ALICJA ZAWISTOWSKA

Uniwersytet w Białymstoku

NIEZAMIERZONE SKUTKI STOSOWANIA TESTÓW WYSOKIEJ STAWKI. LEKCJA Z AMERYKAŃSKIEJ REFORMY NCLB[1]

ABSTRACT. Zawistowska Alicja, *Niezamierzone skutki stosowania testów wysokiej stawki. Lekcja z amerykańskiej reformy NCLB* [Unintended Consequences of High-Stakes Testing. Lesson from NCLB]. Studia Edukacyjne nr 42, 2016, Poznań 2016, pp. 39-64. Adam Mickiewicz University Press. ISSN 1233-6688. DOI: 10.14746/se.2016.42.3

Educational decision makers willingly draw on solutions adopted in other countries. It was also the case in Polish educational reform started in late 90s. Since the introduction of the reform, Poland joined countries whose educational system is divided into three levels, each ending with an exit exams and core curriculum is set to teaching standards. The exams seem to be the most important element of the Polish reform. While the designers of educational policies are often inspired by the experiences of other countries during the planning phase, they are less willing to learn from them when it comes to predicting outcomes of the reform. A good case to analyze potential consequences of high-stakes testing is United States, where standardized tests have been administered since the beginning of the era of mass education. In this paper I will analyze the effects of the last, most controversial federal reform, commonly known as *No Child Left Behind* introduced in 2002. Findings of the study might be used to predict potential unintended effects of using the high stakes tests for accountability policy. The article addresses the problem of test scores inflation as well as the factors which may accelerate it.

Key words: NCLB, inflation, accountability, exit exams, reform

Wprowadzenie

Wprowadzenie w Polsce systemu standaryzowanych egzaminów zewnętrznych było głównym elementem zmian rozpoczętych w 1999 roku.

[1] Artykuł powstał w zakresie projektu „Niezamierzone skutki reform edukacyjnych", zrealizowanego w ramach stypendium przyznanego przez Polsko-Amerykańską Fundację Fulbrighta i przygotowanego w Stanford Univeristy.

Pierwszy egzamin tego typu gimnazjaliści pisali w 2002 roku, a w kolejnych latach ujednolicony test zastąpił wcześniejszą formułę matury pisemnej. Obydwa egzaminy, zarówno gimnazjalny, jak i ten pisany na koniec szkoły średniej, należą do testów wysokiej stawki, a więc takich, które wpływają na dalszą karierę edukacyjną uczniów. Egzamin gimnazjalny, mimo że w założeniu nie ma na celu selekcji uczniów, jest „biletem wstępu" do szkoły średniej, a więc kluczowego progu selekcji w całej karierze. Wysoka stawka egzaminu maturalnego wyraża się natomiast w tym, że ma próg zaliczenia, a liczba uzyskanych punktów wpływa na możliwość wyboru uczelni wyższej. Naukowe badania na temat wpływu tej formy egzaminów na funkcjonowanie szkół koncentrują się głównie na dwóch problemach. Z jednej strony, analizie poddawane są praktyki szkolne opisujące, jak uczniowie, nauczyciele i kadra zarządzająca szkołą przystosowują się do zmian[2]. Drugi obszar badań dotyczy właściwości samych testów – analizy odpowiadają na pytania dotyczące ich rzetelności, trafności oraz szukają różnego rodzaju obciążeń występujących w zadaniach testowych[3]. Lektura tych i innych prac wskazuje, że system egzaminów zewnętrznych ciągle jest na etapie budowy, a szkoły stopniowo przechodzą proces adaptacji do zmieniających się warunków. Niewiele wiadomo na razie na temat potencjalnych ukrytych efektów stosowania testów w Polsce, ale doświadczenia krajów, w których funkcjonują one od dziesięcioleci nie pozostawiają wątpliwości, że towarzyszą temu zawsze różnego rodzaju niezamierzone „efekty uboczne". Do najczęściej wymienianych efektów z zakresu praktyk dydaktycznych należy uczenie pod test, do testu, czy selekcja treści ważniejszych z punktu widzenia egzaminu kosztem pozostałych. Nasilenia tych praktyk można spodziewać się szczególnie tam, gdzie wprowadzeniu egzaminów testowych towarzyszy również zmiana polityki rozliczalności, polegająca na przypisaniu testom większej roli. Stawka testu rośnie wtedy, gdy na podstawie jego wyników podejmuje się decyzje o odebraniu lub przekazaniu dodatkowych środków finansowych szkołom albo wręczeniu premii finansowych nauczycielom za dobre wyniki uczniów. Krajem, który jest najbardziej doświadczony w mierzeniu się z tymi kwestiami są Stany Zjednoczone, w których testy jako narzędzie służące do pomiaru wiedzy uczniów stosowane są na masową skalę od lat 50. XX wieku. W niniejszym artykule przeanalizuję wnioski płynące z tych doświadczeń i przedyskutuję je w kontekście zmian w polskim systemie oświaty. Posłużyć to może do przewidzenia potencjalnych, nieoczekiwanych skutków stosowania testów wysokiej stawki w naszym kraju.

[2] K. Konarzewski, *Przygotowanie uczniów do egzaminu: pokusa łatwego zysku*, Warszawa 2008; J. Choińska-Mika i in., *Realizacja podstawy programowej z historii w gimnazjach*, Warszawa 2013.

[3] A. Pokropek, *Matura z języka polskiego. Wybrane problemy psychometryczne*, Kraków 2011.

Oczywiście, przełożenia amerykańskiej perspektywy do polskich warunków można dokonać tylko w pewnym zakresie, ponieważ systemy szkolne tych krajów różnią się w zbyt wielu aspektach. Znacząco odmienne są też czynniki społeczne mające wpływ na edukację, jak na przykład nierówności społeczne, gospodarka, czy struktura etniczna. Inna jest też konstrukcja prawno-organizacyjna, która daje każdemu stanowi (czasem również dystryktowi) przywilej samodzielnego, i niezależnego od innych stanów oraz centralnego rządu, kształtowania systemu edukacji. Różnice te nie powinny być jednak przeszkodą w odszukiwaniu paralel w skutkach omawianych tu reform. Rdzeń każdego systemu szkolnego składa się z podobnych komponentów, a instytucje szkolne wszędzie działają według analogicznych zasad[4]. Można oczekiwać więc, że w dobie globalnej uniwersalizacji zasad funkcjonowania instytucji, reakcje „organizmu" na podobne „bodźce" będą zbliżone. W artykule tym przeanalizuję niektóre skutki reformy nazywanej potocznie *No Child Left Behind* dla funkcjonowania amerykańskich szkół, ze szczególnym uwzględnieniem problemu inflacji wyników testów oraz wpływu tej reformy na nierówności edukacyjne. Pomijam tym samym inne istotne dla amerykańskiej edukacji problemy, wśród których utworzenie jednolitych narodowo standardów zajmuje czołowe miejsce.

Dojrzewanie systemu testowego w USA

Egzaminy testowe zaczęto stosować w amerykańskich szkołach dopiero wówczas, kiedy udowodniły swoją użyteczność w armii i badaniach klinicznych na początku XX wieku[5]. Skutkiem narastającego przekonania, że standaryzowany test jest lepszym narzędziem pomiaru umiejętności uczniów, zidentyfikowania przyczyn problemów i sortowania uczniów, niż inne formy ewaluacji, było powstanie kilku pozarządowych organizacji, których celem było tworzenie i doskonalenie standaryzowanych testów. Jakość edukacji, jak wierzono, można podnieść właśnie poprzez stosowanie precyzyjnego pomiaru umiejętności, dzięki któremu możliwa byłaby eliminacja problemów trapiących szkoły. Organizacje te, wśród nich *Educational Testing Service*, skupiły wokół siebie grono wybitnych statystyków i specjalistów od psychometrii, którzy wnieśli istotny wkład do współczesnej teorii testów,

[4] D.B. Tyack, L. Cuban, *Tinkering toward utopia: a century of public school reform*, Cambridge 1995; H. Meyer, B. Rowan, *The new institutionalism in education*, Albany 2006.

[5] C.J. Gallagher, *Reconciling a Tradition of Testing with a New*, Educational Psychology Review, 2003, 15(1), s. 83-99.

między innymi Frederica Lorda czy Donalda Rubina[6]. Jednak wraz z tym, jak „ojcowie nowoczesnych testów" dopracowywali metodologię pomiarową, coraz częściej mówiono o kryzysie panującym w amerykańskich szkołach. Za podstawowy problem uznano niewystarczającą efektywność szkół, a ówczesne tło historyczne miało dostarczać ilustracji. W 1957 roku, kiedy Związek Radziecki jako pierwszy w historii wysłał sztucznego satelitę na orbitę okołoziemską, amerykańskie elity dość jednogłośnie orzekły, że przyczyną porażki w rywalizacji z ZSRR był właśnie zły stan szkolnictwa. W odpowiedzi na rosnącą przewagę oponenta zza żelaznej kurtyny, rząd amerykański zainicjował kilka reform i inicjatyw edukacyjnych, których nadrzędnym celem było wzmocnienie konkurencyjności gospodarki. Na fali „postsputnikowych" zmian zwiększono między innymi nakłady finansowe na przedmioty ścisłe i informatykę oraz zaoferowano większe wsparcie dla zdolnych uczniów. W tym okresie znaczenie standaryzowanych testów było jeszcze stosunkowo małe, ze względu na ich diagnostyczne przeznaczenie. W latach 50. były tworzone i używane przede wszystkim w celu zidentyfikowania słabych stron procesu nauczania oraz miały stanowić rzetelne źródło informacji potrzebnej do rozwiązania konkretnych problemów szkolnych[7]. Funkcja testów, z diagnostycznej na selekcyjną, zaczęła się powoli zmieniać w latach 60. i 70. W tych dekadach rząd federalny poprzez serię ustaw wprowadził konieczność ewaluacji postępów uczniów z użyciem standaryzowanych testów oraz powołał program okresowej oceny umiejętności na podstawie reprezentatywnej próbki (*National Assessment of Educational Progress*). Najważniejszym aktem prawnym tego okresu, jak i całej współczesnej historii edukacji w USA, była jednak ustawa *Elementary and Secondary Act* z 1964 roku. Stanowiła ona komponent programu „Walki z ubóstwem" zainicjowany przez prezydenta Lyndona Johnsona. U podłoża ustawy leżało przekonanie, że szkoły w niewystarczającym stopniu wspierają dzieci defaworyzowane, co nieuchronnie prowadzi do wzrostu zróżnicowania osiągnięć między uczniami o różnym statusie socjoekonomicznym. Szkoły natomiast nie są odpowiednio rozliczane ze swoich działań. Odpowiedzią na ten pierwszy problem miały być zapisy w *Title I*, gdzie mowa była o przekazaniu odpowiednich funduszy do szkół, w których uczy się duży odsetek dzieci z ubogich rodzin. O tym, czy pomoc ta przynosi założone rezultaty miały informować z kolei testy przeprowadzane dwukrotnie w ciągu roku wśród uczniów, którzy z tej pomocy korzystali[8]. Postanowie-

[6] R.A. Horn, *Understanding educational reform: a reference handbook*, Santa Barbara 2002.

