Edukacja 2012, 1 (117), 23-43
ISSN 0239-6858

Szacowanie wpływu liczebności klasy na osiągnięcia edukacyjne uczniów z wykorzystaniem eksperymentu ex post facto

MACIEJ KONIEWSKI*

W analizach wpływu liczebności klasy na osiągnięcia edukacyjne wykorzystano dane z badań zrealizowanych w 2006 r. przez Okręgową Komisję Egzaminacyjną w Krakowie. Zmienne wyjaśniające wyniki gimnazjalistów zidentyfikowano za pomocą analizy regresji. Model wyjaśnia 71% wariancji wyników egzaminu. Zmienne te wykorzystano w procedurach wyłonienia statystycznych bliźniąt. Ich przydział do grupy eksperymentalnej i kontrolnej przeprowadzono na trzy sposoby: przez warstwowanie z wykorzystaniem odległości Mahalanobisa, łączenie „jeden do wielu" oraz „jeden do jednego" metodą *k*-średnich; ten ostatni okazał się najskuteczniejszy. Wpływ liczebności klasy na osiągnięcia edukacyjne badanych uczniów był nieistotny statystycznie. Jednak uczniowie z klas poniżej 23 osób osiągali na egzaminie gimnazjalnym średnio o 0,039 odchylenia standardowego lepsze wyniki niż ich rówieśnicy w większych klasach.

Efektywna polityka edukacyjna państwa, której zadaniem jest racjonalna alokacja dostępnych zasobów w celu ciągłego podnoszenia jakości kształcenia, powinna z namysłem korzystać z wyników badań naukowych. Rolą badań edukacyjnych jest natomiast dostarczanie wiedzy na temat siły i kierunku zależności między jakością nauczania a innymi czynnikami, szczególnie tymi na które mamy wpływ, poprzez decyzje administracyjne i finansowanie. Jednym z tych czynników jest liczebność klasy. Ten problem rzadko jest poruszany w polskiej debacie publicznej, od czasu do czasu wychodzi jednak z cienia ważkich tematów społecznych i politycznych.

Liczebnością klasy, jako wielkością mogącą podlegać optymalizacji, zainteresowani są zarówno rodzice, jak i nauczyciele, dyrektorzy i organy prowadzące szkoły. Rodzice i nauczyciele cenią sobie klasy małoliczne. Pierwsi – ponieważ wierzą, że w mniejszych klasach dzieci uczą się bardziej efektywnie. Drudzy – ponieważ w małolicznych klasach pracuje się bardziej komfortowo. Dyrektorzy i organy prowadzące szkoły (samorządy) zainteresowani są raczej utrzymaniem bardziej licznych klas, z uwagi na oszczędności, gdyż pensje nauczycielskie są głównym składnikiem wydatków na edukację.

Wiedza o charakterze zależności między liczebnością klasy a osiągnięciami edukacyj-

Artykuł napisany na podstawie pracy magisterskiej przygotowanej pod kierunkiem dra hab. Jarosława Górniaka w Instytucie Socjologii Uniwersytetu Jagiellońskiego w Krakowie. Adres do korespondencji: Maciej Koniewski, Zakład Socjologii Gospodarki, Edukacji i Metod Badań Społecznych, Instytut Socjologii UJ, ul. Grodzka 52, 31-044 Kraków. Adres e-mail: maciej.koniewski@uj.edu.pl

* Instytut Socjologii, Uniwersytet Jagielloński w Krakowie

nymi uczniów jest istotna w procesie decydowania o liczebności klas – zagadnieniu efektywnego wydatkowania środków publicznych i podnoszenia jakości edukacji. Dyskusja na temat efektu liczebności klasy wydaje się daleka od ostatecznego i uniwersalnego rozwiązania, ponieważ jest to zagadnienie niezwykle złożone i kontekstowe. Wiele zmiennych, potencjalnie mogących wpływać na osiągnięcia edukacyjne uczniów, leży wciąż poza zasięgiem narzędzi pomiarowych, jak i pomysłowości badaczy. Niewątpliwie jednak problem optymalnej liczebności klasy pozostanie trwałym elementem polityki edukacyjnej państwa i jako taki powinien być ciągle pogłębiany.

Wyniki dotychczasowych badań nad efektem małej klasy

Efekt liczebności klasy na osiągnięcia edukacyjne jest przedmiotem wielu badań od początku XX w. Pierwsze badanie na ten temat przeprowadził Joseph Mayer Rice (1902). Wskazać można na dwa wiodące podejścia do tego problemu. Pierwsze to badania eksperymentalne. Pozwalają one z dużą precyzją uchwycić, czy na zmianę wyników poszczególnych uczniów miała wpływ liczebność klasy, czy inne uwarunkowania. Wadą tej metody jest jej duże uzależnienie od kontekstu badania oraz stosunkowo mała liczba osób poddanych obserwacji. Odmienną tradycją w studiach nad efektem liczebności klasy są analizy ekonometryczne. Korzysta się w nich z danych na temat rzeczywistej liczebności klas, częściej jednak jest to stosunek liczby uczniów do nauczycieli w szkole i modeluje związki między liczebnością klas a wielkością przyrostu osiągnięć edukacyjnych (mierzoną za pomocą ogólnokrajowych testów wiedzy i kompetencji). Analizy takie często prowadzone są na danych populacyjnych, jednak zakres możliwości kontrolowania czynników kontekstowych jest w nich ograniczony.

Wśród niektórych autorów istnieje zgoda co do pozytywnego wpływu małych klas na podnoszenie wyników nauczania: „Mimo że rezultaty zarówno randomizowanych badań eksperymentalnych, jak i analiz ekonometrycznych zgodnie wskazują na pozytywny efekt małych klas, niektórzy badacze uważają te dowody za niejednoznaczne" (Nye i in., 2000, s. 124). Inni, jak na przykład Eric Hanushek (1999, 2002) czy Allan Odden (1990) twierdzą, że redukcja liczebności klas jest przedsięwzięciem niewspółmiernie kosztownym w stosunku do uzyskiwanych rezultatów. Kontrowersje wśród badaczy budzi siła tego efektu[1].

Najbardziej wyczerpująca metaanaliza badań nad efektem liczebności klasy to praca Glassa i Smitha (1978), opisana także skrótowo w artykule tychże autorów z 1979 r. Zakwalifikowali oni do swojej analizy 77 badań przeprowadzonych na przestrzeni 70 lat. W sumie we wszystkich włączonych do analizy badaniach wzięło udział 900 000 uczniów. Główny wniosek ich pracy donosi o występowaniu pozytywnego wpływu małych klas, liczących poniżej 23 uczniów, na wyniki nauczania. Wpływ ten pozostaje niezależny od nauczanego przedmiotu, poziomu IQ uczniów oraz podstawowych cech demograficznych. Autorzy ustalili, że nauczanie indywidualne jest o 0,565 odchylenia standardowego miary osiągnięć edukacyjnych skuteczniejsze niż nauczanie w klasach 40-osobowych. Ujemny związek między liczebnością klasy a osiągnięciami edukacyjnymi jest silniejszy w badaniach, w których uczniowie zostali przypisani do

[1] Wyniki dotychczasowych badań nad efektem liczebności klasy zebrane zostały w metaanalizach i przeglądach systematycznych (Bridle i Beliner, 2004; Educational Research Service 1980; Glass i in., 1982; Glass i Smith 1978; Graue i in., 2005; Hedges i Stock, 1983; Molnar, Smith i Zahori, 2000; Nye, Hedges i Konstantopoulos, 2001; Pillmer i Light, 1980; Robinson, 1990; Robinson i Wittebols, 1986; Slavin, 1986).

różnej wielkości klas w sposób losowy, niż w badaniach, w których proces doboru nie był losowy. W badaniach prowadzonych przed 1940 r. nie odnotowywano związku liczebności klas z osiągnięciami uczniów, natomiast silny związek tych dwóch zmiennych odnotowywano w badaniach prowadzonych od lat 60. Fakt ten można tłumaczyć rozwojem bardziej zaawansowanych i dokładnych metod pomiaru, jak i wypracowaniem złożonych schematów eksperymentalnych.

Ciekawa jest także praca Glena Robinsona i Jamesa Wittebolsa (1986). Autorzy zastosowali analizę skupień do klasyfikacji badań na temat efektów liczebności klas, realizowanych w latach 1950–1985. Spośród 124 prac ujętych w metaanalizie Robinsona i Wittebolsa – 54 (44%) faworyzowały mniejsze klasy, w 60 pracach (48%) nie ustalono kierunku związku liczebności klasy z osiągnięciami edukacyjnymi, a 10 kolejnych prac (8%) faworyzowało duże klasy. Najbardziej widoczny, ujemny związek między liczebnością klasy a wynikami uczniów, zaobserwowano wśród ośmio- i dziewięciolatków. Pozytywne efekty małej klasy dotyczyły przede wszystkim umiejętności czytania i liczenia. Efekty były wyraźne w klasach 22-osobowych i mniejszych. Jednakże, jak wspominają autorzy, pozytywne efekty małych klas często nie są stabilne w czasie. Robinson w zwięzłej formie prezentuje także inne wnioski z tych analiz w artykule z 1990 r. Według niego redukcja liczebności klas ma stosunkowo niewielki pozytywny efekt na wyniki uczniów w porównaniu do innych (mniej kosztownych) interwencji lub strategii mających na celu podniesienie poziomu nauczania. Liczebność klasy jako czynnik samodzielny ma niewielki wpływ na wyniki uczniów niezależnie od nauczanego przedmiotu, zwłaszcza w klasach 23–30-osobowych. W klasach o zredukowanej liczebności, której potencjału nauczyciel nie wykorzystuje odpowiednio dostosowując program i metody nauczania, należy spodziewać się braku lub niewielkiej poprawy wyników uczniów.

