Edukacja
An interdisciplinary approach 2013, 1
ISSN 0239-6858
pp. 41–58

Estimating the effect of class size on academic achievement by ex post facto experiment

Maciej Koniewski
Institute of Sociology, Jagiellonian University*

The analyses of influence of class size on academic achievement used data from study conducted in 2006 by the Regional Examination Board in Cracow (Poland). The variables explaining the achievements of lower secondary school pupils were identified using regression analysis. The model explains 71% of variance of exam results. These variables were used to identify statistical twins. Their assignment to the experimental and control group was performed in three ways: by stratification using Mahalanobis distance, matching one-to-many and one-to-one using *k*-means method. The last method proved the most successful. The effect of class size on student outcomes proved statistically insignificant. However, pupils from classes with below 23 pupils achieved higher mean scores than their peers from larger classes by 0.039 standard deviation.

Keywords: sociology of education, class size effect, academic achievement measurement, quasi-experimental research.

Effective educational policy, aiming at rational allocation of available resources in order to maintain improvement of the quality of education, should interpret research results with caution. One role of educational research is to provide knowledge on the strength and direction of the relation between the quality of teaching and other factors, particularly those that can be influenced by administrative and financial decisions. One of those factors is class size. This problem is rarely raised in public debate in Poland but periodically emerges from the shadow of other important social and political topics.

Class size, as a quantity that can potentially be optimised, is of interest to parents, teachers, headmasters and governing bodies of schools. Parents and teachers both favour small classes. The former, because they believe that in smaller classes children learn more effectively and the latter, because in small classes their work is more comfortable. As remuneration of teachers is the main factor in expenditure on education, headmasters and governing bodies of schools are generally interested in keeping classes larger due to savings.

Article based on M.A. thesis prepared under supervision of Prof. Jarosław Górniak, Ph. D., in the Institute of Sociology of the Jagiellonian University, Cracow. This article was published primarily in Polish language in *Edukacja, 117*(1) 2012.

* Mail address: Maciej Koniewski, Zakład Socjologii Gospodarki, Edukacji i Metod Badań Społecznych, Instytut Socjologii UJ, ul. Grodzka 52, 31-044 Kraków, Poland. E-mail: maciej.koniewski@uj.edu.pl

An important issue concerning effective public spending and raising the quality of education, knowledge about the relation between class size and academic achievements of pupils is significant in deciding on size of classes. Since this issue is complex, the debate on the impact of class size is ongoing and unresolved. Many variables, potentially of bearing on academic achievements are still beyond the reach of measurement tools and inventiveness of researchers. Undoubtedly, however, the problem of the optimum number of pupils in a class will remain a constant factor for state education policy and as such should be constantly explored.

The results of the existing research on small class effect

The influence of class size on academic achievements has been the subject of many studies since the beginning of the 20th century. The first study on this topic was conducted by Rice (1902). Experimental studies allow precise assessment of variation of student achievement in relation to class size but are limited in their dependence on study context and relatively small sample sizes. A second approach is econometric analysis with class size data, although more frequently student-teacher ratio is used, modelling relations between class size and academic achievement are measured by national-wide assessments. Such analyses are frequently conducted on population data but the possibility of controlling contextual factors is limited.

Some authors agree on the positive influence of small classes on the quality of teaching: "Although the results of small scale randomised experiments and large-scale econometric studies point to positive effects of small classes, some scholars have seen the evidence as ambiguous" (Nye et al., 2000, p. 123). Hanushek (1999; 2002) and Odden (1990) claim that the cost of reduction in class size is incommensurate with the results obtained. The strength of the effect itself is-controversial[1].

The most exhaustive meta-analysis of research on class size is contained in the work of Glass and Smith (1978; 1979). Their analysis included 77 studies conducted over 70 years. In total in all analysed researches involved 900 000 pupils. The main conclusion of their work supports the existence of the positive small class effect on achievement of classes fewer than 23 pupils. The effect is independent of subject taught, pupils intelligence and influence of basic demographic characteristics. These authors showed that individual teaching is more effective than teaching in classes of 40 pupils by *SD* 0.565 academic achievement measure. Negative correlation between class size and achievement was stronger when pupils were randomly assigned to classes of various sizes. In research conducted before 1940 no relation between class size and achievement was observed, while a strong relation was shown in studies after 1960. This may be interpreted as the effect of advances in measurement methods and more sophisticated experimental schemes.

Robinson and Wittebols (1986) applied cluster analysis to classification of research on effects of class size conducted in 1950–1985. Among 124 studies included in their meta-analysis, 54 (44%) favoured smaller classes, 60 (48%) showed no relation and 10 (8%) favoured large classes. The clearest negative correlation between class size and achievement was observed among eight and nine year old pupils. Positive effects of small class size related mostly to reading and arithmetic. Effects were more pronounced in classes of fewer than 23. However, as they mention, the

---

[1] The results of research on class size effect were collected in meta-analyses and systematic reviews (Briddle and Beliner, 2004; Educational Research Service, 1980; Glass et al., 1982; Glass and Smith, 1978; Graue et al., 2005; Hedges and Stock, 1983; Molnar, Smith and Zahori, 2000; Nye, Hedges and Konstanopoulos, 2001; Pillmer and Light, 1980; Robinson, 1990; Robinson and Wittebols, 1986; Slavin, 1986).

positive effects of smaller classes are frequently unstable. Robinson (1990) concludes that reduction of class size has a relatively small positive effect compared with other cheaper interventions or strategies aimed at improving teaching. Class size as an independent factor has little influence, irrespective of subject, especially in classes of 23 to 30. Without teaching suitably adapted to smaller class size, little improvement can be anticipated.