[7] D.M. Koretz, *Measuring up: what educational testing really tells us*, Cambridge 2008.

[8] L. Crocker, *Teaching for the test: How and why test preparation in appropriate*, [w:] *Defending standardized testing*, red. R. Phelps, New York 2005, s. 159-174.

niom tym przyświecało zdroworozsądkowe przekonanie, że skoro wydatki publiczne na edukację zostały przekazane na konkretny cel, to szkoły z nich korzystające powinny udowodnić, że spożytkowały je w odpowiedni sposób. Na nauczycieli nałożyło to presję utrzymania pożądanych wyników w grupie ubogich uczniów pod groźbą utracenia dotacji. Pojawiło się zatem pytanie, jak te dobre wyniki osiągnąć.

Ruch reformatorów opowiadający się za powszechniejszym stosowaniem testów wypierał powoli zwolenników utworzenia jasnych standardów określających wiedzę uczniów na poszczególnych etapach nauki – konkurencyjnego podejścia do reformowania edukacji. Kluczowe znacznie dla rozprzestrzenienia się tej idei miał raport *Nation at Risk*, przygotowany przez specjalnie powołaną do tego komisję rządową w 1983 roku. Była to analiza problemów amerykańskich szkół, do których zaliczono między innymi malejące wyniki w teście SAT, słabe przygotowanie nauczycieli, czy niską jakość podręczników. Argumentowano, że amerykańscy uczniowie wypadali blado na tle rówieśników z Japonii, a rozwój Związku Radzieckiego stanowi realne zagrożenie dla powodzenia USA. Raport dawał też upust niezadowoleniu z efektów, jakie przyniosły programy przeciwdziałające nierównościom, które mimo wysokich kosztów nie zmniejszyły luki w osiągnięciach uczniów z różnych szczebli struktury społecznej. Zawierał również szereg sugestii dotyczących kierunków, w jakich powinny zmierzać amerykańskie reformy – nawoływano do tego, aby edukacja w większym niż dotychczas stopniu skupiła się na nauce pięciu podstawowych dyscyplin, a postępy uczniów i efektywność szkół powinny być weryfikowane w bardziej rygorystyczny sposób[9]. Pesymistyczny ton raportu sprawił, że jego treść w istotny sposób wpłynęła na kierunek decyzji podejmowanych w kolejnych latach, choć on sam nie miał charakteru aktu prawnego. Szczególnie uważnie został wysłuchany apel dotyczący większej dyscypliny w zakresie oceny uczniów i efektywności szkół.

Postulaty przedstawione w *Nation at Risk* stały się kluczowe dla planów ówczesnego wiceprezydenta USA – George'a H.W. Busha. Wśród jego pomysłów na naprawę systemu edukacji znalazł się między innymi taki, który polegał na utworzeniu narodowych, wspólnych dla całego kraju standardów nauczania i wprowadzenia dobrowolnych testów z kilku podstawowych przedmiotów, które miały być przeprowadzane trzykrotnie w ciągu dwunastoletniej nauki obejmującej szkołę podstawową i średnią[10]. Projekt *Ameryka 2000* nie znalazł jednak poparcia wśród członków Kongresu i nie zaowoco-

[9] G.M. Jones, B.D. Jones, T.Y. Hargrove, *The unintended consequences of high-stakes testing*, Oxford 2003.

[10] F.M. Hess, M.J. Petrilli, *No Child Left Behind Primer*, New York 2006.

wał aktem prawnym. Jednak w kraju, w którym ministerstwo edukacji na poziomie centralnego rządu utworzone zostało w roku 1980, a i później nie wszyscy prezydenci widzieli sens jego istnienia, reformy typu „top-bottom" traktowane są często jako zagrożenie stanowej niezależności, gwarantowanej w formule państwa federacyjnego. Koncepcje zaproponowane przez G.H.W. Busha nie zostały jednak porzucone. Podobna filozofia przyświecała bowiem inicjatywom podejmowanym przez kolejnych prezydentów. Coraz szerzej podzielane było przekonanie, że szkoły powinny stworzyć wspólne dla całego kraju standardy nauczania, podnieść poziom rozliczalności i w bardziej drobiazgowy sposób dokonywać pomiaru umiejętności. Wobec braku zgody na stworzenie jednolitych standardów, za kadencji Billa Clintona, do ustawy o oświacie z 1965 roku wprowadzono zmianę, mówiącą że każdy stan powinien ustalić własne standardy nauczania dla szkół podstawowych i średnich oraz co roku, z użyciem dobrowolnych testów, sprawdzić, czy uczniowie osiągnęli te wymogi. Określono również, jak duży miał być postęp w ciągu jednego roku szkolnego. Wprowadzenie tych zmian zostało jednak dość szybko wstrzymane, między innymi dlatego, że ze względu na brak grożących sankcji niewiele stanów zastosowało się do tych zaleceń. Bez większego echa przeszła również propozycja wysunięta kilka lat później, kiedy administracja Clintona chciała wprowadzić zapis głoszący, że wyniki uczniów powinny być przedstawiane w podziale na przynależność etniczną, status socjoekonomiczny ucznia i umiejętność posługiwania się językiem angielskim. Również ta decyzja, z powodu braku porozumienia w kwestii finansowania, nie została wdrożona. Na realne zmiany amerykańskie szkoły nie musiały jednak czekać długo. W kampanii prezydenckiej w 2000 roku George W. Bush zaproponował bowiem zaimplementowanie w całym kraju rozwiązań, które w ostatnich latach funkcjonowały z powodzeniem (jak wtedy sądzono) w stanie Teksas. Opierały się one – mówiąc w uproszczeniu – na zasadzie „kija i marchewki", a więc stosowaniu kilku testów wysokiej stawki z poważnymi konsekwencjami dla szkół. Propozycje te nie brzmiały dla amerykańskich obywateli szczególnie szokująco. W latach 80. i 90., w konsekwencji rozprzestrzeniającej się koncepcji rozliczalności szkolnej spopularyzowanej za sprawą *Nation at Risk*, blisko połowa stanów stopniowo zastąpiła bardziej twórcze formy ewaluacji wiedzy, takie jak prace badawcze czy eseje, standaryzowanymi testami oraz wzmacniała ich znaczenie w procesie nauki[11]. Jednocześnie w opinii publicznej narosło przekonanie, że niska jakość edukacji jest wynikiem nieudolnego zarządzania szkołami na

[11] L. Darling-Hammond, F. Adamson, *Beyond the bubble test: how performance assessments support 21st century learning*, San Francisco 2014.

poziomie lokalnym i tylko zewnętrzna presja ze strony organu federalnego może poprawić ten stan rzeczy. Poczucie to wzmacniane było dodatkowo frustracją wywołaną dekadami kosztownych, ale nie przynoszących oczekiwanych rezultatów, reform. Okazało się bowiem, że miliony dolarów przeznaczone w latach 70. i 80. na programy kompensacyjne dla ubogich uczniów nie dały oczekiwanych efektów[12]. Skoro więc zwiększenie funduszy nie pomogło amerykańskim uczniom znaleźć się w czołówce międzynarodowych rankingów, a luka między uczniami o różnym pochodzeniu etnicznym i statusie materialnym nadal była ogromna, decydenci postawili podjąć bardziej radykalne kroki. Polegały one na rezygnacji z obranego w latach 80. i 90. kursu zmierzającego do stworzenia standardów nauczania (*standards-based accountability*) na rzecz stosowania polityki rozliczalności opartej na wynikach testów (*test-based accountability*). Był to jednocześnie zwrot w stronę bardziej restrykcyjnego nadzoru nad szkołami.

Test nie ominie nikogo

W 2002 roku nowo powołany prezydent George W. Bush podpisał ustawę *No Child Left Behind*. Jej celem było zapewnienie równego dostępu do nauki każdemu, bez względu na poziom dochodu, płeć, przynależność etniczną, niepełnosprawność, czy poziom znajomości języka angielskiego – żadne dziecko miało „nie zostać pozostawione samo sobie" w szkole. Po raz kolejny idea równego dostępu do edukacji – niemal obsesyjnie powracająca przy okazji instalowania się każdej nowej administracji w Białym Domu – powróciła, ale tym razem zamierzano ją zrealizować w nieco inny sposób niż poprzednio.

