karwowski M. red. 2013. *Ścieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne: trafność wskaźników edukacyjnej wartości dodanej dla szkół maturalnych*, Wydawnictwo Instytutu Filozofii i Socjologii Polskiej Akademii Nauk, Warszawa.

Konarzewski K. 2012. *TIMSS i PIRLS 2011. Osiągnięcia szkolne polskich trzecioklasistów w perspektywie międzynarodowej*, CKE, Warszawa.

MacBeath J., Schratz M., Meuret D., Jakobsen L. 2005. *Czy nasza szkoła jest dobra?*, tłum. K. Kruszewski, Wydawnictwa Szkolne i Pedagogiczne, Warszawa.

Matuszczak K., Wasilewska O. 2015. *Wyniki egzaminów zewnętrznych w pracy szkoły*, [w:] *Egzaminy zewnętrzne w polityce i praktyce edukacyjnej. Raport o stanie edukacji 2014*, red. R. Dolata, M. Sitek, Instytut Badań Edukacyjnych, Warszawa.

McCaffrey D.F., Lockwood J.R., Koretz D.M., Hamilton L.S. 2003. *Evaluating Value-Added Models for Teacher Accountability*, RAND Corporation, Santa Monica, CA.

National Foundation for Educational Research. 2009. *National Testing of Pupils in England, Wales and NorthernIreland*, dostęp: https://www.nfer.ac.uk/shadomx/apps/fms/fmsdownload.cfm?file_uuid=67EAAF91-C29E-AD4D-07F1-A9373EA17105&siteName=nfer.

Nichols S.L., Berliner D. 2007. *The pressure to cheat in a high-stakes testing environment*, [w:] *Psychology of academic cheating*, red. E.M. Anderman, T.B. Murdock, Elsevier Academic Press, Amsterdam–Boston, s. 289–311.

Niemierko B. 2009. *Diagnostyka edukacyjna: podręcznik akademicki*, Wydawnictwo Naukowe PWN, Warszawa.

OECD. 2008. *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*, Organisation for Economic Co-operation and Development, Paris.

OECD. 2013. *Synergies for better learning: an international perspective on evaluation and assessment*, Organisation for Economic Co-operation and Development, Paris.

Papay J.P. 2011. *Different Tests, Different Answers The Stability of Teacher Value-Added Estimates Across Outcome Measures*, „American Educational Research Journal", nr 481, s. 163–193.

Pfeiffer A. 2015. *Raport zbiorczy z badania uczestniczącego w szkołach podstawowych*, Instytut Badań Edukacyjnych, Warszawa.

Pregler A., Wiatrak E. red. 2011. *Ogólnopolskie Badanie Umiejętności Trzecioklasistów Raport OBUT 2011*, CKE, Warszawa.

Raudenbush S.W. 2004. *What are value-added models estimating and what does this imply for statistical practice?*, „Journal of Educational and Behavioral Statistics", nr 291, s. 121–130.

Raudenbush S.W., Willms J.D. 1995. *The Estimation of School Effects*, „Journal of Educational and Behavioral Statistics", nr 20 (4), s. 307–335.

Reardon S.F., Raudenbush S.W. 2009. *Assumptions of value added models for estimating school effects*, „Education Finance and Policy", nr 4 (4), s. 492–519.

SEO. 2013. *Wymagania państwa wobec szkół*, dostęp: http://www.npseo.pl/action/requirements/wymagania_panstwa_wobec_szkol.

Sinka E. 2010. *OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes. Hungary. Country Background Report*, Hungary, dostęp: http://www.oecd.org/edu/school/50484774.pdf.

Skórska P., Świst K., Pokropek A. 2014. *Indywidualne i grupowe efekty motywacji testowej uczniów*, [w:] *Diagnozy edukacyjne: dorobek i nowe zadania: XX Krajowa Konferencja Diagnostyki Edukacyjnej*, red. B. Niemierko, M.K. Szmigel, Grupa Tomami, Kraków, s. 161–171.

Stożek E. 2012. *Czy egzaminy mogą pomóc szkole w rozwoju?*, „Polityka Społeczna", nr tematyczny 1, s. 30–33.

Szaleniec H. 2004. *Jak wykorzystywać wyniki egzaminów zewnętrznych*, WSiP, Warszawa.

Szaleniec H. red. 2009. *Teoria wyniku zadania IRT: zastosowania w polskim systemie egzaminów zewnętrznych: praca zbiorowa*, CKE, Warszawa.

Vigneau E., Qannari E.M. 2003. *Clustering of Variables Around Latent Components*, „Communications in Statistics – Simulation and Computation", nr 32 (4), s. 1131–1150.

Wasilewska O., Rybińska A., Muzyk A. 2014. *Wykorzystanie ewaluacji zewnętrznej i wewnętrznej przez szkoły*, Instytut Badań Edukacyjnych, Warszawa.

Wise S.L., DeMars C.E. 2005. *Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions*, „Educational Assessment", nr 10 (1), s. 1–17.

Wolf L.F., Smith J.K. 1995. *The Consequence of Consequence: Motivation, Anxiety, and Test Performance*, „Applied Measurement in Education", nr 8 (3), s. 227–242.

Żółtak T. 2013. *Statystyczne modelowanie wskaźników edukacyjnej wartości dodanej – podsumowanie polskich doświadczeń (Raport tematyczny z badania)*, Instytut Badań Edukacyjnych, Warszawa.