Opracowania obejmujące nowsze badania nad efektami liczebności klasy (Graue i in., 2005; Molnar, Smith i Zahori, 2000) dowodzą pozytywnego efektu redukcji liczebności klas na wyniki uczniów. Dodatkowych dowodów na poparcie tezy o pozytywnym wpływie małej klasy na osiągnięcia edukacyjne uczniów – szczególnie wśród uczniów z mniejszości etnicznych i narodowych oraz z grup o niskim statusie społecznym – dostarcza badanie zespołu Barbary Nye (2001). Największy wkład w ustalenie kierunku i siły związku liczebności klas z osiągnięciami edukacyjnymi uczniów mają jednak badania z ostatnich czterech dekad XX w., głównie eksperymenty prowadzone na wielką skalę w Stanach Zjednoczonych. Na podstawie wyników tych badań podjęto decyzje o redukcji liczebności klas w poszczególnych stanach.

Bruce Biddle i David Berliner (2004) podsumowują rezultaty eksperymentalnych badań edukacyjnych prowadzonych na wielką skalę w latach 60.–90. w Stanach Zjednoczonych. Najważniejsze wnioski dowodzą, że dobrze zaplanowane i adekwatnie dotowane programy redukcji liczebności klas na wczesnych etapach kształcenia dają zauważalne korzyści w postaci lepszych wyników w nauce. Są one tym większe i bardziej stabilne, im dłużej uczeń uczęszczał do małej klasy. Pozytywny efekt małej klasy jest wyraźny w klasach szkół podstawowych liczących mniej niż 20 uczniów, niezależnie od płci ucznia i nauczanego przedmiotu, a także od stosowanej miary wiedzy i umiejętności. Beneficjentami małych klas są przede wszystkim uczniowie z rodzin biednych oraz członkowie mniejszości narodowych i etnicznych. Uczniowie małych klas na wczesnych etapach nauczania utrzymują wysokie wyniki także w bardziej licznych klasach w kolejnych etapach kształcenia.

Powyżej przytoczone zostały wyniki kluczowych metaanaliz badań eksperymentalnych nad wpływem liczebności klasy na osiągnięcia edukacyjne uczniów. Drugim najczęściej stosowanym podejściem do pomiaru wpływu liczebności klasy na osiągnięcia edukacyjne są analizy ekonometryczne. Niniejszy artykuł skupia się głównie na badaniach eksperymentalnych, jednak warto przedstawić, choćby krótkie podsumowanie wyników analiz ekonometrycznych. Eric Hanushek (1998) zebrał 90 publikacji spełniających odpowiednio wysokie kryteria merytoryczne i metodologiczne, które zawierały 377 osobnych oszacowań funkcji produkcyjnej szkół. Autor pogrupował zebrany materiał według dodatniego lub ujemnego wyniku estymacji związku ilorazu liczby uczniów do liczby nauczycieli z wynikami uczniów. Trzynaście procent wszystkich estymacji wykazało dodatni i statystycznie istotny związek ilorazu uczniów i nauczycieli z wynikami uczniów, tzn. wykazały, że im więcej nauczycieli przypada na jednego ucznia, tym uczniowie osiągają lepsze rezultaty. Ujemny i statystycznie istotny efekt wykazało 15% wszystkich analiz (Hanushek, 1998).

Niewiele prac polskich autorów podejmuje tematykę efektu liczebności klasy. Warto wspomnieć artykuł Macieja Jakubowskiego i Pawła Sakowskiego (2006). Autorzy przedstawili rezultaty analiz wpływu liczebności klasy na wyniki szkolne uczniów. Zastosowane metody umożliwiły wyabstrahowanie wpływu liczebności klasy w oparciu o analizę danych zastanych, zawierających zmienne charakteryzujące szkoły oraz wyniki egzaminów uczniów szkół podstawowych w województwie mazowieckim z lat 2002–2004. Problem, któremu należy sprostać w tego typu analizach, dotyczy endogeniczności. Pojawia się on wtedy, gdy jedna lub więcej cech badanych uczniów jednocześnie determinuje przypisanie do warunków eksperymentalnych – do małej lub dużej klasy – oraz wpływa na zmienną zależną – wyniki szkolne (Strawiński, 2007). Jakubowski i Sakowski radzą sobie z endogenicznością na dwa sposoby. Pierwszy polega na zastosowaniu średniej liczebności klasy w danym roczniku w danej szkole jako zmiennej instrumentalnej dla faktycznej liczebności klasy. Dodatkowo autorzy kontrolowali różnice między szkołami. Drugim sposobem było zastosowanie pomysłu doboru jednostek do analizy opartego na tzw. regule Maimonidesa (Agrist i Lavy, 1999). Analizowano tylko szkoły, które tworzyły nowe klasy, gdy liczba uczniów w roczniku przekraczała około 29 uczniów lub wielokrotność 29 uczniów. Średnia liczebność klas także tu posłużyła jako zmienna instrumentalna. Uzyskane wyniki, w większości przypadków istotne statystycznie, wskazują na niewielki, choć pozytywny efekt małych klas na osiągnięcia edukacyjne uczniów. Utrzymywanie stosunkowo małolicznych klas jest korzystne szczególnie w szkołach na terenach wiejskich. Inne polskie prace poruszające tematykę liczebności klas to analizy Przemysława Śleszyńskiego (2002) oraz Mikołaja Herbsta i Jana Herczyńskiego (2005). Wykorzystane przez tych autorów metody statystyczne nie pozwalają jednak wyciągać wiążących wniosków na temat efektu liczebności klas.

Prawomocność orzekania o wpływie przyczynowym

Powodem częstego zakłopotania, pojawiającego się przy raportowaniu i interpretacji wyników badań naukowych, jest mieszanie lub mylenie korelacji z relacją przyczynową. Można na przykład zaobserwować związek między codziennym jedzeniem śniadania a wynikami szkolnymi uczniów. Te dwa zjawiska są ze sobą skorelowane, co oznacza, że wystąpienie jednego zjawiska często wiąże się z wystąpieniem drugiego. W ustaleniu tej relacji pomaga na przykład analiza regre-

sji. Współwystępowanie dwóch zjawisk nie musi świadczyć o ich zależności przyczynowo-skutkowej. Mogą na nie wpływać inne zjawiska, które lepiej wyjaśniają zmianę poziomu wyników szkolnych.

Dzieci, które nie jedzą regularnie śniadań, mogą pochodzić z biedniejszych rodzin lub częściej opuszczać lekcje, co z kolei warunkuje ich gorsze wyniki. Relacja między codziennym jedzeniem śniadań a wynikami szkolnymi jest relacją pozorną a związek między tymi zjawiskami wyjaśniają inne zmienne pośredniczące. Twierdzenie, że codzienne jedzenie śniadania poprawia wyniki szkolne uczniów, wymaga zweryfikowania takiej hipotezy empirycznie za pomocą metod badawczych, które gwarantują wysoką trafność wewnętrzną otrzymanych wyników.

Wyniki prezentowane w tym artykule pretendują do zasilenia hipotezy o wpływie przyczynowym liczebności klasy na osiągnięcia edukacyjne uczniów. Orzekanie o wpływie przyczynowym jest prawomocne, tylko jeżeli spełnione są trzy podstawowe wymogi dla wszystkich zależności przyczynowo-skutkowych: przyczyna poprzedza skutek, przyczyna współzmienia się (*covary*) ze skutkiem, oraz alternatywne wyjaśnienia relacji przyczynowo-skutkowej są niemożliwe. Wymogom tym sprosta badanie przeprowadzone z wykorzystaniem metody eksperymentu, uznawanej za „złoty standard" poznania naukowego. W eksperymencie badacz manipuluje bodźcem, aby wymusić jego pojawienie się przed efektem. Współzmienność między przyczyną a skutkiem łatwo sprawdzić w analizie statystycznej. Chcąc spełnić trzecie wymaganie wykorzystuje się randomizowane eksperymenty, które sprawiają, że alternatywne wyjaśnienia nie są możliwe. Zakłada się, że są one losowo rozłożone między warunkami eksperymentalnymi, czyli między porównywanymi grupami.

Podstawowa logika badania eksperymentalnego polega na porównywaniu wartości zmiennej wynikowej u osób, które wystawione były na oddziaływanie bodźca, z wartością zmiennej u tych osób, które nie doświadczyły jego oddziaływania. W idealnych warunkach poziom zmiennej wynikowej powinien zostać zmierzony u osoby, która jednocześnie doświadczyła i nie doświadczyła bodźca. Oczywiście nie jest to możliwe. Problem orzekania o wpływie przyczynowym jest więc problemem wynikającym z braków danych (Heckman, Ichimura i Todd, 1997). Badanym, którzy doświadczyli bodźca (grupa eksperymentalna), przyporządkowuje się osoby, które go nie doświadczyły (grupa kontrolna) i w tej grupie mierzony jest poziom zmiennej wynikowej. Zabieg ten nazywany jest wywołaniem stanu kontrfaktycznego. Innymi słowy, chodzi o to, aby jednostki w grupie poddanej oddziaływaniu bodźca i w grupie kontrolnej były możliwie jak najbardziej do siebie podobne.

W eksperymentach randomizowanych, zwanych także prawdziwymi (*true experiments*), efekt podobnego składu obu grup uzyskuje się poprzez losowy dobór osób do obu grup. Losowy dobór jednostek do porównywanych grup nazywany jest randomizacją. Zastosowanie tej metody daje pewność, że zmienna wynikowa jest niezależna zarówno od obserwowanych, jak i nieobserwowanych czynników, innych niż bodziec, które mogłyby na nią wpływać, ponieważ rozkłady tych zmiennych są losowo rozdystrybuowane między porównywanymi grupami.