Licząca ponad tysiąc stron ustawa NCLB szczegółowo opisywała pakiet zmian, jakim poddane zostaną w najbliższym czasie szkoły. Jedną z najważniejszych było zwiększenie liczby obowiązkowych testów szkolnych. Zgodnie z ustawą, każdy uczeń w klasie od 4. do 8. oraz w 12. roku nauki miał przystąpić do testu z czytania, matematyki (te przedmioty uznano za najważniejsze predyktory sukcesu na rynku pracy) oraz dodatkowo kilka razy w tym okresie nauki – z przedmiotów ścisłych. Liczba punktów uzyskanych w testach, zgodnie z ideą większej dostępności do informacji, miała być sprawozdawana przez szkoły w podziale na grupy etniczne, niepełnosprawność uczniów, status materialny i znajomość języka angielskiego –

[12] G.J. Duncan, R.J. Murnane, *Whither opportunity? Rising inequality, schools, and children's life chances*, New York 2011.

w tym ostatnim kryterium chodziło głównie o monitoring postępów nowo przybyłych imigrantów.

Miarą efektywności szkół miał być odsetek uczniów, którzy każdego roku osiągali w testach poziom *proficiency*, a więc środkowy szczebel osiągnięć – między podstawowym a zaawansowanym. Jednak ustalenie, gdzie konkretnie wypada on w faktycznym rozkładzie wyników, podobnie jak wybór testu, z użyciem którego określano poziom umiejętności uczniów, pozostały w rękach poszczególnych stanów. Każdy stan mógł więc używać innego testu i w każdym poziom *proficiency* odnosić się mógł do innego poziomu wyników ucznia. Władze stanowe same miały też ustalić, ilu uczniów każdego roku powinno osiągnąć ten pułap (wyrażono to w tzw. *Adequate Yearly Progress*), ale istotne było, aby do 2014 roku 100% uczniów go osiągnęło. Ujmując to inaczej – wyniki wszystkich uczniów do tego roku powinny być co najmniej na przeciętnym poziomie.

Reforma zainicjowana przez prezydenta G.W. Busha brzmiała pod wieloma względami podobnie, jak propozycje składane przez jego poprzedników i senatorów tego okresu. Nie był nowością ani postulat powszechniejszego stosowania testów, ani ustalenia „kwot" określających, ilu uczniów każdego roku powinno osiągnąć określony poziom. Istotna różnica dotyczyła natomiast sankcji, jakie pociągało za sobą niewywiązanie się z tych zobowiązań. Szkoły, w których uczniowie nie osiągnęli wystarczającego postępu były zagrożone sankcjami ze strony rządu stanowego, których dotkliwość miała nasilać się w miarę odstępstw od realizacji rocznych planów. Kiedy przez dwa lata odpowiedni odsetek uczniów nie osiągnął poziomu *proficiency*, szkoła musiała przygotować i wdrożyć własne „plany naprawcze", a uczniowie uzyskiwali prawo do przeniesienia się do innej placówki, również prywatnej, w której ich nauka jest finansowana z budżetu publicznego. Po upływie kolejnych lat szkoły zostające „w tyle" poddawane miały być restrukturyzacji obejmującej zmiany personalne, a w skrajnych przypadkach zamykane.

Ustawa wprowadzała jeszcze szereg drobniejszych zmian, programów i inicjatyw, jak na przykład tę, że podstawą podejmowania decyzji edukacyjnych jest oparcie się na badaniach naukowych, albo mówiącą o dodatkowych zajęciach dla potrzebujących uczniów. Mowa była również o obowiązkowym podnoszeniu kwalifikacji nauczycieli. Nic jednak nie wywołało większego niezadowolenia wśród pracowników oświaty, uczniów i rodziców niż nowe zasady rozliczalności, które wiązały wyniki testów z sankcjami[13].

[13] Interesujące, że ustawa ta w momencie jej głosowania zyskała wystarczające poparcie Demokratów, którzy poprzednio manifestowali swoją niechęć wobec stosowania testów na większą skalę. Jak podkreślają badacze (zob. D. Ravitch, *The Death and Life of the Great American*

Po upływie krótkiego czasu atmosfera względnej jedności, towarzysząca przyjęciu ustawy, zniknęła. Ostra krytyka zaczęła być coraz częściej wyrażana przez nauczycieli, rodziców, a nawet polityków, którzy wcześniej popierali jej wprowadzenie. Proces dydaktyczny, jak zwracali uwagę uczestnicy tej debaty, zaczął w jeszcze większym stopniu ograniczać się do naprzemiennego pisania i przygotowywania się do testów. Rzeczywiście, jeśli tylko wziąć pod uwagę liczbę testów wymaganych przez rząd federalny, to ich liczba po wprowadzeniu NCLB wzrosła niemal trzykrotnie[14]. Testy te stanowią zaś tylko część całej ich puli. Poza nimi istnieją również testy administrowane na poziomie dystryktu. Ogólnie, jak wskazują różne szacunki, w ciągu 12 lat nauki, trwającej od przedszkola do końca szkoły średniej, przeciętny uczeń pisze od 60 do 100 testów. W klasach 3-8 przeciętnie jest to 10 testów rocznie, choć są również szkoły z rekordowym wynikiem 20 testów w ciągu jednego roku szkolnego[15].

Nie dziwi więc, że „przetestowanie", obok wprowadzenia kar dla szkół nie nadążających za rocznymi planami, stało się najczęściej dyskutowanym komponentem reformy. Krytycznie nastawieni wobec niej badacze szczególnie podkreślają fakt, że od początku nie niosła ona ze sobą potencjału zmian. Reforma nie proponowała niczego nowego na poziomie praktyk szkolnych, ani programów nauczania, a fundusze przeznaczone dla potrzebujących uczniów rozpływały się pomiędzy firmami świadczącymi usługi edukacyjne[16]. Jedyną realną zmianą było wprowadzenie silniejszego reżimu rozliczalności[17], w którym wyniki testów decydują o „być albo nie być" danej szkoły. Ustawa przeniosła bowiem prawie całą odpowiedzialność za efektywność kształcenia z uczniów i ich rodziców na szkoły i nauczycieli.

Do zróżnicowanej funkcji testów, które w trakcie swojej wieloletniej „kariery" służyły raz celom diagnostycznym, a innym razem były probierzem szkolnych bolączek lub stanowiły sito selekcji do następnej klasy czy szkoły, doszła nowa polegająca na kształtowaniu polityki edukacyjnej. Łatwo zorientować się, że mechanizm nagradzania za „wydajność" wmontowany

---

School System: How Testing and Choice are Undermining Education, New York 2010), tej solidarności sprzyjały prawdopodobnie wydarzenia z 11 września, po których Kongres częściej manifestował swoją jednomyślność.

[14] D. Goodman, R.K. Hambleton, *Some misconceptions about large-scale educational assessments*, [w:] *Defending standardized testing*, red. R.P. Phelps, New York 2005, s. 99.

[15] M. Lazarin, *Testing Overload in America's Schools*. Center for American Progress 2014. https://cdn.americanprogress.org/wp-content/uploads/2014/10/LazarinOvertestingReport.pdf [dostęp: 27.02.2016].

[16] D. Ravitch, *The Death and Life of the Great American School System*.

[17] L. Darling-Hammond, F. Adamson, *Beyond the bubble test*; D. Ravitch, *The Death and Life of the Great American School System*.

w NCLB wpisuje się w nurt nowego sposobu zarządzania instytucjami publicznymi, które garściami czerpią z zasad obowiązujących w sektorze prywatnym. Czy jednak szkoły, jako instytucje publiczne, zareagują na żądanie rozliczalności tak samo, jak nastawione na zysk korporacje? Wyniki badań są tu skrajnie rozbieżne, co zresztą nie jest zaskakujące wobec dużej złożoności problemu. Pewne jest natomiast, że metoda „kija i marchewki" w wydaniu szkolnym prowadzi do szeregu równoległych, nieoczekiwanych konsekwencji.

Inflacja wyników

Niektóre efekty uboczne stosowania testów wysokiej stawki są dobrze udokumentowane. Wśród nich najczęściej wymieniana jest selekcja nauczanych treści[18]. Może polegać ona na poświęcaniu większej uwagi zagadnieniom pojawiającym się w testach, a pomijaniu tych mniej prawdopodobnych, ale również skupieniu się na łatwych do opanowania zagadnieniach, kosztem bardziej złożonych. Inną konsekwencją jest również istnienie podwyższonego poziomu stresu wśród uczniów i nauczycieli, którzy odczuwają presję spowodowaną powiązaniem wyników z konkretnymi skutkami. Istnieje jednak sporo dowodów, że testy wysokiej stawki wpływają nie tylko na komfort nauczania oraz jego metody, ale również na sam rozkład wyników w nich uzyskiwanych. Wielokrotnie zaobserwowano bowiem, że po upływie pewnego czasu od wprowadzenia nowego testu w danej populacji uczniów, przeciętne wyniki w tym teście rosną[19]. Istnieją uzasadnione podejrzenia, że taki wzrost stanowi wynik inflacji, a zatem, że uzyskany na teście wynik jest zawyżony względem prawdziwej wiedzy uczniów.

Pierwszych dowodów na istnienie inflacji wyników testów dostarczono amerykańskiej opinii publicznej pod koniec lat 80. Już wtedy było to zjawisko znane, ale nie podejmowano go poza literaturą dotyczącą psychometrii. Dyskusja rozpoczęła się za sprawą publikacji raportu Johna Cannella, który zauważył, że wszystkie stany osiągnęły wynik powyżej średniej na teście szkolnym, którego wyniki były znormalizowane dla całego kraju[20]. Zjawisko to przeszło do historii pod nazwą efektu „*Lake Wobegon*". Nazwa pochodzi od tytułu audycji radiowej, opowiadającej o fikcyjnym miasteczku w Minne-

---

[18] Zob. G.M. Jones, B.D. Jones, T.Y. Hargrove, *The unintended consequences of high-stakes testing*; D.M. Koretz, *Measuring up: what educational testing really tells us*; Smith i Rottenberg 2011.