Graue et al. (2005) and Molnar, Smith and Zahori (2000) more recently confirm a positive effect from reduced class size. Additional evidence supporting benefits of small classes, particularly for pupils from ethnic and national minorities and groups with low social status, is provided in Nye's 2001 study. The studies from the last four decades of the 20th century, which contributed the most to knowledge about how class size relates to academic achievement, were mainly large scale experiments conducted in the United States, where accordingly it was decided to reduce class size in some states.

Briddle and Berliner (2004) sum up the results of large scale experimental educational research conducted in the United States in the 1960s–1990s. Their most important conclusions confirm that well planned and adequately funded programmes for class size reduction in the early phases of education improve academic achievement. The longer a student attended a small class, the greater and more stable the benefit. The positive effects of small classes are clearly noticeable in primary school classes with fewer than 20 pupils, irrespective of gender, subject taught and measure of achievements. Those most benefiting from small classes were from poor families and members of national and ethnic minorities. Pupils attending small classes during the early stages of education also maintained their better performance in larger classes, later in their education.

The second most frequent approach to measuring class size effect on academic achievement is econometric analysis. This article focuses on experimental studies. However, it is worth to present at least short summary of econometric studies outcomes. Hanushek (1998) collected 90 publications satisfying high subject-matter and methodological criteria, including 377 separate estimates of school production functions. The author grouped the data according to estimated positive or negative correlation between student-teacher ratio and student achievement. Statistically significant cases were 13% positive and 15% negative.

Few works by Polish authors address the issue of class size. Jakubowski and Sakowski (2006) used methods, which allow abstraction of class size effect based on secondary data analysis, which includes variables characterising schools and the exam results of primary school pupils in Mazowieckie voivodeship in 2002–2004. The problem faced in this type of analysis is endogenity which appears when student attributes simultaneously determine assignment to a small or a large class and influence the dependent variable, i.e. test scores (Strawiński, 2007). Jakubowski and Sakowski dealt with endogenity in two ways. This first used mean class size in a given year in a given school as an instrumental variable for the actual class size. Additionally, the authors controlled differences between schools. The second method involved selecting units for analysis based on Maimonides' rule (Angrist and Lavy, 1999). Only schools creating new classes when the number of pupils in a given year exceeded 29 pupils or multiples of that number were analysed. The mean size of classes was used as the instrumental variable. The results obtained, in the majority of cases were statistically significant, indicating a small but positive influence on scholastic academic student achievements. Retaining relatively small classes was most beneficial in schools in rural areas. Other relevant Polish studies include Śleszyński (2002) and Herbst and Herczyński (2005) but their findings were not conclusive.

Drawing causal inferences

Confusion in reporting and interpreting results of scientific research frequently results from confusing correlation with causality. As an example, a relation between eating breakfast daily and scholastic academic achievements of pupils can be shown and the two phenomena are mutually correlated. This means that occurrence of one is frequently connected with occurrence of the other. Regression analysis is helpful in identifying the relation. Concurrence of the two phenomena does not however infer causality.

Children who do not eat breakfast regularly may come from poorer families or more frequently miss classes, which in turn determines their poorer results. The relation between eating breakfast every day and school results is ostensible and the connections between these effects may be explained by other mediator variables. The claim that eating breakfast every day improves school results requires verification by means of methods guaranteeing high internal validity of the results obtained.

The results presented in this article aim at supporting the hypothesis of the influence of class size on academic achievement. Statements of causality are valid only if three basic requirements for all causal relations are met: (1) the cause precedes the effect, (2) the cause co-varies with the effect and (3) alternative explanations of causal relation are impossible. The requirements can be met by experiment, considered to be the "gold standard" for scientific research. In an experiment the researcher manipulates the stimulus in order to force its occurrence before the effect. Covariance of cause and effect can be checked using statistical analysis. The third requirement is met by randomisation.

The basic logic of experimental research allows for comparison of the outcome variable for people exposed to the stimulus with the outcome variable for people not exposed (control group). In ideal conditions an outcome variable should be measured simultaneously in those who have experienced and those who have not experienced the stimulus, which of course, is not possible. The problem of proving causal effect is a problem which results from missing data (Heckman, Ichimura and Todd, 1997). Subjects who have experienced the stimulus (experimental group) are matched with people who have not (control group) and the outcome variable is measured for these groups. This method is called "invoking a counter-factual state". In other words, the most similar people in the control group and in the group subjected to the stimulus are compared.

In randomised experiments, also called true experiments, the effect of similar composition of both groups is obtained by random selection of people to both of them. Random selection to the comparison group is called randomisation and use of this method ensures that the outcome variable is independent of both the observed and unobserved factors (other than the stimulus), as the variables are randomly distributed among the groups.

If randomisation is not possible for financial or ethical reasons or when secondary data are used, a method to prevent the effect of observable factors (alternative explanations of causal relationship in question) is to perform statistical stratification or data matching after the study. These statistical techniques allow the matching of experimental and control groups in respect of variables which correlate with the dependent variable and allow appropriate selection of subjects to the experimental and control groups. A better approximation to the theoretically ideal situation in which units of analysis are randomly assigned to the groups can be achieved in this way. All potential factors other than the stimulus, influencing change in the dependent variable are randomly distributed between the compared groups). After the matching procedure groups are similar in respect of attributes that could

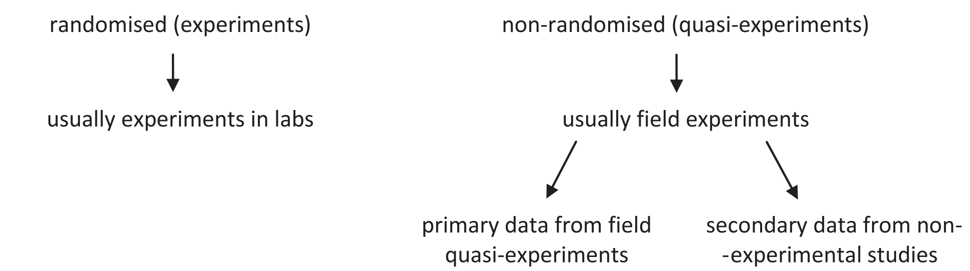

Figure 1. Basic typology of experimental and similar methods.

constitute a potential source for measurement bias in the dependent variable, which results in "resetting" the influence of alternative factors on the dependent variable (the factors which according to theory are expected to correlate with the dependent variable and the stimulus).