W sytuacji gdy randomizacja nie jest możliwa ze względów finansowych, etycznych, technicznych lub gdy operujemy na danych zastanych, sposobem na wykluczenie wpływu czynników obserwowalnych (alternatywnych wyjaśnień związku przyczynowego będącego przedmiotem zainteresowania) jest przeprowadzenie

statystycznego warstwowania (*stratifying*) lub dopasowania (*matching*) danych po badaniu. Wykorzystanie tych technik statystycznych umożliwia dopasowanie grupy eksperymentalnej i kontrolnej do siebie pod kątem zmiennych, które zarówno korelują ze zmienną zależną, jak i wpływają na selekcję osób do warunków, w których oddziałuje bodziec (grupa eksperymentalna). Dzięki temu można zbliżyć się do idealnej sytuacji, w której jednostki analizy są losowo przypisane do porównywanych grup (wszystkie potencjalne czynniki, inne niż bodziec, wpływające na poziom zmiany zmiennej zależnej są losowo rozdystrybuowane między porównywanymi grupami). Po przeprowadzeniu procedury dopasowania uzyskuje się grupy podobne do siebie pod względem cech, mogących stanowić potencjalne źródło obciążenia pomiaru zmiennej zależnej. W ten sposób doprowadza się do „wyzerowania" wpływu alternatywnych czynników na zmienną zależną (tych czynników, które na mocy teorii zostaną zidentyfikowane jako korelujące ze zmienną zależną i faktem doświadczenia bodźca przez jednostkę analizy).

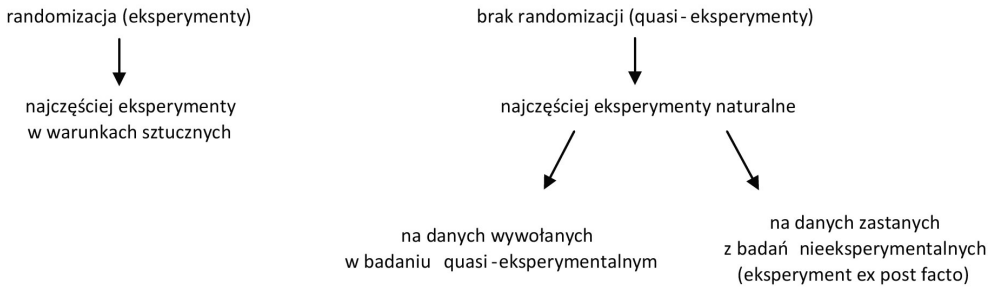
Podstawowe typy badań odwołujące się do logiki eksperymentu

Powyżej przytoczone zostały różne metody badawcze odwołujące się do logiki badań eksperymentalnych. Wymagają one nazwania i zdefiniowania. Wspomniano eksperymenty randomizowane, zwane także prawdziwymi. Najczęściej prowadzone są one w warunkach sztucznych (laboratoria). David Freedman, Robert Pisani i Roger Purves (1997) wskazali trzy atrybuty charakteryzujące randomizowane eksperymenty. Po pierwsze, reakcja grupy eksperymentalnej na bodziec jest porównywana z reakcją grupy kontrolnej na warunki kontrolowane, czyli brak obecności bodźca. Po drugie, przypisanie jednostek do grup eksperymentalnych jest losowe. Po trzecie, manipulowanie bodźcem jest kontrolowane przez badacza. Te trzy kryteria odgrywają kluczową rolę w eksperymentalnym modelu przyczynowości.

W sytuacji gdy nie przeprowadza się randomizacji, należy mówić o quasi-eksperymentach. Niewątpliwie największy wkład w popularyzowanie samego pojęcia quasi-eksperymentu, jak i schematów quasi-eksperymentalnych miał Donald Campbell. Jak słusznie zauważa Thad Dunning (2008, s. 289), Campbell pod pojęciem quasi-eksperymentu, rozumiał „przybliżenie wzorca prawdziwego eksperymentu", czyli porównanie reakcji jednostek w warunkach ekspozycji na bodziec oraz warunkach braku ekspozycji na bodziec.

Mimo że w sytuacji quasi-eksperymentu nie mamy do czynienia z losowym przydziałem jednostek analizy do porównywanych grup, to w przeciwieństwie do innych nieeksperymentalnych metod badacz może pod pewnymi warunkami twierdzić, że przypisanie jednostek do warunków obecności bodźca i kontrolnych jest „takie jak" losowe („*as if" random*) (Dunning, 2008). Uprawomocnienie takiego twierdzenia może mieć umocowanie zarówno w argumentacji *a priori*, jak i w dowodach empirycznych. Te drugie pozwalają kontrolować czynniki potencjalnie wpływające na zmienną zależną, które są obserwowalne. Wykluczenie możliwego wpływu na zmienną zależną czynników nieobserwowalnych nie jest możliwe w quasi-eksperymentach. Możliwa jest natomiast ich kontrola pośrednia oparta na wiedzy płynącej z silnych założeń teoretycznych. Podsumowując – główną i czasami jedyną różnicą quasi-eksperymentów w stosunku do „prawdziwych" eksperymentów jest nielosowe przypisanie jednostek do porównywanych grup.

Eksperymentalne badania edukacyjne charakteryzuje głównie prowadzenie ich w natu-



Rysunek 1. Podstawowa typologia metod eksperymentalnych i pochodnych.

ralnym środowisku badanych. W odróżnieniu od eksperymentów prowadzonych w warunkach sztucznych (laboratorium), eksperymenty naturalne to badania prowadzone w środowisku badanych lub wykorzystujące dane pochodzące z obserwacji naturalnie występującego zjawiska. Jako że w takich warunkach badacz nie jest w stanie manipulować bodźcem, eksperymenty naturalne są tak naprawdę badaniami obserwacyjnymi (Dunning, 2008). W sytuacji, która wymaga wykorzystania danych zastanych pochodzących z obserwacji naturalnie występującego zjawiska, stosuje się szczególny rodzaj badania quasi-eksperymentalnego – eksperyment *ex post facto*. Służy on przekształceniu danych w takie, które spełniają wymagania danych kwalifikowanych do analiz eksperymentalnych. Nazwę „eksperyment ex post facto" zaproponował Francis Stuart Chapin (1946) do opisania badania polegającego na przekształceniu danych nieeksperymentalnych, na przykład z badań przekrojowych lub wzdłużnych, w dane eksperymentalne.

Metodologia szacowania efektu małej klasy

Celem tego opracowania jest ocena wpływu liczebności klasy na osiągnięcia edukacyjne uczniów z wykorzystaniem metody eksperymentu *ex post facto*. Wyższość badań odwołujących się do logiki eksperymentu, a takim jest zastosowane tu podejście, nad badaniami korelacyjnymi polega na możliwości orzekania o związku przyczynowym. Innymi słowy, na podstawie takich badań można mówić o efekcie czy też wpływie, jaki konkretna zmienna niezależna (w tym przypadku: liczebność klasy) ma na zmienną zależną (wyniki szkolne uczniów).

Ramy analiz, których celem jest oszacowanie efektu przyczynowego wyznacza model przyczynowy Rubina (*Rubin Causal Model* – RCM). Można go zobrazować w następujący sposób: efekt przyczynowy dla ucznia (i) w klasie małej (T) versus klasie dużej (C) dla zmiennej wynikowej Y wynosi $E_i = Y_i(T) - Y_i(C)$. Objęcie programem (Zi) nie określa wartości wyniku oczekiwanego (przewidywanego) pary Yi(T), Yi(C), ale będzie determinować, który z nich może być zaobserwowany. Wynik Yi(T) może być obserwowany tylko wtedy, gdy uczeń jest w małej klasie (grupa eksperymentalna); wynik Yi(C) może być obserwowany tylko wtedy, gdy uczeń jest w dużej klasie (grupa kontrolna). Średni efekt przyczynowy szacujemy poprzez uczestnictwo uczniów w programie $E = Y(T) - Y(C)$.

Losowe przypisanie do grupy eksperymentalnej implikuje, że średni wynik posttestu w grupie eksperymentalnej yT jest trafnym i nieobciążonym oszacowaniem Y(T), a średni wynik

i	Zi	Yi(T)	Yi(C)
1	T	*	?
2	T	*	?
3	T	*	?
4	C	?	*
5	C	?	*
6	C	?	*

* oznacza dane empiryczne (obserwowalne)
? oznacza brak danych

Rysunek 2. Fragment bazy danych z badania eksperymentalnego. Dane dla sześciu uczniów.

posttesu w grupie kontrolnej yC jest trafnym i nieobciążonym oszacowaniem Y(C). Dodatkowo różnica między średnimi w grupach: yY – yC jest trafnym i niobciążonym szacunkiem średniego efektu przyczynowego (E).

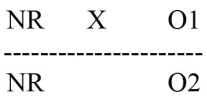
W szacowaniu średniego efektu oddziaływania bodźca przyjmuje się założenie o niezależności zmiennej wynikowej od mechanizmów przypisania jednostek do warunków eksperymentalnych. *The Stable Unit Treatment Value Assumption* (SUTVA) jest założeniem *a priori*, mówiącym, że wartość zmiennej wynikowej Y dla ucznia (i) wystawionego na oddziaływanie bodźca i(T) będzie stała, bez względu na mechanizm przypisania ucznia (i) do warunków T, a także bez względu na to, na oddziaływanie jakich bodźców wystawieni są inni uczniowie (Morgan i Winship, 2007, s. 37). Średni efekt oddziaływania bodźca na wszystkie jednostki w próbie nazywany jest w literaturze ATE (*Average Treatment Effect*). Jego odpowiednikiem dla jednostek w grupie eksperymentalnej jest średni efekt oddziaływania bodźca na jednostki poddane oddziaływaniu – ATT (*Average Treatment for the Treated*), z kolei dla osób w grupie porównawczej średni efekt oddziaływania bodźca na jednostki niepoddane oddziaływaniu – ATC (*Average Treatment Effect for the Controls*).

Celem oszacowania wpływu liczebności klasy na osiągnięcia edukacyjne uczniów została przeprowadzona analiza danych z badań zrealizowanych przez Okręgową Komisję Egzaminacyjną (OKE) w Krakowie tuż po egzaminie gimnazjalnym w maju 2006 r. wśród uczniów gimnazjów. Próba do badania OKE została dobrana losowo, przy użyciu schematu losowania warstwowego. Badanie zrealizowano w 28 szkołach, w 83 oddziałach klas trzecich, w których przeprowadzano ankiety audytoryjne. Łącznie zebrano 1757 pełnowartościowych ankiet, w 1733 przypadkach ich wyniki udało połączyć się z wynikami ankietowanych z egzaminu gimnazjalnego.