[19] Tamże.

[20] J.J. Cannell, *Nationally Normed Elementary Achievement Testing in America's Public Schools: How all 50 states are above the national average?* Educational Measurement, 1988, 7(2).

socie, w którym wszyscy mieszkańcy są pod względem jakichś cech powyżej średniej. Osiągnięcie takich rezultatów nie jest jednak możliwe w standaryzowanych testach. Późniejsze analizy potwierdziły istnienie tej powszechnej „nadprzeciętności", ale ustalenie jej przyczyn podzieliło badaczy. Niektórzy uważali, że wyniki zostały celowo zmanipulowane na etapie przeprowadzenia testów (które były przygotowane przez prywatne firmy), albo w momencie podawania ich do wiadomości publicznej. Nieco późniejsze hipotezy[21] kierowały uwagę na praktyki nauczycieli, między innymi: wybieranie spośród dostępnych testów tego, który najlepiej będzie pasował do programu nauczania w danym dystrykcie, albo „trenowanie" uczniów do lepszego zdania testów. Inni wskazywali na zaistnienie artefaktu spowodowanego odnoszeniem wyników uczniów do grupy referencyjnej, której dla danego roku nie stanowili inni uczniowie, ale poprzednie roczniki[22]. Przy względnej przewidywalności zadań i braku instytucjonalnej kontroli nad procedurą ich przeprowadzania, ta ostatnia hipoteza była bardzo prawdopodobna. Wskazywano też na wpływ motywacji uczniów, którzy byli bardziej zmobilizowani pisząc ważny dla nich egzamin zaliczeniowy, niż test normalizujący. Ten ostatni służył do wyznaczenia ogólnonarodowej średniej, ale w odróżnieniu od testu szkolnego nie miał on wysokiej stawki.

Wśród interpretacji efektu *Lake Wobegon* nie pojawiła się ani razu taka, która mówiłaby, że stoją za nim pozytywne zmiany w szkołach. Stało się oczywiste, że wyniki testów informują o czymś więcej niż tylko o umiejętnościach uczniów – zawarta jest w nich również informacja, jak egzaminy testowe można „ograć". Testy zyskały tym samym nowych przeciwników, których sceptyczne nastawienie ugruntowane zostało w końcu naukowymi dowodami. „Wojna o testy" w amerykańskiej oświacie weszła w nową fazę, w której – jak podkreśla Phelps[23] – obawy o manipulacje rozciągnęły się na wszystkie egzaminy testowe, bez względu na konsekwencje, jakie ze sobą niosły oraz standardy, według których były prowadzone.

Testy nie przestały być jednak podstawowym narzędziem oceny wiedzy uczniów. Nic nie wskazywało również, aby podjęto udane próby przeciwdziałania przyczynom ich iluzorycznego wzrostu. Oliwy do ognia dolały, opublikowane kilka lat po ogłoszeniu wyników Cannella, badania zrealizowane pod kierunkiem Davida Koretza[24]. W jednym z amerykańskich dystryktów badacze ci zauważyli, że wyniki testów wysokiej stawki wśród

---

[21] G.W. Phillips, *The Lake Wobegon Effect*, Educational Measurement: Issues and Practice, 1990, 3(9).

[22] J.J. Cannell, *Nationally Normed Elementary Achievement Testing*, s. 5-9.

[23] R.P. Phelps, *Defending standardized testing*, New York 2005.

[24] D.M. Koretz, *Measuring up: what educational testing really tells us*.

uczniów trzecich klas systematycznie rosły, ale tylko do momentu, kiedy w szkole zaczęto stosować inny, niemal identyczny test – wtedy znacząco spadły. Rozkład tych wyników przypominał kształtem odwróconą literę V, albo jak wolą amerykańscy badacze – ząbki piły. To ostatnie porównanie odnosi się nie tylko do kształtu rozkładu wyników, ale również jego cykliczności. Dlaczego uczniowie na nowym teście poradzili sobie gorzej niż na starym? Według zespołu badawczego Koretza było to właśnie wynikiem inflacji. Aby obronić swoją hipotezę, niedługo po przeprowadzeniu tam obowiązującego testu wysokiej stawki, uczniowie w losowo wybranych szkołach napisali inną wersję testu. Okazało się, że wyniki tego ostatniego były niemal o połowę niższe aniżeli wyniki testu przeprowadzonego zaledwie dwa tygodnie wcześniej. Badacze doszli do podobnych wniosków analizując wyniki czwartoklasistów w innym stanie. W nowo wprowadzonym teście uczniowie w ciągu krótkiego czasu imponująco poprawili swoje wyniki, czego nie odnotowano w przeprowadzonym w tym samym okresie teście porównawczym o zbliżonej konstrukcji. W tym drugim średnia liczba punktów nawet nieznacznie spadła[25]. Badacze wyciągnęli z tego jednoznaczne wnioski: uczniowie byli przygotowani do zaliczenia jednego, konkretnego testu, ale ich wiedza nie podlegała generalizacji i kiedy tylko zmieniła się formuła testu, radzili sobie gorzej. Badania zespołu D. Koretza na nowo ożywiły dyskusję na temat wpływu testów na proces nauki, ale tym razem uwaga przesunęła się bardziej w kierunku różnego rodzaju praktyk, które powodują wzrost wyników. Jeśli bowiem to nie wiedza uczniów się zmieniła, to co?

Cud nad testem

Na początku pierwszej dekady XXI wieku uwagę opinii publicznej zwróciły szybko rosnące wyniki uczniów w Teksasie. Wystrzeliły one wówczas, kiedy na początku lat 90. XX wieku wprowadzono tam nowe zasady polityki szkolnej rozliczalności, które stanowiły reakcję na pesymistyczne konkluzje zawarte w raporcie *Nation at Risk*. Podobnie jak w późniejszej reformie NCLB, zasadzały się na przekonaniu, że wprowadzenie standaryzowanych testów pomoże zidentyfikować przyczyny niskiej efektywności szkół oraz umożliwi skierowanie pomocy do najbardziej potrzebujących uczniów. Kluczowym elementem nowej polityki było wprowadzenie standaryzowanego testu TAAS[26], który zarówno dla uczniów, jak i nauczycieli

[25] Tamże.

[26] *Texas Assessment of Academic Skills* został wprowadzony w roku 1990/1991, ale system egzaminów w Teksasie rozwinął się na dobre w 1994 roku. Od tego roku uczniowie zdawali

miał charakter testu wysokiej stawki. Uczniowie musieli go zdać z kilku przedmiotów kilkakrotnie w ciągu nauki w szkole podstawowej i średniej, aby uzyskać dyplom ukończenia szkoły. Wyniki uzyskane przez uczniów stanowiły jednocześnie ważny komponent oceny nauczycieli i dyrektora szkoły. Dobre placówki, spełniające konkretne kryteria ilościowe, mogły spodziewać się finansowej nagrody, a słabsze musiały liczyć się z sankcjami ze strony instytucji nadzorującej. Szkoły nie pozostały pasywne. Po wprowadzeniu zmian postępy uczniów rosły tak spektakularnie, że sytuację tę określono mianem „teksańskiego cudu". W dekadzie lat 90. wskazywały na to cztery ważne wskaźniki[27]: zwiększał się odsetek uczniów zdających test (z 52% w roku 1994 do ponad 70% w roku 1998), zmniejszała się różnica punktów między grupami etnicznymi, coraz mniej uczniów porzucało naukę, a dobre wyniki w TAAS znalazły potwierdzenie także w krajowym teście NAEP. Ten ostatni pokazał, że czwartoklasiści z Teksasu dokonali większego postępu w latach 1992-1996 w matematyce niż ich rówieśnicy z pozostałych stanów. Zaczęto badać przyczyny owego „cudu".

Nie powinno zaskakiwać, że odpowiedzią nauczycieli i uczniów na nowe zasady było bardziej intensywne przygotowywanie do testu. Treningi tego rodzaju były szczególnie nasilone w szkołach z przeciętnie niższymi wynikami, które były najbardziej narażone na sankcje[28]. Jednak uczenie się testów to za mało, aby można było utrzymać wysoki poziom w szkołach o zróżnicowanej kompozycji socjoekonomicznej i społecznej uczniów. Walt Haney[29] przekonywał, że w Teksasie dokonywano selekcji uczniów przystępujących do egzaminu. Mianowicie, po wprowadzeniu reformy liczba latynoskich i afroamerykańskich uczniów powtarzających ostatnią klasę szkoły podstawowej wzrosła, co – ze względu na relatywnie niższe wyniki uczniów należących do tych kategorii – miało zagwarantować uzyskanie korzystniejszych rezultatów na teście w 10 klasie. Ten ostatni był szczególnie ważny, ponieważ służył jako wskaźnik efektywności szkół średnich. Systematycznie powiększała się też liczba uczniów, którzy klasyfikowani byli jako posiadający „specjalne potrzeby", a których wyniki nie były brane pod uwagę w systemie rozliczania szkół. Wykluczanie słabszych uczniów z przystąpie-

---

testy z czytania i matematyki w klasach 3, 4, 5, 6, 7, 8 i 10 oraz dodatkowo z innych przedmiotów w wybranych latach nauki. Test zawierał głównie pytania wielokrotnej odpowiedzi (Klein i Haney 2000).

[27] W. Haney, *The Myth of the Texas Miracle in Education*, Education Policy Analysis Archives, 2000, 8(41).

[28] J.H. Vasquez, L. Darling-Hammond, *Accountability Texas-Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context*, Educational Evaluation and Policy Analysis, 2008, 30(2), s. 75-110.

[29] W. Haney, *The Myth of the Texas Miracle in Education*.

nia do testu miało być również przyczyną wysokiej pozycji Teksasu w krajowym teście NAEP, którego wyniki służą do porównań między stanami. Inne badania wskazują, że słabsi uczniowie byli wręcz nakłaniani do przerwania nauki, aby liczba zdobytych przez nich punktów nie zaniżała szkolnej średniej[30].