Application of experimental logic

Freedman, Pisani and Purves (1997) pointed out three attributes characterising randomised experiments. First, reaction of the experimental group to the treatment can be compared to the reaction of the control group to the controlled conditions, i.e. lack of treatment. Second, assignment to experimental groups is random. Third, stimulus is controlled by the researcher. The three criteria play a crucial role in the experimental model of causality.

When randomisation is not performed, the situation is a quasi-experiment. The greatest contribution to popularisation of the notion of quasi-experiment and quasi-experimental schemes was that of Campbell. According to Dunning (2008, p. 289), by quasi-experiment Campbell understood, "an approximation to the real experimental template", i.e. comparing reaction of units in conditions of exposition to a treatment to that without exposition to the treatment.

Even though there is no random assignment in a quasi-experiment, the researcher can still, under certain conditions, claim that assigning subjects to conditions of the presence of the stimulus or control conditions is "as if" random (Dunning, 2008), as opposed to other non-experimental methods. Such a claim can be justified both from a priori argumentation and empirical evidence. The latter allows control of (observable) factors potentially influencing the dependent variable. Excluding possible influence on the dependent variable is not possible in quasi-experiments, however, indirect control, founded on knowledge from strong theoretical assumptions is possible. The main and sometimes only difference between quasi-experiments and "real" experiments is non-random assignment (of subjects) to the control and treatment (experimental) groups (Figure 1).

Experimental educational research is characterised by being conducted in the natural environment of the subjects. Unlike experiments conducted in artificial conditions (a laboratory), natural experiments are studies conducted in the environment of the subjects or using data from observation of a naturally occurring phenomenon. Since in such conditions a researcher is not able to manipulate the stimulus, natural experiments are actually observational studies (Dunning, 2008). When the need to use data from observations of naturally occurring phenomena occurs, a specific type of quasi-experimental study is used, ex post facto experiment. It transforms data such that they can be analysed. The term "ex post facto experiment" was proposed by Chapin (1946) to describe studies transforming non-experimental data into experimental data, e.g. from data originating in cross-sectional or longitudinal studies.

i	Z_i	$Y_i(T)$	$Y_i(C)$
1	T	*	?
2	T	*	?
3	T	*	?
4	C	?	*
5	C	?	*
6	C	?	*

* marked as empirical data (observable)
? marked as missing data

Figure 2. A database sample from the experimental study. Data for six pupils.

Methodology of the small class effect estimation

The aim of this study is to estimate the influence of class size on academic achievements of pupils using ex post facto experiment method. Superiority of research using the experimental logic, including the approach used here, over correlational research is in its potential to allow drawing causal inference. In other words, findings could "prove" the existence of influence of the specific independent variable (class size) on the dependent variable (academic achievement).

The framework of analyses aiming at estimating the causal effect is determined by the Rubin Causal Model (RCM). It can be pictured as follows (Figure 2): causal effect for a pupil (i) in a small class (T) versus in a large class (C) for outcome variable Y amounts to $E_i = Y_i(T) - Y_i(C)$. Inclusion into a small class (Z_i) does not determine the value of the expected (predicted) result for $Y_i(T)$, $Y_i(C)$ pair, but will determine which of them can be observed. Result $Y_i(T)$ can be observed only when a pupil is in a small class (experimental group); result $Y_i(C)$ can be observed only when a pupil is in a large class (control group). The mean causal effect is estimated as: E = Y(T) – Y(C).

Random assignment to the experimental group implies that the mean result of post-test in the experimental group yT is a valid and unbiased estimation of Y(T), and the mean result of the post sub-test in control group yC is a valid and unbiased estimate of Y(C). Additionally, the difference between means in both groups: yT – yC is a valid and unbiased estimate of causal effect (E).

In estimating the mean effect of the stimulus, independence of the outcome variable from the selection mechanisms to experimental conditions is assumed. The Stable Unit Treatment Value Assumption (SUTVA) is an a priori assumption stating that the value of the outcome variable Y for a pupil (*i*) subjected to stimulus *i*(T) will be constant, irrespective of the mechanism of assignment of the pupil (*i*) to conditions T, as well as irrespective of effect of other stimuli that other pupils may be subjected to (Morgan and Winship, 2007, p. 37). In literature the mean effect of stimulus on all units in the sample is called ATE (Average Treatment Effect). Its counterpart for units in the experimental group is average stimulus effect on units subjected to the stimulus – ATT (Average Treatment for the Treated), and for those in the comparison group the average stimulus effect on units is not subject to stimulus – ATC (Average Treatment Effect for the Controls).

In order to estimate the effect of class size on academic achievement, data was used from research of lower secondary school pupils by the Regional Examination Board *(Okręgowa Komisja Egzaminacyjna* – OKE) in Cracow, Poland shortly after their lower secondary school exam in May, 2006. The

NA X O_1

NA O_2

Where:

- X – stimulus
- O_1, O_2 – posttest
- Line (-----) – NA (nonrandom assignment) students have not been randomly assigned to the compared groups.

Figure 3. Experiment with post-test and non-equivalent groups.

sample for the OKE study was random, drawn using a stratified sampling scheme. The study was mass conducted in 28 schools, in 83 third grade classes. In total 1757 completed questionnaires were collected. In 1733 cases it was possible to match questionnaire data with pupils' lower secondary school exam scores.

The data from cross-sectional studies were transformed into "experimental" data. The methodology of data preparation, as well as the analysis itself, conformed to an implementation of an ex post facto experiment according to the logic of quasi-experimental schemes, with final measurement (post-test) and non-equivalent groups (experimental and control group). The scheme is presented in Figure 3.