Dane z badań przekrojowych zostały przetransformowane w dane eksperymentalne. Metodologia przygotowania danych, jak i samej analizy, odpowiada realizacji eksperymentu *ex post facto*, zgodnie z logiką schematu quasi-eksperymentalnego z pomiarem końcowym (posttestem) i grupami nieekwiwalentnymi (grupą eksperymentalną i jedną grupą kontrolną). Schemat ten przedstawiony został na Rysunku 2.

Potencjalne zagrożenia dla trafności wyniku

Zastosowany schemat eksperymentu *ex post facto* niesie ze sobą co najmniej dwa zagrożenia dla trafności wewnętrznej. Pierwszym jest brak losowego przypisania jednostek analizy do warunków eksperymental-



Gdzie:
X – bodziec
O1, O2 – posttest
linia (--------) oznacza, że grupy nie były utworzone losowo
NR (*nonrandom assignment*)

Rysunek 3. Eksperyment z posttestem i grupami nieekwiwalentnymi.

nych (grupy wystawionej na oddziaływanie bodźca – w tym wypadku klasy małolicznej, oraz grupy na którą nie oddziaływał bodziec – w tym wypadku klasy wielolicznej). Wiąże się z tym problem występowania alternatywnych zmiennych wyjaśniających zmianę zmiennej wynikowej. Identyfikacja tych zmiennych powinna odbywać się z odwołaniem do teorii. Jednak w przypadku eksperymentu *ex post facto*, istnieją spore ograniczenia związane ze zmiennymi występującymi w bazie danych. Tak czy inaczej, im więcej zmiennych, które podzielają wariancję ze zmienną zależną uda się zidentyfikować i kontrolować, tym trafniejsze rezultaty uda się osiągnąć.

Nakreślony tu schemat eksperymentu *ex post facto* nie pozwala w pełni zadośćuczynić założeniu o niezależności jednostek obserwacji od mechanizmów selekcji do warunków eksperymentalnych (SUTVA). Kontrolowane są tylko jawne i zmierzone w badaniu OKE czynniki selekcyjne, które jednocześnie korelują z poziomem osiągnięć edukacyjnych. Czynnikami takimi są np.: płeć i miejsce zamieszkania. Poza kontrolą znajdują się pozostałe czynniki ukryte czy też jawne, ale niezmierzone w badaniu OKE. Wprowadzenie quasi-rynkowych mechanizmów w systemie finansowania polskiej edukacji rozluźniło sztywne zasady rejonizacji. Rodzice, opiekunowie, dzieci mogą sami wybierać szkołę. Za selekcję uczniów odpowiadają także dyrektorzy i nauczyciele. Brak kontroli tych czynników, każe z dozą ostrożności interpretować prezentowane tu dane.

Drugim potencjalnym źródłem obciążenia prezentowanych tu wyników jest brak pomiaru pierwotnego (pretestu) zmiennej zależnej (w tym przypadku zmienną zależną jest wynik ucznia z egzaminu gimnazjalnego). Pomiar pierwotny pozwala uzyskać wiedzę o różnicach „na wejściu" wśród uczestników eksperymentu. Redukcja obciążenia wywołanego brakiem randomizacji, jak i brakiem pretestu możliwa jest dzięki przeprowadzeniu dopasowania statystycznego jednostek w grupie eksperymentalnej i kontrolnej. Brak pretestu zostanie zrekompensowany włączeniem do zmiennych, które będą uwzględnione w procedurze dopasowania, zmiennej niosącej informację o wcześniejszych osiągnięciach edukacyjnych uczniów (oceny z siedmiu przedmiotów na pierwszy semestr trzeciej klasy gimnazjum).

Oprócz możliwych źródeł obciążania związanych z samym schematem eksperymentalnym, także jakość danych z badania OKE wymaga co najmniej trzech uwag krytycznych. Po pierwsze, nie było to badanie ogólnopolskie, obejmowało jedynie województwa: małopolskie, lubelskie i podkarpackie. Druga uwaga dotyczy doboru próby, który został przeprowadzony według schematu z 2004 roku. Zastosowany schemat nie uwzględniał zmiany w sieci szkół, jaka na-

stąpiła do roku 2006, w którym to roku zrealizowano badanie. Schemat był losowy tylko na poziomie szkół. Następnie, w wylosowanej do badania szkole przeprowadzano ankietę audytoryjną wśród wszystkich trzecioklasistów obecnych w szkole w dniu badania. Z tym wiąże się kolejna uwaga. Nie wiemy nic o tym, ilu uczniów było nieobecnych w szkole w dniu badania. Liczebność klas nie była zmienną w bazie. Została wyliczona na podstawie liczby uczniów z danej klasy, którzy wzięli udział w badaniu, czyli łącznie 1757 uczniów. Stąd też liczebności klas przyjęte w analizach w niektórych przypadkach mogą być zaniżone o liczbę uczniów, którzy w dniu badania nie byli w szkole.

Identyfikacja kowariantów osiągnięć edukacyjnych

W planowaniu eksperymentów bez losowego przypisania jednostek do porównywanych grup szczególnie ważna jest identyfikacja i kontrola potencjalnych źródeł wariancji zmiennej wynikowej, innych niż wpływ bodźca. W eksperymentach *ex post facto* liczba alternatywnych wyjaśnień zmiennej zależnej determinowana jest liczbą dostępnych w bazie danych zmiennych. Jest to spore ograniczenie, zwłaszcza gdy badacz nie ma dostępu do zmiennych, które na gruncie teorii należałoby uznać za potencjalne zmienne wyjaśniające wariancję zmiennej zależnej.

Na podstawie dorobku badań nad edukacją należy wskazać następujące czynniki, które potencjalnie mogą mieć wpływ na wyniki szkolne uczniów:

- indywidualne (np. cechy genetyczne, samoocena, aspiracje edukacyjne, motywacje, zainteresowania, dotychczasowe osiągnięcia szkolne, czas przeznaczany na naukę, inteligencja, stan zdrowia);
- środowiskowe (sytuacja rodzinna oraz otoczenie koleżeńskie, np.: wykształcenie rodziców, status społeczno-ekonomiczny rodziny, model rodziny, liczba rodzeństwa, miejsce zamieszkania, warunki pracy domowej, stosunek rodziców do nauki, aspiracje rodziców, współdziałanie rodziców ze szkołą, środowisko rówieśnicze, osiągnięcia szkolne rówieśników, ich kapitał kulturowy, ekonomiczny i społeczny, cechy kultury szkoły determinowane jej składem społecznym);
- instytucjonalne i pedagogiczne (np. model i program szkoły, liczba uczniów w klasie, zasoby materialne szkoły, rozkład zajęć, organizacja lekcji i pracy domowej, wykształcenie i doświadczenie nauczycieli, współpraca między nauczycielami, metody nauczania i sprawdzania osiągnięć, doskonalenie zawodowe nauczycieli, stosunek nauczycieli do uczniów, podręczniki i programy nauczania, organizacja zajęć pozalekcyjnych).

Nie ma zgody co do tego, które zmienne determinują sukcesy uczniów w sposób jednoznaczny. Za najsilniej wpływające na wyniki uczniów zmienne uznaje się: status społeczny, który wiąże się z miejscem zamieszkania i środowiskiem rodzinnym ucznia; poziom wewnętrznej motywacji; aspiracje znaczących innych i wpływ grupy rówieśniczej. Wpływ środowiska rodzinnego i rówieśniczego na wyniki uczniów wskazują najdobitniej badania Jamesa Colemana (1966) oraz Erica Hanushka (1992, 1997). Nowsze badania wskazują jednak na możliwość przeszacowania wpływu rodziny (środowiska wspólnego), kosztem wpływu czynnika genetycznego (Byrne i in., 2010; Harris, 2000; Hart, Petrill, Kamp Dush, 2010).

W wyniku badań przeprowadzonych przez Okręgową Komisję Edukacyjną w Krakowie w 2006 r. wśród uczniów ostatniej klasy gimnazjów zgromadzono szereg danych, które umożliwiają ocenę statusu społecznego ucznia, jego motywacji do nauki, aspira-

Tabela 1
Statystyki podsumowania modelu regresji

Blok	*R*	*R*-kwadrat	Skorygowane *R*-kwadrat	Standardowy błąd oszacowania
1	0,843	0,710	0,708	8,800
2	0,844	0,713	0,710	8,768

Zmiana *R*-kwadrat	*F*	*df*1	*df*2	Istotność
0,710	274,209	14	1567	0,000
0,003	3,916	4	1563	0,004

cji rodziców i rówieśników, a także uprzednich osiągnięć edukacyjnych. Po przeglądzie dostępnych zmiennych, w celu identyfikacji alternatywnych wyjaśnień zmienności wyników egzaminu przystąpiono do specyfikacji modelu regresji, w którym zmienną zależną był wynik egzaminu gimnazjalnego. Uczeń łącznie z części humanistycznej i matematyczno-przyrodniczej egzaminu mógł otrzymać od 0 do 100 punktów. W badanej próbie minimalna wartość zmiennej wynosi 9, a maksymalna 99. Mediana wynosi 55 punktów, a średnia 55,9. Rozkład jest lekko prawoskośny (0,025). Z ponad stu pytań zawartych w kwestionariuszu oraz ich różnych kombinacji czynnikowych do modelu regresji wybrano ostatecznie 15 zmiennych niezależnych. Zaskakiwać może, że ze względu na nieistotną zmianę współczynnika determinacji, do modelu nie włączono takich zmiennych jak: status społeczny, wykształcenie rodziców, aspiracje znaczących innych, aspiracje grupy rówieśniczej. Zmienne te, uważane przez badaczy za ważne determinanty osiągnięć edukacyjnych, tu okazały się niezwiązane z wynikami uczniów.