Warto odnotować, że nie wszyscy badacze zgadzali się z tymi wnioskami. Richard Phelps[31], jeden z niewielu badaczy, który bierze testy w obronę, dokonał reanalizy danych W. Haneya i wyczytał z nich inną historię – podważył między innymi argument głoszący, że liczba uczniów nie uzyskujących promocji w Teksasie wyróżniała się na tle innych stanów. Nie potwierdził też, aby szkoły w tym stanie wykluczyły szczególnie dużą liczbę uczniów z niższymi wynikami z testu NAEP. Głos ten nie stał się jednak zbyt donośny, ze względu na mnożące się dowody na manipulacje przy TAAS oraz doniesienia z innych stanów. Przykładem tego ostatniego może być badanie przeprowadzone na Florydzie[32] w okresie, kiedy obowiązywał tam test wysokiej stawki dla uczniów szkoły podstawowej i średniej (K12). D.N. Figlio analizował długość okresu zawieszenia w prawach ucznia, które uniemożliwiało uczestnictwo w zajęciach, w tym egzaminach. Najważniejszy wniosek z tego badania nie dotyczył nawet tego, że uczniowie ze słabszymi wynikami byli zawieszani na dłuższy okres niż ci lepsi (za tę samą przewinę), ale tego, że ta luka znacznie powiększała się podczas sesji egzaminacyjnych: uczniowie słabsi zostawali w domach, gdy ich lepiej rokujący rówieśnicy przystępowali do testu.

Wyniki tych i innych badań jasno pokazały, że uruchomienie systemu surowej rozliczalności, opartej jedynie na wynikach testów, stanowi bodziec do manipulacji i nadużyć[33]. Warto zwrócić uwagę, że przyczyną tych negatywnych zjawisk nie było samo stosowanie testów, ale zastosowanie ich wyłącznie do rozliczania szkół. Tworzenie rankingów, karanie, nagradzanie oraz sortowanie uczniów dokonywane było z użyciem tego samego narzędzia, które w założeniach służyło do pomiaru cech ukrytych, jakimi są umie-

---

[30] L. Darling-Hammond, F. Adamson, *Beyond the bubble test*; A.L. Amrein, D.C. Berliner, *High-stakes testing, uncertainty, and student learning*, Education Policy Analysis Archives, 2002, 10(18).

[31] R.P. Phelps, *Kill the messenger: the war on standardized testing*, New York – New Brunswick 2003.

[32] D.N. Figlio, *Testing, Crime, and Punishment*, Journal of Public Economics, 2006, 90(4-5), s. 837-851.

[33] Najgłośniejszy w ostatnich latach ujawniony przypadek oszustwa szkolnego miał miejsce w Atlancie w 2001 roku. Uwagę mediów zwróciły wtedy nieregularności wyników uczniów w niektórych dystryktach miasta. Dochodzenie ujawniło, że w ponad czterdziestu na pięćdziesiąt sześć szkół objętych badaniem nauczyciele i dyrektorzy poprawiali odpowiedzi uczniów na arkuszach egzaminacyjnych.

jętności. Do tego celu, a nie do innego, były one projektowane. Testy zaliczyć można wobec tego do tych „wynalazków" współczesnej nauki, które podobnie jak dynamit, nie zawsze są stosowane zgodnie z pierwotnym przeznaczeniem.

Poza kwestią szkolną, przypadek Teksasu ma jeszcze jeden wymiar. Publikacja wyników na ten temat nastąpiła tuż przed wprowadzeniem reformy NCLB, dla której inspiracją były właśnie rozwiązania rozwijane w Teksasie w latach 1995-2000, kiedy gubernatorem był G.W. Bush. Jest to tym bardziej ciekawe, że podobny „cud" zdarzył się też w Północnej Karolinie, jednak to nie ten stan znalazł się w centrum uwagi badaczy tuż przed wyborami.

Źródła inflacji wyników

Przytoczone wcześniej badania, w których istnienie inflacji stwierdza się poprzez porównanie wyników dwóch testów (w przypadku USA jest to zazwyczaj krajowy NAEP oraz test używany w danym stanie lub dystrykcie) mają kilka wad. Po pierwsze, do testów wysokiej stawki – inaczej niż do testów o mniejszym znaczeniu lub zrównujących – uczniowie przygotowują się dłużej, a podczas ich wypełniania są bardziej zmotywowani do udzielania poprawnej odpowiedzi. Większa mobilizacja na „prawdziwych" egzaminach może w pewnej mierze wyjaśniać różnicę w wynikach uzyskanych w porównywanych testach. Drugi problem odnosi się do konstrukcji testów. Wysoka korelacja między dwoma testami świadczyłaby o niskim prawdopodobieństwie wystąpienia inflacji, ponieważ umiejętności uczniów bez względu na użyte narzędzie są zbliżone. Słaby związek między dwoma testami wcale nie daje jednak pewności, że istnieje inflacja. Dwa testy, nawet jeśli obejmują podobny materiał z programu nauczania i mają zbliżone własności psychometryczne, mogą różnić się w drobnych, ale istotnych szczegółach. Jako przykład posłużyć może analiza wyników z egzaminu gimnazjalnego przeprowadzonego w Polsce w 2013 roku[34]. Egzamin ten przeprowadzany został w dwóch wariantach, różniących się kolejnością poprawnych odpowiedzi w niektórych pytaniach wielokrotnego wyboru. Badacze założyli, że uczniowie wypełniający test mogą sugerować się rozkładem symbolu odpowiedzi w kolejnych pytaniach – np. kiedy w trzech następujących po sobie pytaniach poprawne odpowiedzi to A, A, A mogą potraktować to jako mało prawdopodobne i zmienią odpowiedzi, aby uzyskać bardziej „cha-

[34] M. Koniewski, P. Majkut, P. Skórska, *Zróżnicowane funkcjonowanie zadań testowych ze względu na wersję testu*, Edukacja, 2014, 1(126), s. 79-94.

otyczny" wzór (np. A, B, A). Intuicje te potwierdziły się. W wariancie testu, w którym pojawiała się seria takich samych symboli odpowiedzi uczniowie rzadziej udzielali poprawnej odpowiedzi niż w tym, gdzie była ona zróżnicowana. Ponadto, widząc serię takich samych symboli, zmieniali odpowiedź w pytaniu, które uznawali za najtrudniejsze[35]. Z punktu widzenia zagadnienia inflacji, wyniki tego badania mogą być interpretowane jako symptom ograniczonej zdolności generalizacji wiedzy (jeśli bowiem uczniowie mają dobrze opanowany materiał, powinni odpowiedzi opierać wyłącznie na jego znajomości). Treningi do testów, polegające na nauce wskazywania poprawnej odpowiedzi poprzez eliminację mało prawdopodobnych, są tylko jednym tego przykładem. Inne związane są z różnego rodzaju obciążeniami poznawczymi, które uruchamiają się przy wypełnianiu zadań testowych. Metodologia tworzenia testów dąży do wyeliminowania takich efektów. Wydaje się także, że nie wszyscy uczniowie są jednakowo podatni na pułapki zawarte w testach i jednakowo wyczuleni na ich luki. Może to dotyczyć szczególnie tych uczniów, którzy w sytuacji niepewności bardziej polegają na swojej intuicji.

O ile jednak wpływ konstrukcji testu na wyniki można zwykle stosunkowo precyzyjnie oszacować, trudniej dokonać tego samego w odniesieniu do różnego rodzaju praktyk ułatwiających pozytywne przejście egzaminów[36]. D.M. Koretz[37] wskazuje dwie zasadnicze kategorie występujących praktyk. Do jednej grupy zalicza oczywiste nadużycia, m.in. podpowiadanie uczniom, poprawianie odpowiedzi. Do drugiej należą te metody pracy uczniów i nauczycieli, które są ściśle zorientowane na skuteczność zaliczenia testu. Ich negatywny wpływ polega np. na tym, że niektóre treści programu szkolnego są całkowicie pomijane na rzecz innych, a więc dochodzi do istotnej relokacji czasu i zaangażowania. Wyniki podlegają wtedy inflacji, ponieważ te brakujące elementy uniemożliwiają generalizację wiedzy i sprawiają, że uczeń jest przygotowany jedynie do konkretnego testu.

Adherenci egzaminów testowych bronią jednak prawa do intensywnego przygotowania się do testów argumentując, że warunki wystąpienia realnej inflacji są bardziej złożone. W ich przekonaniu sposób przygotowywania uczniów jest bardziej pochodną cech testów – m.in. poziomu ich wszechstronności i przewidywalności. „Uczenie pod test" występuje wówczas, kiedy doskonalenie umiejętności analitycznych oraz wiedzy staje się mniej opłacalne niż szlifowanie strategii wyboru poprawnej odpowiedzi opartej na

---

[35] Tamże.

[36] T. Haladyna, S. Nolen, N. Haas, *Raising standardized achievement test scores and the origins of test score pollution*, Educational Researcher, 1991, 20(5).

[37] D.M. Koretz, *Measuring up: what educational testing really tells us*.

poprzednich testach. Metody „optymalizacji" procesu edukacyjnego nie zrodziły się zresztą w epoce testów, ani nie są dla niej osobliwe. Najdawniejszym odkrytym dowodem na stosowanie „bryków", na długo przed pojawieniem się nowoczesnych testów, jest anonimowy rękopis studenta Uniwersytetu w Paryżu, datowany na lata 1230-1240. Zawierał on listę najbardziej prawprawdopodobnych pytań zadawanych podczas egzaminu ustnego (Madaus i in., 2009). Doniesienia o ograniczaniu nauczanych zagadnień podczas przygotowania do testu pochodzą też z Chin – kraju, który jako pierwszy wprowadził surowe egzaminy wysokiej stawki dla urzędników na stanowiska państwowe. Około 70 lat po wprowadzeniu tych egzaminów, w roku – bagatela – 681 n.e., szef komisji egzaminacyjnej doniósł cesarzowi, że kandydaci „wkuwają" na pamięć zagadnienia z poprzednich lat, bez zrozumienia treści[38]. Intensywne przygotowanie się do egzaminu, polegające na pomijaniu niektórych zagadnień kosztem innych – jak twierdzą adwokaci egzaminów testowych – nie jest skorelowane z istnieniem testów wielokrotnej odpowiedzi, ale wynika raczej z samej stawki egzaminu.