Potential threats to validity of the result

The ex post facto experiment introduced here, introduces at least two threats to internal validity. The first is lack of random assignment of units of analysis to experimental conditions (the group subjected to the treatment, in this case a small class, and the control group is a large class). This also contributes to the problem of alternative variables explaining the change in outcome variable. Identification of these variables could be based on theory. However, in the case of ex post facto experiment there are significant restrictions connected with variables available in the database. The more identifiable and controllable variables sharing the variance with the dependent variable, the more credible the results obtained.

The ex post facto experiment scheme presented above did not satisfy the requirement for complete independence of observations from selection to experimental conditions (SUTVA). Only the explicit and measured selection factors, for example, gender and place of residence in the OKE study were controlled and which simultaneously correlated with the level of academic achievement. Other implicit factors, or those which are explicit but not measured in the study were beyond control. Introduction of quasi-market mechanisms into the financing system of Polish education had allowed flexible catchment areas; carers and children to choose between schools and headmasters and teachers responsibility for selection of pupils. Lack of control over these factors requires great caution in the interpretation of such data.

The second potential source of bias in the results presented here is lack of initial measurement (pre-test) of the dependent variable. In this case the dependent variable is the pupil score from the lower secondary school exam. The initial measurement would provide knowledge on "entry" differences among participants of the experiment. Reduction in bias resulting from lack of randomisation as well as lack of pre-test is possible by means of conducting statistical matching of units in experimental and control groups. Lack of pre-test is compensated for by inclusion of variables taken into account by the matching procedure, i.e. earlier results of the pupils (marks from seven subjects in the first semester of third grade of lower secondary school).

Apart from possible sources of bias connected with the experimental scheme alone, there are important criticisms of the OKE study: (a) This was not a whole-country study, as it only included Małopolskie, Lubelskie and Podkarpackie voivodeships; (b) Sample selection based on the scheme from 2004 did not account for the change in the school network that took place before 2006, i.e. the year the study was conducted; (c) The scheme was random only at a school level; (d) In schools drawn for the study, questionnaires were only administered to third grade pupils present at school on that day; (e) The number of absences on that day were not known; (f) The size of classes was not a variable in the study; instead it was a calculation derived from the number of pupils in a given class who participated in the study, i.e. the total of 1757 pupils. Therefore, the assumed size of any class could have been higher, owing to absence.

Identification of co-variants of academic achievements

When planning experiments without random assignment, identification of potential sources of variance for the outcome variable other than from the stimulus and controlling for them is vitally important. In ex post facto experiment the number of alternative explanations for the dependent variable follows the available number of variables. This is a significant limitation to the method, especially if the researcher does not have access to variables which could potentially account for variance.

Based on existing studies the following factors have potential influence on academic achievement:

- individual factors (e.g. genetic, self-assessment, aspirations, motivation, interests, previous school results, time devoted to study, intelligence, state of health);
- environmental factors (family situation and peer environment, e.g. parents' education, socio-economic status, family model, number of siblings, place of residence, conditions for doing homework, parental attitude towards learning, parental aspirations, cooperation between the parents and school, peer environment, school results of peers, their cultural, economic and social capital, attributes of school culture determined by its social composition);
- institutional and pedagogical factors (e.g. model and programme of the school, number of pupils in a class, material resources of the school, timetable, planning of classes and homework, education and experience of teachers, cooperation between teachers, teaching methods and methods of checking progress, in-service training of teachers, teachers' attitudes towards pupils, textbooks and curricula, after-school activities).

There is no agreement as to which variables unambiguously favour positive results. The variables deemed to have the greatest influence on results are social status, connected with place of residence and family environment; internal motivation; aspirations of "significant others" and influence of the peer group.

Influence of family and peer environment on academic achievements is the strongest as indicated by the work of Coleman (1966) and Hanushek (1992; 1997). More recent research, however, suggests overestimation of the impact of family (common environment) at the expense of influence of the genetic factor (Byrne et al., 2010; Harris, 2000; Hart, Petrill and Kamp Dush, 2010).

As a result of study conducted by OKE in 2006 among the pupils of the last grade of lower secondary school, data was collected to allow assessment of social status, motivation

Table 1
Statistics summing up the regression model

Block	R	R^2	R^2 adjusted	SE
1	0.843	0.710	0.708	8.800
2	0.844	0.713	0.710	8.768

R^2 change	F	$Df1$	$Df2$	$Sig.$
0.710	274.209	14	1567	0.000
0.003	3.916	4	1563	0.004

for learning, aspirations of parents and peers and previous academic achievement. After review of available variables, to identify alternative explanations of variability of exam results, a regression model was defined in which the dependent variable was the lower secondary school exam result. A student could obtain from 0 to 100 total points total from the humanities and the maths and science exam papers. In the sample under consideration the minimum value of the variable was 9 and the maximum 99. The median score was 55 and the mean, 55.9 points. The distribution is slightly right-skewed (0.025). From over one hundred questions included in the questionnaire and various combinations of their components, 15 independent variables were finally incorporated into the regression model. Contrary to the expectations social status, education of parents, aspirations of significant others, aspiration of the peer group were not shown to be significant.

The model explained 71% of variance in lower secondary school exam results (Table 1). The variables were introduced to the model in two blocks. In the first block all variables measured at quantitative levels were introduced. Variables representing the categories of ordinal variable such as, "Parents checked the homework" were introduced in the second block. These are so-called dummy variables, coded by 0 or 1. The variable representing the category "never" served as a reference category. The table presents regression analysis outcomes.

The β coefficients were calculated based on standardised variables and is thus independent of units of the individual variables. This allows comparison of the strength of relation between particular variables and the explained variable. The strongest predictors in the model were previous academic achievement and the expected exam score.