Model wyjaśnia 71% wariancji wyników egzaminu gimnazjalnego. Mówi o tym wartość skorygowanego współczynnika determinacji. Standardowy błąd oszacowania wskazuje wielkość błędu przewidywania wyników egzaminu gimnazjalnego. Jest on niespełna dwukrotnie mniejszy niż odchylenie standardowe zmiennej zależnej. Oznacza to, że przewidywanie wyników egzaminu dla konkretnego ucznia na podstawie modelu regresji jest dwukrotnie skuteczniejsze, niż w oparciu o średni wynik egzaminu wszystkich uczniów.

Zmienne zostały wprowadzone do modelu w dwóch blokach. W pierwszym wprowadzone zostały wszystkie zmienne mierzone na poziomach ilościowych, w tym pytania z kafeterią typu Likerta. W drugim bloku wprowadzone zostały zmienne reprezentujące kategorie zmiennej porządkowej: „Rodzice sprawdzali odrobione lekcje". Są to tak zwane zmienne pomocnicze (*dummy variables*), kodowane zerojedynkowo. Zmienna reprezentująca kategorię *nigdy*, posłużyła jako kategoria odniesienia. W tabeli zaprezentowano wartości współczynników regresji dla poszczególnych zmiennych niezależnych, wartości testu *t*, korelacje oraz statystyki współliniowości.

Współczynnik β został wyliczony na podstawie zmiennych standaryzowanych, tym samym nie jest zależny od jednostek, w których wyrażone są konkretne zmienne. Umożliwia on porównywanie siły związku poszczególnych zmiennych ze zmienną wyjaśnianą. Najsilniejszymi predyktorami w modelu są uprzednie osiągnięcia edukacyjne uczniów oraz spodziewana liczba punktów z egzaminu.

Współczynnik korelacji semicząstkowej po podniesieniu do kwadratu informuje o tym,

Tabela 2
Wyniki analizy regresji

		B	Błąd st.	Beta	*t*	Istotność	Semi-cząstkowa	Tolerancja	VIF
Czynniki indywidualne	Stała	9,741	1,790		5,44	0,000			
	Uprzednie osiągnięcia edukacyjne[1]	1,375	0,055	0,541	24,98	0,000	0,338	0,391	2,558
	Spodziewana liczba punktów z egzaminu[2]	0,344	0,021	0,297	16,34	0,000	0,222	0,556	1,798
	Twoje sposoby zdobywania wiedzy podczas lekcji w szkole (1=Zdecydowanie nie … 5=Zdecydowanie tak)	1,102	0,250	0,063	4,42	0,000	0,060	0,890	1,124
	Mam zwyczaj pytać nauczycieli o wszystko, co jest niezrozumiałe (1=Zdecydowanie nie … 5=Zdecydowanie tak)	–1,012	0,203	–0,073	–4,94	0,000	–0,068	0,849	1,178
	Ważne jest dla mnie tylko to, żeby przejść do następnej (klasy) szkoły (1=Zdecydowanie nie … 5=Zdecydowanie tak)	–0,959	0,160	–0,093	–6,01	0,000	–0,081	0,767	1,305
	Płeć (0=Chłopiec/1=Dziewczynka)	–1,995	0,488	–0,061	–4,09	0,000	–0,055	0,817	1,224
	Miejsce zamieszkania (0=Miasto/1=Wieś)	–1,419	0,471	–0,044	–3,01	0,003	–0,041	0,877	1,141
	Dysleksja (0=Nie/1=Tak)	3,607	0,838	0,060	4,30	0,000	0,058	0,950	1,053
	Czy w gimnazjum brałeś(aś) udział w zawodach sportowych? (0=Nie/1=Tak)	–1,750	0,455	–0,054	–3,84	0,000	–0,052	0,937	1,067
	Motywacja	–3,676	1,495	–0,047	–2,46	0,014	–0,033	0,511	1,957
Czynniki środowiskowe	Uczniom w mojej klasie zależy, aby na lekcjach jak najwięcej się nauczyć	–0,549	0,209	–0,040	–2,63	0,009	–0,036	0,801	1,248
	Rodzice sprawdzali odrobione lekcje kilka razy w roku	0,277	0,487	0,016	0,57	0,570	0,008	0,239	4,186
	Rodzice sprawdzali odrobione lekcje kilka razy w miesiącu	1,493	0,519	0,084	2,88	0,004	0,039	0,216	4,632
	Rodzice sprawdzali odrobione lekcje przynajmniej raz w tygodniu	–0,562	0,513	–0,032	–1,10	0,274	–0,015	0,220	4,550
	Rodzice sprawdzali odrobione lekcje kilka razy w tygodniu	–1,607	0,558	–0,090	–2,88	0,004	–0,039	0,190	5,266
Czynniki instytucjonalne i pedagogiczne	Sposób prowadzenia lekcji zachęca mnie do aktywności	–1,190	0,242	–0,079	–4,92	0,000	–0,067	0,715	1,398
	Oceń stopień trudności zadań rozwiązywanych w klasie w stosunku do zadań na egzaminie[3]	5,157	0,872	0,083	5,91	0,000	0,080	0,930	1,075
	Czy na lekcjach w szkole rozwiązywałaś(łeś) zadania z języka polskiego i matematyki podobne do zadań na egzaminie?[4]	2,562	0,761	0,048	3,37	0,001	0,046	0,898	1,113

[1] Suma stopni szkolnych z języka polskiego, historii, matematyki, biologii, chemii, fizyki, geografii, otrzymanych na pierwszy semestr klasy trzeciej w gimnazjum.
[2] Suma odpowiedzi na pytania o spodziewaną liczbę punktów z części humanistycznej i matematyczno-przyrodniczej.
[3] 1 = Zadania rozwiązywane w szkole były łatwiejsze od tych na egzaminie; 2 = Zadania rozwiązywane w szkole były podobne pod względem trudności do tych na egzaminie; 3 = Zadania rozwiązywane szkole były trudniejsze od tych na egzaminie; 4 = Trudno powiedzieć – nie potrafię tego ocenić. To czynnik utworzony na podstawie odpowiedzi na pytania o stopień trudności zadań rozwiązywanych w klasie z chemii (ładunek czynnikowy równy 0,759), fizyki (0,759), matematyki (0,734), biologii (0,695), języka polskiego (0,676), geografii (0,674), historii (0,642). Czynniki wyodrębniony został metodą głównych składowych dla zmiennych kategorialnych (Categorical Principal Component Analysis – CATPCA). Obserwacje z brakami danych były wyłączane parami. Pierwszy wymiar wyjaśnia 0,7592 + 0,7592 + 0,7342 + 0,6952 + 0,6762 + 0,6742 + 0,6422=50% początkowej wariancji całkowitej. Alfa Cronbacha wynosi 0,831. Zmienna została sprowadzona do zakresu 0–100 (normalizacja min-max).
[4] 1 = Nie; 2 = Tak – od czasu do czasu; 3 = Tak – dość często; 4 = Trudno powiedzieć. Czynnik utworzony na podstawie odpowiedzi na pytania dotyczące języka polskiego (ładunek czynnikowy równy 0,824) i matematyki (0,838). Czynnik wyodrębniony został metodą głównych składowych dla zmiennych kategorialnych (CATPCA). Obserwacje z brakami danych były wyłączane parami. Wymiar wyjaśnia 0,8242 + 0,8382 = 69% wariancji całkowitej tych dwóch zmiennych. Zmienna została sprowadzona do zakresu 0–100 (normalizacja min–max).

jaka część całkowitej wariancji zmiennej zależnej jest sprowadzalna do wyłącznego wpływu danej zmiennej niezależnej. Najwięcej, bo aż 11% ($0{,}338^2$) wariancji wyników egzaminu gimnazjalnego wyjaśniają uprzednie osiągnięcia edukacyjne uczniów. Spodziewana liczba punktów z egzaminu wyjaśnia 5% ($0{,}222^2$) wariancji wyników egzaminu.

Poszukiwanie statystycznych bliźniąt

Dobór przypadków do grupy eksperymentalnej i kontrolnej przeprowadzono kilkoma metodami, aby móc porównać jakość uzyskanych dopasowań. Przeprowadzono dopasowanie z wykorzystaniem odległości Mahalanobisa. Pełni ona rolę syntetycznej miary, która jest nośnikiem informacji o charakterystykach obiektów. Wyraża odległość obserwacji od centroidu, który jest punktem równowagi w wielowymiarowej przestrzeni wyznaczonej przez zmienne niezależne, uwzględnione w modelu regresji. Zdecydowano się zastosować tę miarę, ponieważ uwzględnia ona skorelowanie zmiennych niezależnych. Dystans Mahalanobisa został wyliczony dla 1546 uczniów w wyniku analizy regresji. Uczniów dobierano w pary tak, aby odległość Mahalanobisa między dobranymi w parę uczniami była jak najmniejsza, a różnili się tylko faktem uczęszczania do klasy poniżej 23 uczniów (grupa eksperymentalna) lub do klasy powyżej 22 uczniów (grupa kontrolna). Zdecydowano się na taki podział, ponieważ jak wykazali Glass i Smith (1978) na podstawie metaanalizy 77 badań na temat efektu małej klasy, liczebność klasy ma wpływ na poprawę wyników edukacyjnych uczniów w klasach mniejszych niż 23-osobowe. Podobny wniosek sformułowali Robinson i Wittebols (1986) na podstawie metaanalizy 124 prac z lat 1950–1985.