Z praktyką przygotowywania się do testu wiąże się także inny, ważniejszy być może problem. Jej intensywność oraz metody różnią się między szkołami i nauczycielami[39], uniemożliwiając tym samym rzetelne porównywanie wyników testów. Nie wiadomo, ile „dodatkowych" punktów na teście uzyskuje uczeń ze szkoły elitarnej lub słabszej, uczący się pod okiem doświadczonego lub rozpoczynającego pracę nauczyciela. Nie wiemy też, jak na wyniki testów wpływają pozaszkolne zajęcia, takie jak korepetycje, dodatkowe kursy, czy materiały. One również w nieznany sposób różnicują wyniki uczniów, uzależniając je nie tylko od posiadanej wiedzy, ale przede wszystkim od pochodzenia społecznego. Warto sobie jednak zdawać sprawę z tego, że osiągnięcie idealnej trafności testów, a więc sytuacji, w której osoby o takich samych umiejętnościach mają takie same prawdopodobieństwo odpowiedzi na to samo pytanie, jest niezmiernie trudne. Wyniki testów mierzą, oprócz umiejętności, również jakąś frakcję pochodzenia społecznego ucznia i związanego z tym kapitału kulturowego.

Polityka rozliczalności a efekt inflacji

O kwestii rozliczalności opartej na wynikach testów można myśleć również w kategoriach „prawa Campbella". Donald Campbell (1976) w artykule

[38] H.K. Suen, L. Yu, *Chronic Consequences of High-Stakes Testing? Lessons from the Chinese Civil Service Exam*, Comparative Education Review, 2006, 50(1), s. 46-65.

[39] G.M. Jones, B.D. Jones, T.Y. Hargrove, *The unintended consequences of high-stakes testing*.

poświęconym sposobom oceny zmian społecznych opisał mechanizm inflacji wskaźników. Główną konkluzję płynącą z artykułu można sparafrazować w następujący sposób: im bardziej dany wskaźnik służy za podstawę podejmowania decyzji społecznych, tym bardziej prawdopodobne, że będzie poddany zafałszowaniu i zanieczyści wiedzę o procesie społecznym, który miał monitorować. Jest wiele obszarów, w których można zaobserwować istnienie tego zjawiska. Przykładu dostarcza choćby służba zdrowia, gdzie instytucje rozliczane bywają na podstawie liczby skutecznie wyleczonych pacjentów. Wskaźnik ten można podnieść poprzez stosowanie innowacyjnych metod leczenia, ale również poprzez odmowę leczenia pacjentów w ciężkim stanie. Bardziej drastycznym przykładem działania prawa Campbella była „merytokratyczna" zasada nagradzania amerykańskich żołnierzy podczas wojny w Wietnamie zależnie od liczby zabitych przeciwników. Jak się później okazało, ofiarą tej okrutnej reguły nierzadko stawali się cywile.

W odniesieniu do edukacji, prawo Campbella stosuje się bardzo bezpośrednio: jeśli wynik w teście staje się jedynym wskaźnikiem jakości nauczania, traci wówczas swoją rzetelność i nie wskazuje już tego, co miał wskazywać – podlega zanieczyszczeniu przez stosowanie praktyk skrojonych wyłącznie pod pomiar. Wydaje się, że w politykę rozliczalności opartej na testach wpisany jest pewien paradoks. Polityka ta nastawiona jest na uzyskanie szybkiego wzrostu wyników, ale kiedy faktycznie tak się dzieje, obserwatorzy są zazwyczaj zgodni, że tendencja taka jest najprawdopodobniej iluzoryczna. Trudno bowiem w innych kategoriach wyjaśnić szybki wzrost wyników obserwowany zaraz po wprowadzeniu zmian. Wątpliwy jest zwykle nagły wzrost motywacji uczniów w porównaniu ze starszymi kolegami i koleżankami, szczególnie jeśli dany rocznik na wcześniejszych etapach nauki nie wykazywał podobnych sukcesów. Trudno też – zważywszy na znaną odporność instytucji na interwencje zewnętrzne – spodziewać się, że wyższe wyniki w publicznych, masowych szkołach zawdzięczać można nagłej, radykalnej zmianie nawyków i sposobu pracy nauczycieli[40]. Wydaje się, że zgodnie z tym co sugeruje J. Lee (2010), w systemie surowej rozliczalności szkoły mają do wyboru dwie strategie: albo stosując różne sposoby „ograniа systemu" osiągnąć dobre wyniki na konkretnym teście wysokiej stawki, albo działać w granicach etyki zawodowej i zadowolić się stabilnymi wynikami, które jedynie w długiej perspektywie mogłyby ulec stopniowej, nieznacznej poprawie, często niezauważalnej wobec wzrostu średniej międzyszkolnej

Sceptycyzm w ocenie zasad rozliczalności opartej na testach wysokiej stawki pojawił się w USA już w erze pierwszej generacji takich systemów, to

[40] D.B. Tyack, L. Cuban, *Tinkering toward utopia.*

jest w latach 80.[41], a po wdrożeniu NCLB jeszcze się umocnił. Zgoda w ich ocenie nigdy nie była jednak pełna. W jednym z pierwszych badań na ten temat Audrey Amrein i David Berliner[42] porównali wyniki testu NAEP z wynikami testów w stanach, które posługiwały się rygorystyczną polityką rozliczalności. Do jej elementów należały nagrody finansowe dla szkół lub nauczycieli, stypendia dla wyróżniających się uczniów, a z drugiej strony – możliwość zamknięcia szkoły czy wymiany jej personelu. Z przeprowadzonych przez nich analiz nie wyłaniała się wyraźna tendencja – w okresie objętym badaniem wyniki niektórych kategorii uczniów w części stanów wzrosły, a w innych spadły[43]. W studium tym nie skorzystano jednak w pełni z możliwości porównania stanów posiadających system silnej i słabej rozliczalności, jaką daje swoboda wyboru ustroju szkolnego w USA. Wziął to natomiast pod uwagę Barak Rosenshine[44]. Badacz ten odróżnił stany, w których testy miały „bez wątpienia" wysoką stawkę, od wszystkich pozostałych i nie potwierdził wcześniejszych ustaleń A. Amrein i D. Berlinera[45]. Wskazywał bowiem, że stany stosujące większy reżim sankcji i nagród podniosły swoje wyniki w porównaniu z pozostałymi, choć ten przyrost nie wszędzie był jednakowo duży[46].

Do podobnych wniosków doszli również Martin Carnoy i Susanna Loeb (2002). Ich analiza dotyczyła wyników w testach z matematyki przeprowadzonych na krajowej próbce uczniów w latach 1996-2000. M. Carnoy i S. Loeb utworzyli sześciopunktowy indeks oparty na analizie polityki rozliczalności w poszczególnych stanach. Okazało się, że jej ostrzejsza wersja jest pozytywnie skorelowana z kilkupunktowym wzrostem w teście NAEP w ósmym roku nauki. Wynik ten utrzymał się również przy kontroli struktury populacji uczniów, która z testu została wykluczona, co w przypadku NAEP ma na ogół znaczenie dla ogólnych rezultatów[47].

Powstały jednak poważne wątpliwości, jak interpretować ten wzrost w kontekście opisanego wcześniej problemu inflacji. Szybka zmiana trendu wyników – czy to w górę albo jak się rzadziej zdarza w dół – może sygnalizować skuteczniejsze opanowanie praktyk „ograniа testów", co w systemie

---

[41] J. Lee, T. Reeves, *Revisiting the Impact of NCLB High-Stakes School Accountability, Capacity, and Resources: State NAEP 1990-2009 Reading and Math Achievement Gaps and Trends*, Educational Evaluation and Policy Analysis, 2012, 34, s. 209-231.

[42] A. Amrein, D. Berliner, *High-stakes testing, uncertainty, and student learning*.

[43] Tamże.

[44] B. Rosenshine, *High-stakes testing: Another analysis*, Education Policy Analysis Archives, 2003, 11(24).

[45] A. Amrein, D. Berliner, *High-stakes testing, uncertainty, and student learning*.

[46] B. Rosenshine, *High-stakes testing: Another analysis*.

[47] Tamże.