The semi-partial correlation coefficient after squaring shows what part of the total variance of the dependent variable is reducible to exclusive influence of a given independent variable. As much as 11% (0.338^2) of variance of lower secondary school exam results is explained by previous academic achievement. The estimated exam score predicts 5% (0.222^2) variance of exam results.

Identifying statistical twins

Selection of experimental and control groups was conducted using various methods to allow comparison of the quality of matching approaches. Matching was performed using Mahalanobis distance, a measure of the characteristics of objects. It expresses the distance of observation from a centroid which is a point of balance in multidimensional space established by independent variables taken into account in the regression model. This measure was adopted as it accounts for the correlation of independent variables. The Mahalanobis distance was calculated for 1546 pupils in regression analysis. Pupils were matched into pairs such that the Mahalanobis

Table 2

Regression analysis (detailed outcomes)

		B	*SE*	β	T	*Sig.*	Zero-order	Partial	Part
	(Intercept)	9.741	1.790		5.441	0.000			
Individual factors	Prior student achievements[a]	1.375	0.055	0.541	24.968	0.000	0.762	0.534	0.338
	Expected exam score[b]	0.344	0.021	0.297	16.354	0.000	0.682	0.382	0.222
	Your methods of learning during school classes (1 = Definitely no/…/5 = Definitely yes)	1.102	0.250	0.063	4.412	0.000	0.236	0.111	0.060
	I have tendency to ask teachers about everything I do not understand (1 = Definitely no/…/5 = Definitely yes)	-1.012	0.203	-0.073	-4.994	0.000	-0.043	-0.125	-0.068
	Only promotion to the next grade/school is important for me (1 = Definitely no/…/5 = Definitely yes)	-0.959	0.160	-0.093	-6.006	0.000	-0.427	-0.150	-0.081
	Sex (0 = Boy/1 = Girl)	-1.995	0.488	-0.061	-4.090	0.000	0.078	-0.103	-0.055
	Place of residence (0 = Urban area/1 = Rural area)	-1.419	0.471	-0.044	-3.012	0.003	-0.185	-0.076	-0.041
	Dyslexia (0 = No/1 = Yes)	3.607	0.838	0.060	4.303	0.000	0.075	0.108	0.058
	Did you take part in sports competitions in lower secondary school? (0 = No/1 = Yes)	-1.750	0.455	-0.054	-3.841	0.000	-0.024	-0.097	-0.052
	Motivation	-3.676	1.495	-0.047	-2.459	0.014	0.338	-0.062	-0.033
Environmental factors	Students in my class are set on learning as much as possible during classes	-0.549	0.209	-0.040	-2.632	0.009	-0.200	-0.066	-0.036
	Parents checked homework several times in a year	0.277	0.487	0.016	0.568	0.570	-0.048	0.014	0.008
	Parents checked homework several times a month	1.493	0.519	0.084	2.879	0.004	-0.082	0.073	0.039
	Parents checked homework at least once a week	-0.562	0.513	-0.032	-1.095	0.274	-0.115	-0.028	-0.015
	Parents checked homework several times a week	-1.607	0.558	-0.090	-2.881	0.004	-0.114	-0.073	-0.039
Institutional and pedagogical factors	The way the classes are taught encourages me to be active	-1.190	0.242	-0.079	-4.922	0.000	-0.066	-0.124	-0.067
	Assess the level of difficulty of problems solved in class as compared to those at the exam[c]	5.157	0.872	0.083	5.913	0.000	0.226	0.148	0.080
	Did you solve problems in Polish and maths during classes at school that were similar to those at the exam?[d]	2.562	0.761	0.048	3.368	0.001	0.181	0.085	0.046

(a) Sum of school grades from Polish, history, maths, biology, chemistry, physics and geography obtained at the end of the first semester of third grade of lower secondary school.

(b) Sum of answers to questions regarding the expected number of points from humanities and maths-and-science part.

(c) 1 – Tasks solved at school were easier than those at the exam; 2 – Tasks solved at school were similarly difficult as those at the exam; 3 – Tasks solved at school were more difficult than those at the exam; 4 – Hard to say, I cannot assess that. This is a factor created based on answers to the question on difficulty level of tasks solved in class in chemistry (factor load equals to 0.759), physics (0.759), maths (0.734), biology (0.695), Polish (0.676), geography (0.674), history (0.642). The factors were extracted using Categorical Principal Component Analysis (CATPCA) method. Observations with lacking data were excluded by pairs. The first dimension explains $0.759^2 + 0.759^2 + 0.734^2 + 0.695^2 + 0.676^2 + 0.674^2 + 0.642^2 = 50\%$ of the initial total variance. Cronbach's Alpha amounts to 0.831. The variable was brought to the scale of 0–100 (min-max normalisation).

(d) 1 – No; 2 – Yes, from time to time; 3 – Yes, quite often; 4 – Hard to say. Factor created on the basis of answers to the question regarding Polish (factor load equals to 0.824) and maths (0.838). The factor was extracted using Categorical Principal Component Analysis (CATPCA) method. Observations with lacking data were excluded by pairs. The dimension explains $0.824^2 + 0.838^2 = 69\%$ of the total variance of the two variables. The variable was brought to the scale of 0–100 (min-max normalisation).

distance between the pupils in a pair was as low as possible. The only difference between pupils was class size below 23 (experimental group) or over 22 pupils (control group). This division was justified on the basis of Glass and Smith (1978) conducted on the meta-analysis of 77 studies on small class effect. They found that class size had an influence on academic achievement in classes with fewer than 23 pupils. A similar result was found by Robinson and Wittebols (1986) in a meta-analysis of 124 studies from 1950–1985.