Dopasowanie przeprowadzono osobno w grupach uczniów wydzielonych według miejsca zamieszkania oraz statusu społeczno-ekonomicznego (*Socioeconomic Status* – SES). Mimo że SES okazał się być nieistotnym predyktorem w modelu regresji, zdecydowano się uwzględnić ten czynnik w procedurze dopasowania, ponieważ wiele badań wskazuje na SES jako istotną determinantę osiągnięć edukacyjnych uczniów. Analogicznie, istnieje wiele dowodów wskazujących, że dzieci ze szkół miejskich osiągają lepsze wyniki od dzieci ze szkół wiejskich. Dopasowanie z osobnym uwzględnieniem dodatkowych zmiennych kategorialnych nazywane jest warstwowaniem. Pozwala uzyskać idealne połączenie jednostek analizy pod kątem zmiennych tworzących warstwy. Dopasowanie przeprowadza się w warstwach, których jest dokładnie tyle, ile wynosi iloczyn liczby kategorii zmiennych branych pod uwagę. W każdej z warstw łączeni byli w pary uczniowie z klas małych i dużych. Łącznie uzyskano 413 par uczniów, w których różnica w dystansie Mahalanobisa między uczniami w parze nie przekraczała 0,1 odchylenia standardowego. Łączenie, w którym arbitralnie określa się maksymalną dopuszczalną odległość między obserwacjami nazywane jest *caliper matching*. Różnice wielkości 0,1 odchylenia standardowego odległości Mahalanobisa gwarantują znaczną redukcję obciążenia szacowania efektu wpływu liczebności klasy na osiągnięcia edukacyjne uczniów. Niewątpliwie, im bardziej rygorystycznie zdefiniowana zostanie wartość progowa (*caliper*) tym uzyskane dopasowanie będzie dokładniejsze.

Drugą procedurą było dopasowanie z wykorzystaniem metody *k*-średnich, przeprowadzone na podstawie zmiennych zidentyfikowanych podczas analizy regresji jako istotne determinanty wyników gimnazjalnych uczniów oraz dodatkowo SES, z powodów opisanych wcześniej. Dopuszcza się przeprowadzenie analizy skupień obserwacji według zmiennych mierzonych na różnych

poziomach pomiaru. Warunkiem jest jednak ich wcześniejsza transformacja. Możliwych jest co najmniej kilka transformacji. W opisywanej analizie zastosowana została standaryzacja, w której wszystkie zmienne zostały podzielone przez swoje odchylenia standardowe. Dodatkowo zmienne dychotomiczne zostały po standaryzacji pomnożone przez wartość 0,707 (Bacher, 2002, s. 165), ponieważ jako miara dystansu między obserwacjami, została zastosowana odległość euklidesowa.

Procedurę dopasowania optymalnego z wykorzystaniem metody *k*-średnich opisuje Johann Bacher (2002). Zbiór danych z badań przeprowadzonych przez krakowską OKE został podzielony na dwie części według liczebności klas, do których uczęszczali badani uczniowie. Grupę eksperymentalną tworzyli uczniowie z klas 22-osobowych i mniejszych. Zbiór, z którego wyłoniono statystyczne bliźnięta, tworzyli uczniowie z klas liczniejszych. W analizowanym zbiorze danych grupę eksperymentalną tworzyło 920 uczniów (53% próby), a grupę porównawczą 813 uczniów (47% próby). Ponieważ analizę *k*-średnich przeprowadzono z wyłączeniem braków danych (LISTWISE), do analizy zostały włączone tylko te obserwacje, dla których posiadano informacje dotyczące wartości wszystkich zmiennych wykorzystanych w procedurze dopasowania. W grupie eksperymentalnej było to 700 obserwacji. Utworzono więc 700 skupień i zapisano w osobnym pliku ich centra, czyli punkty w przestrzeni wielowymiarowej, w których krzyżują się średnie wartości wszystkich zmiennych uwzględnionych w analizie dla danego skupienia (w tym przypadku konkretnej obserwacji). Zapisane centra wykorzystano do klasyfikacji obiektów z grupy porównawczej, z której wyodrębniono grupę kontrolną.

Dopasowanie metodą *k*-średnich przeprowadzono w dwóch wariantach. W pierwszym, jednemu uczniowi z klasy małej przypisano kilku uczniów z klas dużych (dopasowanie „jeden do wielu"). Zaletą tej metody jest zachowanie większej liczby przypadków w próbie efektywnej, co umożliwia uzyskanie bardziej trafnych zewnętrznie wyników. Wadą jest natomiast zwiększenie wariancji oszacowania parametrów. W drugim wariancie jednemu uczniowi z klasy małej został przypisany dokładnie jeden uczeń z klasy dużej (dopasowanie „jeden do jednego"). Grupa kontrolna i eksperymentalna są w tym przypadku równoliczne. Metoda ta została zastosowana także w dopasowaniu z wykorzystaniem odległości Mahalanobisa. Korzyścią płynącą z tej metody jest redukcja wariancji oszacowania parametrów. Słabością, natomiast, mniejsza trafność zewnętrzna wyników, bowiem mniejsza jest łączna liczba obserwacji, w oparciu o które szacuje się wartości parametrów.

Porównanie metod dopasowania

Jakość dopasowania można wstępnie ocenić na podstawie odległości przypadków w grupie kontrolnej do centrów skupień wyznaczonych przez ich odpowiedniki w grupie eksperymentalnej. W przypadku dopasowania „jeden do wielu" odległości przypadków z grupy kontrolnej do swoich odpowiedników w grupie eksperymentalnej wahały się między 1,228 a 5,230 odległości euklidesowej. Jedna czwarta przypadków notuje odległość poniżej 2,276 odległości euklidesowej. Połowa przypadków oddalona jest od swoich odpowiedników o 2,730, a trzy czwarte o 3,147. Analogicznie można opisać rezultaty łączenia „jeden do jednego". Średnia odległość euklidesowa obserwacji z grupy kontrolnej do swoich statystycznych bliźniąt w grupie eksperymentalnej wynosi 2,613. Najlepiej dopasowany przypadek jest o 1,228 odległości euklidesowej oddalony od swojego odpowiednika w grupie eksperymentalnej. Przypadek najdalej oddalony, położony jest w odległości 5,230 odległości

euklidesowej od swojego statystycznego bliźniaka. Dla jednej czwartej przypadków notuje się odległość 2,149, a dla połowy 2,580. Z kolei 75 percentyl wyznacza odległość równa 3,040 odległości euklidesowej. Przedstawione wartości są jedynie poglądowe. Nie istnieje jasne kryterium, na podstawie którego można by orzec o dopasowaniu satysfakcjonującym, czy dyskwalifikującym wyniki dalszych analiz. Niemniej jednak, im mniejsze odległości notujemy, tym otrzymane wyniki są mniej obciążone.

Paul Rosenbaum i Donald Rubin (1983) zaproponowali jako miarę jakości dopasowania jednostek analizy w porównywanych grupach procentowy udział różnicy średnich międzygrupowych w średniej wartości odchylenia standardowego, co można wyrazić wzorem:

$$100 * (\overline{X}_E - \overline{X}_P) / \left[\left(s_E^2 + s_P^2 \right) / 2 \right]^{0,5}$$

gdzie:

$\overline{X}_E$ i $\overline{X}_P$ to średnie wartości testowanej zmiennej odpowiednio w grupie eksperymentalnej i porównawczej,

s_P^2 i s_E^2 to wariancje tej zmiennej w grupie eksperymentalnej i porównawczej.

Obciążenie poniżej 5% uważane jest za nieistotne.

Kolejną metodą walidacji procedury łączenia jest stosowanie testu t-Studenta dla prób niezależnych. Stosując tę metodę mamy możliwość porównania średnich zmiennych, użytych w procedurze łączenia obserwacji, w grupie eksperymentalnej i kontrolnej. Aby dopasowanie jednostek było satysfakcjonującej jakości, nie powinny występować istotne statystycznie różnice średnich między grupami.

Tabela 3 przedstawia standaryzowane różnice procentowe, wartości testu *t* oraz poziomy istotności dla porównania (a) uczniów z klas małych i dużych przed dopasowaniem; (b) uczniów z klas małych i dużych po warstwowaniu i dopasowaniu z wykorzystaniem odległości Mahalanobisa; (c) uczniów z klas małych i dużych po dopasowaniu „jeden do wielu" i (d) „jeden do jednego".

Po przeprowadzeniu procedury łączenia z wykorzystaniem odległości Mahalanobisa udało się uzyskać grupy eksperymentalną i kontrolną jednolite pod względem zmiennych mających wpływ na osiągnięcia edukacyjne uczniów (zmienne niezależne w modelu regresji). Ewentualne różnice w średnich międzygrupowych są nieistotne statystycznie. Porównując procentowe obciążenie, jakie wnoszą poszczególne zmienne, radykalnie zredukowano obciążenie, które wnosiło pięć zmiennych, natomiast znacznie wzrosło obciążenie na dwóch zmiennych.

W przypadku 11 zmiennych udało się uzyskać poprawę dopasowania stosując metodę *k*-średnich w wariancie „jeden do wielu", gorzej dopasowanych jest 10 zmiennych. Wykazano, że średnie żadnej zmiennej (proporcje w przypadku zmiennych binarnych, zarówno oryginalnych, jak i pomocniczych – rekodowanych zmiennych porządkowych) nie różnią się istotnie między grupami eksperymentalną i kontrolną. W przypadku dopasowania metodą *k*-średnich w wariancie „jeden do jednego" zbalansowanych zostało 12 zmiennych. Z kolei dla 10 dopasowanie się pogorszyło. Aby możliwe było poddanie ogólnej ocenie przeprowadzonych procedur łączenia, obliczono sumę kwadratów dla standaryzowanych różnic w średnich dla wszystkich zmiennych uwzględnionych w dopasowaniu. Obliczono także analogicznie sumę kwadratów wartości testu *t* i poziomów istotności.