„kija i marchewki" jest szczególnie prawdopodobne. Dobrze ilustrują ten problem badania z Chicago[48], gdzie w 1997 roku wprowadzono program testów wysokiej stawki Iowa Test of Basic Skills (ITBS). Porażka na teście w trzecim, szóstym lub ósmym roku nauki wiązała się dla ucznia z nieuzyskaniem promocji do kolejnej klasy, a w ostatniej klasie uniemożliwiała ukończenie szkoły średniej. Uczniowie mieli możliwość ponownego pisania testu po odbyciu dodatkowych, letnich zajęć, ale w systemie rozliczalności szkoły brany pod uwagę był tylko wynik z pierwszej sesji. Jeśli określona liczba uczniów (więcej niż 15%) uzyskała wyniki poniżej krajowej normy na teście z czytania, a szkoła nie podjęła działań naprawczych, nauczycielom i administracji groziło przeniesienie do innych placówek lub zwolnienie. Po wprowadzeniu tych zasad, wyniki uczniów w testach z matematyki i czytania wzrosły bardziej niż wynikałoby to z wcześniejszych trendów. Dzieci z Chicago cieszyły się też wyższymi wynikami, niż ich rówieśnicy z innych dystryktów miejskich tego stanu, w których nie stosowano podobnej polityki. Równolegle odnotowano jednak także wzrost wyników względem innego testu administrowanego na poziomie stanu (wypełnianego przez te same dzieci). Można było zatem podejrzewać, że te ponadprzeciętne rezultaty spowodowane były specyficznym przygotowaniem do testu. Aby to potwierdzić, badacze założyli, że największego wzrostu wyników można oczekiwać w przypadku zagadnień mało złożonych albo takich, które pojawiają się w teście ITBS częściej niż w teście porównawczym. Ich przypuszczenia potwierdziły się – w ciągu krótkiego czasu uczniowie bardziej poprawili wyniki w łatwych pytaniach matematycznych, niż w złożonych. Innymi słowy, metodą na osiągnięcie sukcesu było położenie przez nauczycieli nacisku na uczenie umiejętności bardziej fundamentalnych, dające im bardziej pewne „zyski" w postaci liczby punktów. Jak twierdzi B.A. Jacob[49], reforma ta negatywnie wpłynęła na umiejętność generalizacji wiedzy, a więc możliwość jej zastosowania w kontekstach poza tym konkretnym testem. Wydaje się jednak, że wypada zachować pewną ostrożność w jej ogólnej ocenie, skoro kierunkiem „optymalizacji" procesu nauczania kierowała zawartość testu, czyniąc zagadnienia podstawowe bardziej opłacalne.

Również wyniki innych badań z zakresu wpływu polityki rozliczalności na wyniki testów nie prowadzą do jednoznacznych konkluzji. Obraz zamazuje się szczególnie wówczas, kiedy analizy prowadzone są w rozbiciu na

[48] B.A. Jacob, *Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools*, Journal of Public Economics, 2005, 89, s. 761-796.

[49] Tamże.

poszczególne kategorie uczniów, przedmioty i lata nauki[50]. Po części może to wynikać z szeregu różnic międzyszkolnych, które nie jest łatwo skontrolować w modelach statystycznych. Sama podatność szkół na zmianę, ich zdolność do organizacyjnej adaptacji stanowi tu ważną zmienną, która nie musi być przecież bezpośrednio związana z czystą jakością kształcenia.

Czy ktoś został w tyle?

Inflację wyników można uznać za główny, niezamierzony skutek uboczny testów wysokiej stawki, który za sprawą NCLB uległ prawdopodobnie wzmocnieniu. Głównym zamierzonym celem tej reformy było natomiast zmniejszenie luki między uczniami z różnych grup etnicznych i o odmiennym statusie materialnym. Jednak w odróżnieniu od rozbieżnych opinii na temat wpływu nowego modelu rozliczalności na efektywność szkół, badania dotyczące zmniejszenia nierówności są jednoznaczne – do poważnych zmian nie doszło[51]. Różnica między wynikami uczniów z różnych grup etnicznych zaczęła wolno maleć na długo przed wprowadzeniem tej reformy, a zmiany rozpoczęte w roku 2002 ani nie osłabiły, ani nie nasiliły tego trendu. Nie inaczej kształtowała się luka między dziećmi z ubogich i zamożnych rodzin. Różnica w osiągnięciach edukacyjnych między tymi grupami amerykańskich uczniów znacznie się zwiększyła w ostatnich dwóch dekadach, ale nic nie wskazuje na to, aby ten wzrost miał coś wspólnego z reformą[52].

Nieskuteczność reform szkolnych w redukcji nierówności nie jest odkryciem nowym – zróżnicowanie szans kształtuje się dużo wcześniej, niż opanowana zostanie umiejętność płynnego czytania i liczenia[53]. Szkoły repro-

---

[50] J. Lee, *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An indepth look into national and state reading and math outcome trends*, Harvard 2006; tegoż, *Is Test-Driven External Accountability Effective? Synthesizing the Evidence from Cross-State Causal-Comparative and Correlational Studies*, Review of Educational Research, 2008, 78(3), s. 608-644; X. Wei, *Are More Stringent NCLB State Accountability Systems Associated With Better Student Outcomes? An Analysis of NAEP Results Across States*, Educational Policy, 2012, 26, s. 268-308, T.S. Dee, B. Jacob, *The impact of No Child Left Behind on student achievement*, Journal of Policy Analysis and Management, 2011, 30(3), s. 418-446.

[51] S.F. Reardon, *The widening academic – achievement gap between the rich and the poor: New evidence and possible explanations*, [w:] *Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low Income Children*, red. R.M. Murnane, G. Duncan, New York 2011.

[52] S.F. Reardon, K. Bischoff, *Income Inequality and Income Segregation*, American, Journal of Sociology, 2011, 116(4), s. 1092-1153.

[53] G.J. Duncan, K. Magnuson, *The Nature and Impact of Early Achievement Skills, Attention Skills, and Behavior Problems*, [w:] *Whither opportunity? Rising inequality, schools, and children's life chances*, red. G.J. Duncan, R.J. Murnane, New York 2011, s. 47-70.

dukują te podziały, robią to jedynie z różnym stopniem skuteczności. W USA do czynników, które potęgują różnice szans należą szczególnie dramatyczne nierówności dochodowe[54]. Nie sposób zagwarantować dzieciom równego startu w szkole wobec systematycznej koncentracji dochodu osiągającej na początku XXI pułap z okresu Wielkiej Depresji[55]. Nie sprzyja temu również segregacja przestrzenna, coraz silniej skorelowana z dochodem, która – biorąc pod uwagę rejonizację szkół – ujednolica strukturę etniczną i materialną uczniów[56]. Ważny jest jeszcze jeden czynnik. Coraz większe znaczenie dla powiększania luki w wynikach szkolnych mają inwestycje rodziców w edukację pozaszkolną. W USA, gdzie oferta takich zajęć jest szczególnie duża, różnica w tym zakresie wzrosła w ostatnich dekadach dramatycznie: od lat 70. zamożne rodziny powiększyły swoje inwestycje o 150%, podczas gdy rodziny z niskim dochodem – tylko o jedną trzecią[57].

Brak wpływu na poziom nierówności można także wyjaśnić biorąc pod uwagę priorytety NCLB. W pierwszej kolejności zmiany celowały w pomiar umiejętności uczniów, a dopiero w drugiej w ich poprawę. Takie podejście, choć krytykowane, nie jest całkowicie pozbawione sensu. Jego zwolennicy posługują się metaforą: „najpierw trzeba zebrać informacje o pacjencie, aby można było rozpocząć odpowiednią kurację". Przeciwnicy częstego testowania odpowiadają im w tej samej konwencji: „jeśli pacjent jest chory, mierzenie temperatury co chwilę nie pomoże mu wyzdrowieć". Rację ma zapewne jedna i druga strona, ponieważ nawet najbardziej drobiazgowe informacje o osiągnięciach uczniów zdadzą się na nic, jeśli nie będą odpowiednio wykorzystane. Z drugiej strony, planowanie polityk edukacyjnych bez wiedzy o jej stanie oraz zróżnicowaniu ma małe szanse powodzenia. Być może zwolennikom i przeciwnikom stosowania testów warto w tym kontekście przypomnieć zasadę Paracelsusa: „to dawka sprawia, że lekarstwo nie jest trucizną". Ostatecznie wypada też zauważyć, że kwestia małej efektywności NCLB w zmniejszaniu nierówności i niezamierzone efekty systemu testów wysokiej stawki są od siebie w znacznej mierze niezależne. Doświadczenia innych krajów, które próbowały osiągnąć większą równość odmiennymi metodami (jak choćby Francji, która w latach 80. objęła opieką uczniów z dzielnic ubóstwa) również prowadzą do mało optymistycznych konkluzji.

---

[54] S.F. Reardon i in., *Left behind? The effect of No Child Left Behind on academic achievement gaps*, Center for Education Policy Analysis, Stanford CA 2013; S.F. Reardon, K. Bischoff, *Income Inequality and Income Segregation*; S.F. Reardon, *The widening academic – achievement gap between the rich and the poor*.

[55] E. Saez, G. Zucman, *Wealth Inequality in the United States since 1913: Evidence From Capitalized Income Data*, Cambridge 2014.

[56] S.F. Reardon, K. Bischoff, *Income Inequality and Income Segregation*.

[57] G.J. Duncan, R.J. Murnane, *Whither opportunity?*

Wnioski

W ostatnich dekadach świat instytucji publicznych przeszedł swoistą rewolucję, szeroko adaptując paradygmat określany jako *New Public Management*. Zmiany, jakich doświadczył, polegają na przyjęciu zasad proefektywnościowych, które dotychczas wykorzystywane były głównie w gospodarce rynkowej. Jedną z najważniejszych jest nacisk na rozliczalność, rozumianą jako dążenie do osiągnięcia zadanych celów oraz wykazanie tego za pomocą konkretnych wskaźników. W amerykańskiej edukacji takim wskaźnikiem stały się wyniki egzaminów testowych. W Europie, gdzie zwrot w kierunku stosowania testów ma znacznie krótszą historię (większość krajów zaczęła wprowadzać systemy egzaminów krajowych dopiero od połowy lat 90., a trend ten przyspieszył na dobre dopiero w roku 2000), testy stosowane są na razie w innym celu. Zazwyczaj wpływają tylko na przebieg karier szkolnych uczniów, mają charakter diagnostyczny lub informacyjny.