Matching based on pupils' place of residence and socio-economic status (SES) was conducted separately in groups of pupils attending small and larger classes. Although SES had proved to be insignificant as a predictor in the regression model, a decision was taken to account for this factor in the matching procedure, as many studies identified SES as an important determinant of academic achievement. Similarly, there are several pieces of evidence indicating that children from urban schools achieve better results than their peers in rural areas. Matching with separate account for additional categorical variables is called stratification. It allows the ideal combination of units for analysis with respect to variables, creating strata. Matching is conducted in strata, the number of which equals exactly the product of the number of categories of variables taken into account. Pupils were combined into pairs from large and small classes in each stratum. In total, 413 pairs of pupils were obtained, for which the difference in Mahalanobis distance between the pupils in a pair did not exceed 0.1. Matching, in which the maximum allowable distance between observations is arbitrarily established, is called a *calliper matching*. Differences in Mahalanobis distance amounting to 0.1 guaranteed a significant reduction in estimation bias of the effect. The more rigorously defined the threshold value (*calliper*), the more precise the obtained matching.

The second matching procedure used was the *k*-means method, which is based on variables identified during regression analysis as significant determinants of lower secondary school results of pupils and additionally SES for the reasons described above. Cluster analysis was performed according to variables measured at different levels of measurement, provided their previous transformation. Several such transformations are possible. The analysis described here used standardization, in which all variables were divided by their standard deviations. Additionally standardised dichotomous variables were multiplied by 0.707 (Bacher, 2002, p. 165), as the measure used for distance between observations was the Euclidean distance.

The procedure of optimal matching using the *k*-means method is described in Bacher (2002). A data set from Cracow OKE was divided according to class size. The experimental group consisted of pupils from classes of 22 pupils and fewer. Statistical twins were chosen from the larger classes. The experimental group included 920 pupils (53% of the sample), and the comparison group 813 pupils (47% of the sample). The *k*-means analysis was conducted excluding missing data (LISTWISE). In the experimental group there were 700 observations, so, 700 clusters were created and their centres, i.e. points in the multi-dimensional crossing space for mean values of all variables (a specific observation) were recorded. The recorded centres were used for classifying objects in the comparison group from which a control group was isolated.

Matching using the *k*-means method was accomplished in two ways. In the first, one student from a small class was assigned to more than one student from large classes (one-to-many matching). The benefit of this method is to maintain a larger number of cases in the effective sample, which allows more externally valid results to be obtained. An increase in the variance of parameter estimation is a disadvantage. In the second variant, one student from a small class was matched to one student from a large class (one-to-one matching). The control and experimental group in this case are equal in size. This method was also used to match Mahalanobis distance. The advantage of this method is the reduction in the variance of parameter estimation. The weakness is lower external validity, as the total number of observations on which the values of parameters are estimated, is lower.

Comparison of matching methods

The quality of matches can be initially assessed on the distance between cases in the control group and the centres of the clusters determined by their counterparts in the experimental group. In the case of one-to-many, matching distances between control group members and their experimental group counterparts varied between 1.228 and 5.230 of the Euclidean distance. One quarter of cases were below 2.276 of the Euclidean distance. Half the cases were 2.730 apart and three-quarters by 3.147. The results of one-to-one matching can be described similarly. The mean Euclidean distance of observations from the control group to their statistical twins in the experimental group was 2.613. The best matched case was placed 1.228 from its counterpart in the experimental group. The greatest pair separation was 5.230. For one quarter of the cases a distance of 2.149 is recorded and for half of 2.580. The 75th percentile is at a distance of 3.040. These values are only illustrative. There is no clear criterion allowing determination of whether a match is satisfactory or not. However, the smaller the distance, the less biased the result.

Rosenbaum and Rubin (1983) proposed a percentage share of the difference between inter-group means in the mean value of standard deviation as a measure of match quality. This can be expressed by the formula:

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

where:

- $\overline{X}_E$ and $\overline{X}_P$ are the mean values of the tested variable in the experimental and control group, respectively;
- s_E^2 and s_P^2 are variances of this variable in the experimental and control groups.

Bias below 5% is accepted as insignificant.

Another method for validation of the matching procedure is to use Student's *t*-test for independent samples. Using this method, means can be compared for variables used in observation matching procedure. For a match to be of satisfactory quality there should be no significant statistical differences in means between the groups.

Table 3 presents standard percentage differences, values of *t*-test and levels of significance levels for comparison of (a) pupils from small and large classes before matching; (b) pupils from small and large classes after stratification and matching using Mahalanobis distance; (c) pupils from large and small classes after one-to-one matching and (d) one-to-many matching.

After the matching procedure using Mahalanobis distance it was possible to obtain a homogeneous experimental and control group in respect of variables impacting on academic achievement (independent variables in the regression model). Differences in means, if any, between the groups were statistically insignificant. When comparing the percentage bias caused by particular variables the bias caused by five variables was radically reduced, while bias for two variables increased.

In the case of 11 variables an improved match was obtained using the *k*-means method in the one-to-many variant. While inferior matching was obtained for 10 variables, it was demonstrated that no variable means (proportions for binary variables, both original and dummy – of recoded ordinal variables) were significantly different