Suma kwadratów różnic średnich oraz suma kwadratów wartości testu *t* dla do-

Tabela 3

Szczegółowe miary jakości przeprowadzonych procedur łączenia

A – Standaryzowana różnica w % (obciążenie w %) B – Test *t* dla prób niezależnych C – Istotność		Zakres	Przed dopasowaniem			Warstwowanie i dopasowanie z wykorzystaniem odległości Mahalonobisa			Dopasowanie „jeden do wielu”			Dopasowanie „jeden do jednego”		
			A	B	C	A	B	C	A	B	C	A	B	C
Uprzednie osiągnięcia edukacyjne		7–42	9	–1,95	0,05	–10	–1,41	0,16	–11	–1,79	0,07	6	–0,77	0,44
Spodziewana liczba punktów z egzaminu		0–100	9	–1,78	0,07	–3	–0,45	0,65	–8	–1,23	0,22	5	–0,65	0,52
Wiedza zdobywana podczas lekcji w szkole		1–5	5	–1,12	0,26	0	–0,07	0,94	–8	–1,27	0,21	3	–0,4	0,69
Mam zwyczaj pytać nauczycieli o wszystko, co jest niezrozumiałe		1–5	0	0,06	0,95	10	1,43	0,15	–4	–0,65	0,52	7	–0,97	0,33
Ważne jest dla mnie tylko to, żeby przejść do następnej klasy (szkoły)		1–5	–1	0,14	0,89	3	0,45	0,65	4	0,7	0,48	–2	0,34	0,73
Płeć		0–1	–1	0,3	0,76	–5	–0,72	0,47	–4	–0,67	0,5	7	–0,93	0,35
Trudność zadań na egzaminie		0–100	5	–0,96	0,34	–7	–1,04	0,3	–5	–0,81	0,42	7	–0,97	0,33
Dysleksja		0–1	0	–0,05	0,96	–12	–1,77	0,08	7	1,09	0,28	–6	0,85	0,4
Udział w zawodach sportowych w gimnazjum		0–1	–2	0,45	0,65	–2	–0,35	0,72	4	0,7	0,48	–4	0,5	0,62
Motywacja do nauki		0–100	–4	0,89	0,37	2	0,35	0,73	–3	–0,55	0,58	4	–0,59	0,56
Uczniom w mojej klasie zależy, aby na lekcjach jak najwięcej się nauczyć		1–5	–2	0,34	0,73	1	0,1	0,92	–1	–0,11	0,91	4	–0,51	0,61
Miejsce zamieszkania		0–1	–5	0,95	0,34	–1	–0,12	0,91	8	1,32	0,19	–3	0,36	0,72
Status społeczno-ekonomiczny	Bardzo niski	0–1	–15	2,82	0	8	1,21	0,23	8	1,19	0,23	–2	0,25	0,8
	Niski	0–1	–3	0,58	0,56	4	0,55	0,58	–1	–0,17	0,87	2	–0,3	0,77
	Średni	0–1	8	–1,54	0,12	–9	–1,31	0,19	–5	–0,76	0,45	2	–0,29	0,77
	Wysoki	0–1	6	–1,1	0,27	–10	–1,45	0,15	2	0,31	0,75	–5	0,67	0,5
Rodzice sprawdzali odrobione lekcje	Nigdy	0–1	6	–1,17	0,24	7	1,08	0,28	–10	–1,5	0,13	8	–1,08	0,28
	Kilka razy w roku	0–1	9	–1,9	0,06	12	1,71	0,09	–2	–0,39	0,7	–1	0,1	0,92
	Kilka razy w miesiącu	0–1	–5	1,06	0,29	6	0,88	0,38	2	0,39	0,7	–4	0,5	0,62
	Przynajmniej raz w tygodniu	0–1	–12	2,51	0,01	–1	–0,1	0,92	10	1,55	0,12	–5	0,73	0,47
	Kilka razy w tygodniu	0–1	–1	0,11	0,91	–1	–0,14	0,89	4	0,69	0,49	–2	0,24	0,81
Podobieństwo zadań na egzaminie do rozwiązywanych w klasie		0–100	3	–0,7	0,48	1	0,15	0,88	–6	–0,87	0,38	–1	0,17	0,86
Sposób prowadzenia lekcji zachęca mnie do aktywności		1–5	–8	1,73	0,08	0	0,06	0,95	6	0,9	0,37	9	–1,23	0,22

Tabela 4
Ogólne miary jakości przeprowadzonych procedur łączenia danych

	Suma kwadratów różnic średnich	Suma kwadratów wartości testu *t*	Suma kwadratów poziomów istotności
Przed dopasowaniem	960	39	6
Po dopasowaniu „jeden do wielu”	860	21	6
Po dopasowaniu „jeden do jednego”	512	10	9
Po warstwowaniu i dopasowaniu „jeden do jednego” z wykorzystaniem odległości Mahalanobisa	965	20	9

brego dopasowana powinny zbiegać do zera. Z kolei im wyższa suma kwadratów poziomów istotności, tym lepiej. Bezapelacyjnie, wśród zestawionych w tabeli najlepszej jakości dopasowaniem jest łączenie *k*-średnich zrealizowane w wariancie „jeden do jednego”.

Wyniki

Po dopasowaniu „jeden do jednego”, wobec braku istotnych różnic między porównywanymi grupami eksperymentalną i kontrolną (przy kontroli zmiennych uwzględnionych w procedurze dopasowania), średni wynik uczniów w klasach 22-osobowych i mniejszych nie różni się istotnie od średniego wyniku egzaminu gimnazjalnego w klasach bardziej licznych. Średni wynik egzaminu gimnazjalnego w klasach małych wynosi 57,44 punktów (odchylenie standardowe 15,34), a w klasach dużych 56,81 punktu (odchylenie standardowe 15,94). Dla oceny istotności różnicy średnich przeprowadzono test *t*-Studenta. Wyniki testu odczytujemy przy spełnieniu założenia równości wariancji. Na poziomie istotności 0,573 należy odrzucić hipotezę, że średni wynik egzaminu gimnazjalnego w klasach małych jest istotnie różny od średniego wyniku w klasach dużych.

Analogiczna analiza przeprowadzona na grupach wyłonionych metodą warstwowania z wykorzystaniem odległości Mahalanobisa daje podobne rezultaty. Istotność statystki *t* wynosi 0,520. Tak więc należy stwierdzić, że liczebność klasy nie ma istotnego statystycznie wpływu na wyniki egzaminów gimnazjalnych. Uczniowie w mniejszych klasach osiągają średnio o 1,1 punktu (0,068 odchylenia standardowego) lepsze rezultaty niż ich rówieśnicy z klas dużych. W przypadku wyników w grupach wyłonionych procedurą dopasowania „jeden do wielu” metodą

Tabela 5
Porównanie średnich wyników uczniów w grupie eksperymentalnej i kontrolnej za pomocą testu t-Studenta

Test Levene’a jednorodności wariancji		Test *t* równości średnich				
F	Istotność	*t*	*df*	Istotność (dwustronna)	Różnica średnich	Błąd standardowy różnicy
0,490	0,483	0,563	778	0,573	0,630	1,119

Tabela 6
Efekt małej klasy w zależności od zastosowanej procedury łączenia danych

	Jakość dopasowania	ATT	Istotność testu *t*
Po warstwowaniu i dopasowaniu „jeden do jednego" z wykorzystaniem odległości Mahalanobisa	+	0,068	0,52
Po dopasowaniu „jeden do wielu" z wykorzystaniem alaszy skupień	++	0,093	0,13
Po dopasowaniu „jeden do jednego" z wykorzystaniem alaszy skupień	+++	0,039	0,57

k-średnich, średni wynik uczniów w małych klasach wynosi 57,44 punktów (odchylenie standardowe 15,34), a w dużych 55,93 punkty (odchylenie standardowe 15,79). Różnica 1,51 punktu okazała się nieistotna statystycznie.

Aby oszacować efekt liczebności klasy na osiągnięcia edukacyjne (mierzone ogólnopolskim egzaminem gimnazjalnym), wyliczono różnicę średnich wyników egzaminów w porównywanych grupach. Średni efekt oddziaływania bodźca (małej klasy) dla jednostek w grupie poddanej oddziaływaniu bodźca – w zależności od metody dopasowania – wynosi między 0,039 a 0,093 odchylenia standardowego wyników egzaminu. Średni efekt oddziaływania bodźca dla wszystkich obserwacji w próbie wynosi 0,067 (obliczony dla „surowych" danych przed dopasowaniem) (Tabela 6).

Zmienna wynikowa stanowi sumę punktów, jaką dany uczeń uzyskał z części humanistycznej i matematyczno-przyrodniczej egzaminu. Zobaczmy, czy efekt małej klasy zależy od nauczanego przedmiotu. Także statystycznie nieistotny, jednak różny wynik obserwujemy dla części humanistycznej i matematyczno-przyrodniczej w klasach małych i dużych. Uczniowie z klas małych osiągają średnio o 0,69 punktu lepsze wyniki z części matematyczno-przyrodniczej egzaminu niż ich rówieśnicy z klas dużych oraz średnio o 0,06 punktu mniej z części humanistycznej. Jest to potwierdzenie zdroworozsądkowego przypuszczenia, że w małej klasie uczniowie efektywniej uczą się przedmiotów ścisłych niż humanistycznych (Tabela 7).

Zobaczmy jeszcze, gdzie na skali staninowej, służącej do porównywania osiągnięć edukacyjnych, lokują się uczniowie z grupy eksperymentalnej i kontrolnej, wyłonieni w procedurze dopasowania *k*-średnich w wariancie „jeden do jednego" (Rysunek 4). W środkowym (piątym) staninie skali znajdują się uczniowie, którzy osiągnęli przeciętny wynik na egzaminie. Uczniowie z klas małych lepiej poradzili sobie na części matematycznej egzaminu. Była już o tym mowa, tu mamy natomiast prezentację graficzną. W staninie czwartym jest 5% uczniów więcej z klas dużych, z kolei w staninie szóstym 4% więcej uczniów z klas małych. Biorąc pod uwagę wyniki z części humanistycznej, w staninie czwartym jest 2% więcej uczniów z klas małych, a w staninie szóstym 3% uczniów więcej z klas dużych.