Casus amerykański może wydawać się na tym tle odległy, ale właśnie poprzez dokonanie porównań z przypadkiem tak odmiennym można sformułować kilka przewidywań dotyczących zjawisk, które na naszym gruncie dopiero kiełkują. Warto się im przyglądać choćby po to, aby podobne skutki, oprócz statusu niezamierzonych, nie uzyskały w Polsce charakteru nieprzewidzianych. Szkoły w okresie reform stają się zawsze swego rodzaju poligonem, na którym w sposób doświadczalny sprawdza się metody rozwiązywania wielu problemów naraz, takich jak niska efektywność kształcenia, nieobiektywność ocen, czy duża nierówność szans. Wprowadzenie w Polsce testowych egzaminów zewnętrznych było jednym z takich „naturalnych eksperymentów". Analiza wyników amerykańskich badań skłania do przypuszczenia, że ryzykiem, jakie niosą za sobą testy wysokiej stawki, jest pojawienie się inflacji wyników. Można jednak pokusić się o hipotezę, że ryzyko to jest tym większe, im poważniejsze konsekwencje zostaną tym testom przypisane. Podwyższony reżim rozliczalności wywiera presję na krótkookresowy wzrost wyników, który przy zachowaniu innych czynników na tym samym poziomie, może być jedynie iluzoryczny. Doświadczenia amerykańskie wyraźnie pokazują, że surowa rozliczalność przyczynia się do rozmywania granicy między etycznymi a nieetycznymi praktykami szkolnymi, tym samym potrafi „psuć morale" kadry pedagogicznej. Testy mogą się wówczas stać – paradoksalnie – bezużyteczne, bowiem przestają mierzyć to co mierzyć powinny, a więc umiejętności uczniów. Jest to najważniejsza sprzeczność edukacyjnych polityk rozliczalności i towarzyszących im systemów testowych: gdy wyniki rosną, pojawiają się także wątpliwości doty-

czące realnych przyczyn ich osiągnięcia. Sukcesy osiągane krótkookresowo zawsze witane są z entuzjazmem, przede wszystkim przez zwolenników reformy, jednak realnej oceny skutków dokonać można dopiero w relatywnie długim okresie – w Stanach Zjednoczonych była to perspektywa dziesięcioleci. Również w Polsce duże znaczenie dla oceny dotychczas wprowadzonych reform będą miały kolejne edycje pomiarów edukacyjnych – począwszy od badań PISA, a na maturze skończywszy.

W Polsce, jak na razie, wyniki testów nie służą podejmowaniu kluczowych decyzji dotyczących zarządzania szkołami i odwrotnie niż w NCLB – mają bardziej realne konsekwencje dla uczniów, niż dla nauczycieli. Wydaje się więc, że przynajmniej obecnie ryzyko inflacji jest względnie małe. Kolejne lata pokażą, czy ewentualne zmiany w zakresie polityki rozliczalności to zmienią.

BIBLIOGRAFIA

Amrein A.L., Berliner D.C., *High-stakes testing, uncertainty, and student learning*, Education Policy Analysis Archives, 2002, 10(18).

Cannell J.J., *Nationally Normed Elementary Achievement Testing in America's Public Schools: How all 50 states are above the national average?* Educational Measurement, 1988, 7(2).

Carnoy M., Loeb S.D., *External Accountability Affect Student Outcomes? A cross-state analysis*, Education and Evaluation and Policy Analysis, 2002, 24(4).

Choińska-Mika J. i in., *Realizacja podstawy programowej z historii w gimnazjach*, Instytut Badań Edukacyjnych, Warszawa 2013.

Crocker L., *Teaching for the test: How and why test preparation in appropriate*, [w]: *Defending standardized testing*, red. R. Phelps, L. Erlbaum Associates, New York 2005.

Darling-Hammond L., *From "Separate but equal" to "No child left behind": the collision of new standards and old inequalities*, [w:] *Many children left behind: how the No Child Left Behind Act is damaging our children and our schools*, red. D. Meier, G.H. Wood, Beacon Press, Boston 2004.

Darling-Hammond L., Adamson F., *Beyond the bubble test: how performance assessments support 21st century learning*, Jossey-Bass, San Francisco 2014.

Dee T.S., Jacob B., *The impact of No Child Left Behind on student achievement*, Journal of Policy Analysis and Management, 2011, 30(3).

Dolata R., *Cicha rewolucja w oświacie – proces różnicowania się gimnazjów w dużych miastach*, Edukacja, Studia, Badania, Innowacje, 2010, 1(105).

Duncan G.J., Magnuson K., *The Nature and Impact of Early Achievement Skills, Attention Skills, and Behavior Problems*, [w:] *Whither opportunity? Rising inequality, schools, and children's life chances*, red. G.J. Duncan, R.J. Murnane, Russell Sage Foundation, New York 2011.

Duncan G.J., Murnane R.J., *Whither opportunity? Rising inequality, schools, and children's life chances*, Russell Sage Foundation, New York 2011.

Eurydice Network, *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*, 2009.

Figlio D.N., *Testing, Crime, and Punishment*, Journal of Public Economics, 2006, 90(4-5).
Gallagher C.J., *Reconciling a Tradition of Testing with a New*, Educational Psychology Review, 2003, 15(1).
Goodman D., Hambleton R.K., *Some misconceptions about large-scale educational assessments*, [w:] *Defending standardized testing*, red. R.P. Phelps, Mahwah, Lawrence Erlbaum Associates, New York 2005.
Haladyna T., Nolen S., Haas N., *Raising standardized achievement test scores and the origins of test score pollution*, Educational Researcher, 1991, 20(5).
Haney W., *The Myth of the Texas Miracle in Education*, Education Policy Analysis Archives, 2000, 8(41).
Hess F.M., Petrilli M.J., *No Child Left Behind primer*, Peter Lang, New York 2006.
Holcombe R., Jennings J., Koretz D., *The roots of score inflation: An examination of opportunities in two states' tests*, [w:] *Charting reform, achieving equity in a diverse nation*, red. G. Sunderman, CT: Information Age Publishing, Greenwich 2013.
Horn R.A., *Understanding educational reform: a reference handbook*, ABC-CLIO, California, Santa Barbara 2002.
Jacob B.A., *Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools*, Journal of Public Economics, 2005, 89.
Jones G.M., Jones B.D., Hargrove T.Y., *The unintended consequences of high-stakes testing*, Lanham, Md.: Rowman & Littlefield Pubishers, Oxford 2003.
Klein S.P., Hamilton L.S., McCaffrey D.F., Stecher B.M., *What Do Test Scores in Texas Tell Us?* CA: RAND Corporation, Santa Monica 2000.
Konarzewski K., *Przygotowanie uczniów do egzaminu: pokusa łatwego zysku*, ISP, Warszawa 2008.
Koniewski M., Majkut P., Skórska P., *Zróżnicowane funkcjonowanie zadań testowych ze względu na wersję testu*, Edukacja, 2014, 1(126).
Koretz D.M., *Measuring up: what educational testing really tells us*, Mass. Harvard University Press, Cambridge 2008.
Lazarin M., *Testing Overload in America's Schools*. Center for American Progress 2014. https://cdn.americanprogress.org/wp-content/uploads/2014/10/LazarinOvertestingReport.pdf [dostęp: 27.02.2016].
Lee J., *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An indepth look into national and state reading and math outcome trends*, The Civil Rights Project at Harvard University, Harvard 2006.
Lee J., *Is Test-Driven External Accountability Effective? Synthesizing the Evidence from Cross-State Causal-Comparative and Correlational Studies*, Review of Educational Research, 2008, 78(3).
Lee J., Reeves T., *Revisiting the Impact of NCLB High-Stakes School Accountability, Capacity, and Resources: State NAEP 1990-2009 Reading and Math Achievement Gaps and Trends*, Educational Evaluation and Policy Analysis, 2012, 34.
Madaus G.F., Russell M.K., Higgins J., *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. IAP. 2009.
Meyer H., Rowan B., *The new institutionalism in education*, State University of New York Press, Albany 2006.
Phelps R.P., *Kill the messenger: the war on standardized testing*, Transaction Publishers, New York – New Brunswick 2003.
Phelps R.P., *Defending standardized testing*, Mahwah, L. Erlbaum Associates, New York 2005.

Philips G.W., *The Lake Wobegon Effect*, Educational Measurement: Issues and Practice, 1990, 3(9).

Pokropek A., *Matura z języka polskiego. Wybrane problemy psychometryczne*, XVII Konferencja Diagnostyki Edukacyjnej, Kraków 2011.

Ravitch D., *The Death and Life of the Great American School System: How Testing and Choice are Undermining Education*, Basic Books, New York 2010.

Reardon S.F., *The widening academic – achievement gap between the rich and the poor: New evidence and possible explanations*, [w:] *Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low Income Children*, red. R.M. Murnane, G. Duncan, Russell Sage Foundation, New York 2011.

Reardon S.F., Bischoff K., *Income Inequality and Income Segregation*, American, Journal of Sociology, 2011, 116(4).

Reardon S.F., Greenberg E.H., Kalogrides D., Shores K.A., Valentino R.A., *Left behind? The effect of No Child Left Behind on academic achievement gaps*, Center for Education Policy Analysis, Stanford CA 2013.

Rosenshine B., *High-stakes testing: Another analysis*, Education Policy Analysis Archives, 2003, 11(24).

Saez E., Zucman G., *Wealth Inequality in the United States since 1913: Evidence From Capitalized Income Data*, Working paper 20625, National Bureau of Economic Research, Cambridge 2014.

Suen H.K., Yu L., *Chronic Consequences of High-Stakes Testing? Lessons from the Chinese Civil Service Exam*, Comparative Education Review, 2006, 50(1).

Tyack D.B., Cuban L., *Tinkering toward utopia: a century of public school reform*, Mass. Harvard University Press, Cambridge 1995.

Thomas M.H., Bobbit Nolen N.S., Haas N.S., *Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution*, Educational Researcher, 1991, 20(5).

Vasquez J.H., Darling-Hammond L., *Accountability Texas-Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context*, Educational Evaluation and Policy Analysis, 2008, 30(2).

Wei X., *Are More Stringent NCLB State Accountability Systems Associated With Better Student Outcomes? An Analysis of NAEP Results Across States*, Educational Policy, 2012, 26.