Table 3

Detailed measures of quality of the conducted matching procedures

(A) Standarised difference in % (bias expressed in %) (B) Student's *t*-test for independent samples (C) Significance			Before matching			After stratifying and matching based on Mahalanobis distance			After one-to-many matching			After one-to-one matching		
		Scale range	A	B	C	A	B	C	A	B	C	A	B	C
Prior student achivements		7–42	9	-1.95	0.05	-10	-1.41	0.16	-11	-1.79	0.07	6	-0.77	0.44
Expected exam score		0–100	9	-1.78	0.07	-3	-0.45	0.65	-8	-1.23	0.22	5	-0.65	0.52
Methods of learning during school classes		1–5	5	-1.12	0.26	0	-0.07	0.94	-8	-1.27	0.21	3	-0.4	0.69
Tendency to ask teachers about everything. which is not understandable		1–5	0	0.06	0.95	10	1.43	0.15	-4	-0.65	0.52	7	-0.97	0.33
Only promotion to the next grade/school is important for me		1–5	-1	0.14	0.89	3	0.45	0.65	4	0.7	0.48	-2	0.34	0.73
Gender/Sex		0–1	-1	0.3	0.76	-5	-0.72	0.47	-4	-0.67	0.5	7	-0.93	0.35
Difficulty of tasks at the exam		0–100	5	-0.96	0.34	-7	-1.04	0.3	-5	-0.81	0.42	7	-0.97	0.33
Dyslexia		0–1	0	-0.05	0.96	-12	-1.77	0.08	7	1.09	0.28	-6	0.85	0.4
Taking part in sports competitions in lower secondary school		0–1	-2	0.45	0.65	-2	-0.35	0.72	4	0.7	0.48	-4	0.5	0.62
Motivation to learn		0–100	-4	0.89	0.37	2	0.35	0.73	-3	-0.55	0.58	4	-0.59	0.56
Students in my class are set on learning as much as possible during classes		1–5	-2	0.34	0.73	1	0.1	0.92	-1	-0.11	0.91	4	-0.51	0.61
Place of residence		0–1	-5	0.95	0.34	-1	-0.12	0.91	8	1.32	0.19	-3	0.36	0.72
Socio-economic status	Very low	0–1	-15	2.82	0	8	1.21	0.23	8	1.19	0.23	-2	0.25	0.8
	Low	0–1	-3	0.58	0.56	4	0.55	0.58	-1	-0.17	0.87	2	-0.3	0.77
	Average	0–1	8	-1.54	0.12	-9	-1.31	0.19	-5	-0.76	0.45	2	-0.29	0.77
	High	0–1	6	-1.1	0.27	-10	-1.45	0.15	2	0.31	0.75	-5	0.67	0.5
Parents checked homework	Never	0–1	6	-1.17	0.24	7	1.08	0.28	-10	-1.5	0.13	8	-1.08	0.28
	Several times a year	0–1	9	-1.9	0.06	12	1.71	0.09	-2	-0.39	0.7	-1	0.1	0.92
	Several times a month	0–1	-5	1.06	0.29	6	0.88	0.38	2	0.39	0.7	-4	0.5	0.62
	At least once a week	0–1	-12	2.51	0.01	-1	-0.1	0.92	10	1.55	0.12	-5	0.73	0.47
	Several times a week	0–1	-1	0.11	0.91	-1	-0.14	0.89	4	0.69	0.49	-2	0.24	0.81
Similarity between problems solved at the exam and solved during classes		0–100	3	-0.7	0.48	1	0.15	0.88	-6	-0.87	0.38	-1	0.17	0.86
The way the classes are taught encourages me to be active		1–5	-8	1.73	0.08	0	0.06	0.95	6	0.9	0.37	9	-1.23	0.22

Table 4
General measures of quality of the data matching procedures

	Mean differences sum of squares	Student's *t*-test sum of squares	Significance levels sum of squares
Before matching	960	39	6
After one-to-many matching using *k*-means cluster analysis	860	21	6
After one-to-one matching using *k*-means cluster analysis	512	10	9
After stratifying and one-to-one matching based on Mahalanobis distance	965	20	9

Table 5
Comparison of mean results of pupils in experimental and control group using Student's t-test

Levene test of variance homogenity		Student's *t*-test				
F	*Sig.*	*T*	*Df*	*Sig.*	Mean difference	*SE*
0.490	0.483	0.563	778	0.573	0.630	1.119

between the experimental and the control group. In the case of matching using one-to-one *k*-means, 12 variables were equal. Otherwise, for 10 variables, the match was poorer. In order to present a general assessment of the matching procedures, a sum of squares was calculated for standardised differences in means for all the variables included in the match. The sum of the squares of the *t*-test values and significance levels were calculated in the same way. The best quality match presented in the table was obtained by *k*-means matching in the one-to-one variant.

Results and discussion

After one-to-one matching, lacking significant differences between the experimental and control group (with control of variables included in the matching procedure) the mean result for pupils in classes with 22 pupils or fewer was not significantly different from the mean result of lower secondary school exam in larger classes. The mean score of lower secondary school exam in small classes was 57.44 points (*SD* 15.34) and in large classes to 56.81 points (*SD* 15.94). For assessment of the significance of difference between means a Student's *t*-test was performed. The results of the test should be interpreted assuming equal variance. With a significance level of 0.573 the hypothesis that the mean score at the lower secondary school exam in small classes would be significantly different from the result in large classes should be rejected.

A parallel analysis of groups stratified using the Mahalanobis distance yielded similar results. Significance of *t*-test statistics amounted to 0.520. Consequently, class size did not demonstrate a statistically significant effect on the results of lower secondary school exams. Pupils in smaller classes obtained better results on average by 1.1 point (*SD* 0.068) than their peers from large classes. In the results for groups created using the one-to-many matching procedure using the *k*-means method, the mean result for pupils in small classes was 57.44 (*SD* 15.34) and in large classes 55.93 points (*SD* 15.79). The

Table 6
Small class effect by data matching procedure used

	Quality of the matching	ATT	Student's *t*-test sig.
After stratifying and one-to-one matching based on Mahalanobis distance	+	0.068	0.52
After one-to-many matching using *k*-means cluster analysis	++	0.093	0.13
After one-to-one matching using *k*-means cluster analysis	+++	0.039	0.57

Table 7
Small class effect by exam subject (maths-and-science and humanities)

After one-to-one matching using *k*-means cluster analysis	Scores difference	ATT	Student's *t*-test sig.
Maths-and-science part of the exam	0.69	0.069	0.66
Humanistic part of the exam	-0.06	-0.008	0.33

difference of 1.51 point was also statistically insignificant.