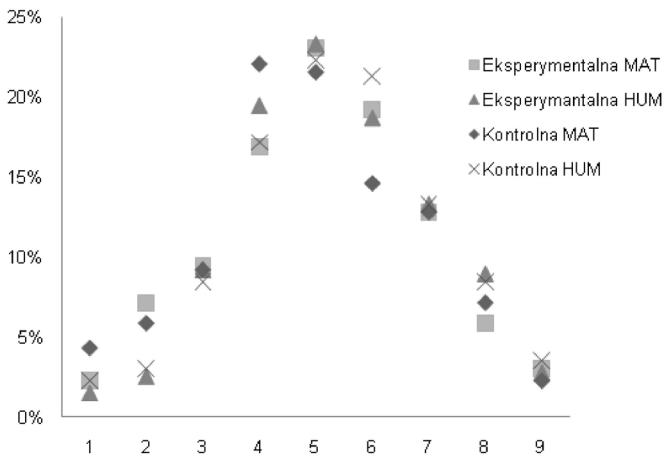
Na podstawie przeprowadzonych analiz należy stwierdzić, że wpływ liczebności klasy na osiągnięcia edukacyjne wśród badanych uczniów nie jest istotny statystycznie. Niemniej jednak uczniowie w klasach małych osiągali na egzaminie gimnazjalnym średnio o 0,039 odchylenia standardowego lepsze wyniki niż ich rówieśnicy z klas ponad 22-osobowych. Przy zachowaniu rygory-

Tabela 7
Efekt małej klasy w zależności od części egzaminu (matematyczno-przyrodniczej i humanistycznej)

Po dopasowaniu „jeden do jednego" z wykorzystaniem alaszy skupień	Różnica punktów	ATT	Istotność testu *t*
Część matematyczno-przyrodnicza	0,69	0,069	0,66
Część humanistyczna	–0,06	–0,008	0,33

stycznych procedur statystycznych i kontroli zmiennych kontekstowych jest to ważny wynik, który wpisuje się w rezultaty publikowane w ramach głównego nurtu literatury światowej traktującej na ten temat. Generalnie raportują one o pozytywnym, jednak trudno obserwowalnym wpływie małolicznej klasy na osiągnięcia edukacyjne uczniów.

Uzyskany wynik nie powinien być jednak interpretowany bez zwrócenia uwagi na możliwe źródła jego zakłócenia. Wśród nich największym jest sygnalizowany wcześniej brak kontroli wpływu nauczycielskiego. Jest to poziom wykształcenia nauczyciela, odbyte kursy, cechy osobowości, zaangażowanie w prowadzenie zajęć, sposób prowadzenia zajęć, umiejętności, doświadczenie pedagogiczne itp. Przeprowadzone analizy były ograniczone pulą zmiennych dostępnych w bazie danych. Brakowało także zmiennych uwikłanych na poziomie szkoły (dostępne pomoce naukowe, infrastruktura itp.). Zarówno wpływ nauczycielski, jak i zmienne na poziomie szkoły były jedynie pośrednio kontrolowane przez wykorzystanie w procedurach dopasowania zmiennej dotyczącej miejsca zamieszkania uczniów. We wszystkich przypadkach miejsce zamieszkania uczniów pokrywało się z położeniem szkoły. Przyjęto tu założenie, że miejsce położenia szkoły jest dobrym wskaźnikiem wpływu nauczycielskiego i wpływu szkoły, szkoły miejskie cechują się bowiem wyższym poziomem nauczania oraz lepszą bazą dydaktyczną niż szkoły wiejskie.



Rysunek 4. Odsetek uczniów w danym staninie.

Należy pamiętać, że uzyskane rezultaty są prawomocne jedynie w zakresie danych. Uzyskane wyniki mogą być zakłócone ograniczeniami związanymi z jakością danych, zastosowanym schematem eksperymentalnym, brakiem kontroli specyficznych mechanizmów selekcji uczniów do klas małych i dużych. Ponieważ w większości są to mechanizmy indywidualne dla danej szkoły, ich pełna kontrola byłaby możliwa tylko w sytuacji randomizowanego eksperymentu. Świadomość wskazanych powyżej ograniczeń otwiera drogę dalszym analizom i badaniom, mającym na celu szacowanie efektu małej klasy.

Literatura

Angrist, J. D. i Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievment. *The Quarterly Journal of Economics, 114*(2), 533-575.

Bacher, J. (2002). *Cluster analysis*. Lecture Notes. Nuremberg: University of Erlangen-Nuremberg.

Biddle, B. J. i Berliner D. C. (2004). Small class size and its effects. *Educational Leadership, 59*(5), 12–23.

Byrne, B., Coventry, W. L., Olson, R. K., Wadsworth, S. J., Samuelsson, S., Petrill, S. A., Willcutt, E. G. i Corley, R. (2010). Teacher effects in early literacy development: evidence from a study of twins. *Journal of Educational Psychology*, *102*(1), 32–42.

Chapin, F. S. (1946). An application of ex post facto experimental design. *Sociometry*, *9*(2/3), 133.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, F., Mood, A. M. i Weinfeld, F. D. (1966). *Equality of educational opportunity*. Washington: U.S. Goverment Funding Office.

Dunning, T. (2008). Improving causal inference: strengths and limitations of natural experiments. *Political Research Quarterly, 61*(2), 282–293.

Educational Research Service. (1980). Class size research: a critique of recent meta-analyses. *The Phi Delta Kappan, 62*(4), 239–241.

Freedman, D., Pisani i R., Purves, R. (1997). *Instructors' Manual for Statistics (3rd ed)*. Department of Statistics, University of California, Berkeley, New York: Norton.

Glass, G. V. i Smith, M. L. (1978). *Meta-analysis of research on the relationship of class–size and achievement. The class size and instruction project*. San Francisco: Far West Laboratory for Educational Research and Development.

Glass, G. V. i Smith, M. L. (1979). Meta-analysis of research on the relationship of class-size and achievement. *Educational Evaluation and Policy Analysis*, *1*, 2-16.

Glass, G. V., Cahen, L. S., Smith, M. L. i Filby, N. N. (1982). *School class size: research and policy*. Sage Publications.

Graue, E., Oen, D., Hatch, K., Rao, K. i Fadali, E. (2005). Perspectives on class size reduction. [Odczyt zaprezentowany 12.04.2005 na sympozjum *Early childhood policy in practice: the case of class size*, w ramach dorocznego spotkania American Educational Research Association]. Montreal, Canada.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*(1), 84–117.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Evaluation and Policy Analysis, 19*(2), 141–164.

Hanushek, E. A. (1998). Conclusions and controversies about the effectiveness of school resources. *FRBNY Economic Policy Review, 4*(1), 11–27.

Hanushek, E. A. (1999). The evidence on class size. W: S. E. Mayer i P.E. Peterson (red.), *Earning and learning: how schools matter* (s. 131–168), Washington, DC: Brookings Institution.

Hanushek, E A. (2002). Evidence, politics, and the class size debate. W: L. Mishel i R. Rothstein (red.), *The class size debate*. Washington, DC: Economic Policy Institute, s. 37–66.

Harris, J. R. (2000). *Geny czy wychowanie?*. Warszawa: Wydawnictwo Czarna Owca.

Hart, S., Petrill, S. i Kamp Dush, C. (2010). Genetic influences on language, reading and mathematics skills in a national sample: an analysis using the National Longitudinal Survey of Youth. *Language, Speech, and Hearing Services in Schools, 41*(1), 118–128.

Heckman, J. J., Ichimura, H. i Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605–654.

Hedges, L. V. i Stock, W. (1983). The effects of class size: an examination of rival hypotheses. *American Educational Research Journal, 20*(1), 63–65.

Herbst, M. i Herczyński, J. (2005). *School choice and*

student achievement. evidence from Poland. Warsaw: Warsaw University.

Jakubowski, M. i Sakowski, P. (2006). Quasi-experimental estimates of class size effect in primary schools in Poland. *International Journal of Educational Research, 45*(3), 202–215.

Molnar, A., Smith, P. i Zahori, J. (2000). *The 1999–2000 evaluation results of the student achievement guarantee in education (SAGE) Program*, CERAI. University of Wisconsin–Milwaukee.

Morgan, S. L. i Winship, C. (2007). *Counterfactuals and causal inference: methods and principles for social research*, Cambridge: Cambridge University Press.

Nye, B., Hedges, L. V. i Konstantopoulos, S. (2000). The effects of small classes on achievement. The results of the Tennessee class size experiment. *American Educational Research Journal, 37*(1), 123–151.

Nye, B., Hedges, L. V. i Konstantopoulos, S. (2001). Are effects of small classes cumulative? Evidence from a Tennessee Experiment. *Journal of Educational Research, 94*(6), 336–345.

Odden, A. (1990). Class size and student achievement. Research-based policy alternatives. *Educational Evaluation and Policy Analysis, 12*(2), 213–227.

Pillmer, D. B. i Light, R. J. (1980). Synthesing outcomes: how to use research from many studies. *Harvard Education Review*, *50*, 170–189.

Rice, J. M. (1902). Educational research: a test in arithmetic. *The Forum*, *34*, 281–297.

Robinson, G. E. (1990). Synthesis of research on the effects of class size. *Educational Leadership*, *47*(7), 80–90.

Robinson, G. E. i Wittebols, J. H. (1986). *Class size research: A related cluster analysis for decision making*. Arlington, Virginia: Educational Research Service.

Rosenbaum, P. R. i Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, *29*, 159–183.

Slavin, R. E. (1986). *Student team learning. An overview and practical guide*. Washington, DC: Professional Library National Education Association.

Sleszyński, P. (2002). *Ekonomiczne uwarunkowania wyników sprawdzianu szóstoklasistów i egzaminu gimnazjalnego przeprowadzonych wiosną 2002 roku*. Ekspertyza wykonana dla MENiS.

Strawiński, P. (2008). Quasi-eksperymentalne metody ewaluacji. W: A. Haber (red.), *Środowisko i warsztat ewaluacji* (s. 193–220). Warszawa: PARP.

Strawiński, P. (2007). Przyczynowość, selekcja i endogeniczne oddziaływanie. *Przegląd Statystyczny*, *4*, 49–61.