In order to assess the effect of class size on academic achievement (as measured by the whole-country lower secondary school exam) the difference between the means of exam results was calculated. The mean effect of the small class on units in the group, depending on the method of matching, was between 0.039 and 0.093 standard deviations. The mean effect of the treatment for all observations in the sample was 0.067 (calculated for "raw" data before matching) (Table 6).

The outcome variable is the sum of scores that a given student obtained from the humanities and maths-and-science. It is relevant to consider how the small class effect is influenced by subject taught. Again statistically insignificant, the result was observed for humanities and maths-and-science in small and large classes. Pupils from small classes obtained a result 0.69 points better in maths-and-science than their peers from large classes and on average scored 0.06 point lower in humanities. This confirms common sense presumptions that in a small class, pupils learn science more effectively than humanities (Table 7).

It is also relevant to consider distributions on the stanine (STAndard NINE) scale (often

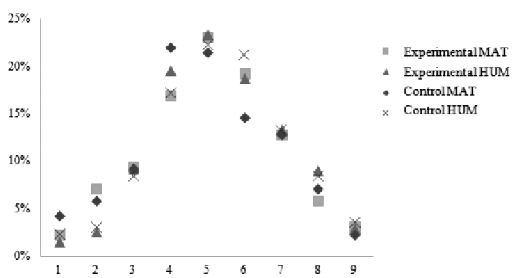

Figure 4. Share of pupils in a given stanine.

used for comparing academic achievements in Poland) from the experimental and control groups selected in *k*-means matching procedure in one-to-one variant (Figure 4). In the middle stanine five on the scale represents pupils who achieved average results at the exam. Pupils from small classes performed better in maths. In the fourth stanine there were 5% more pupils from large classes, while in the sixth there were 4% more from small classes. Taking into account the results of humanities, in the fourth stanine there were 2% more pupils from small classes, and in the sixth 3% more from large classes.

Conclusions

Based on the analyses, it can be observed that the effect of class size on academic achievement was not statistically significant in the sample investigated. However, at the lower secondary school exam, pupils from small classes scored better mean results than their peers from classes of over 22 by *SD* 0.039. This result, obtained using rigorous statistical procedures and control of contextual variables, conforms with the results of published research. It generally confirms a positive influence of small class on academic achievement but which is difficult to observe.

Many papers reporting such statistically insignificant results are unpublished but such evidence remains important as a balance to overestimation due to the resulting publication bias. In randomised experiments, considered the "gold standard" of research methods, the small class effect is strong and positive but many econometric studies as well as the ex post facto experiment reported here fail to confirm a positive effect.

Therefore, the obtained result should not be interpreted without interest in possible sources of bias. The greatest deficiency in this study was seen as lack of control for teacher effect, as already mentioned. Education of teachers, courses attended, personality, involvement with teaching, teaching methods, skills and experience are all contributing. Analyses were limited by the pool of variables available in the database. Confounding variables at the school level were also missing (e.g. available teaching aids, infrastructure). Both teachers and school variables were only indirectly controlled by application of matching procedures to pupils' place of residence. In all cases, place of residence coincided with location of the school. It was assumed here that location of the school is a good indicator of teacher and school influence, as schools in urban areas are associated with a higher standard of education and have better resources than rural schools.

The results obtained are valid only for this data and may have been compromised by data quality, experimental scheme used and lack of control over specific mechanisms to select pupils into small and large classes. As in the majority of cases these mechanisms are unique to a given school, their proper control would only be feasible in randomised experiment. Awareness of these limitations is preparatory to further work to assess the small class effect and logically therefore also important factors to understanding the education system and its development.

Literature

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

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

Biddle, B. J. and 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. and 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. and Weinfeld, F. D. (1966). *Equality of educational opportunity*. Washington: U.S. Government 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 and R., Purves, R. (1997). *Instructors' Manual for Statistics* (3rd ed.). Department of Statistics, University of California, Berkeley, New York: Norton.

Glass, G. V. and 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. and 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. and Filby, N. N. (1982). *School class size: research and policy*. Sage Publications.

Graue, E., Oen, D., Hatch, K., Rao, K. and Fadali, E. (2005). Perspectives on class size reduction. [Presentation on 12 April 2005 at *Early childhood policy in practice: the case of class size* Symposium, as part of annual meeting of the 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. In S. E. Mayer and P. E. Peterson (eds.), *Earning and learning: how schools matter* (pp. 131–168), Washington, DC: Brookings Institution.

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

Harris, J. R. (2000). *Geny czy wychowanie?* [*Genes or upbringing?*] Warszawa: Wydawnictwo Czarna Owca.

Hart, S., Petrill, S. and 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. and 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. and Stock, W. (1983). The effects of class size: an examination of rival hypotheses. *American Educational Research Journal*, *20*(1), 63–65.

Herbst, M. and Herczyński, J. (2005). *School choice and student achievement. Evidence from Poland*. Warsaw: Warsaw University.

Jakubowski, M. and 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. and 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. and Winship, C. (2007). *Counterfactuals and causal inference: methods and principles for social research*. Cambridge: Cambridge University Press.

Nye, B., Hedges, L. V. and 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. and 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. and 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. and Wittebols, J. H. (1986). *Class size research: a related cluster analysis for decision making*. Arlington, Virginia: Educational Research Service.

Rosenbaum, P. R. and 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* [*Economic determinants of results of six-graders' test and lower secondary school exams conducted in spring 2002*]. Expert opinion commissioned by the Ministry of Education.

Strawiński, P. (2008). Quasi-eksperymentalne metody ewaluacji [Quasi-experimental evaluation methods] In A. Haber (ed.), Środowisko *i warsztat ewaluacji* [*Evaluation environment and methods*] (pp. 193–220). Warszawa: Polska Agencja Rozwoju Przedsiębiorczości.

Strawiński, P. (2007). Przyczynowość, selekcja i endogeniczne oddziaływanie [Causality, selection, and endogeneic effect]. *Przegląd Statystyczny*, *4*, 49–61.