Pokropek *Large-scale Assess Educ (2016) 4:4*
DOI 10.1186/s40536-016-0018-2

Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

Introduction to instrumental variables and their application to large-scale assessment data

Artur Pokropek*

*Correspondence:
artur.pokropek@gmail.com
Department of Philosophy and Sociology, Polish Academy of Sciences, Nowy Świat 72, 00-330 Warsaw, Poland

Abstract

In the social sciences, estimating causal effects is particularly difficult. Gold standards are set by randomized experiments in many cases expensive, unenforceable for ethical and practical reasons. Recent research has drawn attention to techniques that under some conditions, could estimate causal effects on non-experimental observable data. One technique is the instrumental-variables (IVs) approach. This approach is used to determine variation that is exogenous in treatment and to estimate causal inferences. This paper begins by explaining the logic of IVs and then reviews the literature on the use of the IVs approach in the educational context. The most common types of IVs and the guidelines for selecting appropriate variables are explained. The statistical background of IVs estimation is described, which is followed by a discussion of the assumptions that underlie statistical procedures. Finally, empirical examples that use data from the Polish extension of the Programme for International Student Assessment are presented to estimate the effects on student learning outcomes of having at least one neighborhood friend in the classroom.

Keywords: Instrumental variables, Large-scale assessment, PISA, Friendship, Causal inference

Causal inference

The majority of questions posed in the natural, physical as well in social sciences are causal in nature. We want to know whether one event is a consequence of another event or whether the treatment influences the outcome. We ask whether the drug really fights the disease, if policy programs actually reduce inequalities, and whether class size influences student achievement. Exploring causal relationships gives us the opportunity to understand our world and the tools for promoting effective change.

In the social sciences, making causal inferences about mindful objects that are responsive and impossible to control is difficult, but it is possible. The gold standard of making compelling causal inferences depends on experimental designs in which the assignment of participants to treatments is "exogenous" rather than "endogenous." Exogenous is defined as related to external causes, whereas endogenous is defined as related to internal causes. Endogenous assignment of participants to treatments simply means that the probability of a particular treatment is related to the outcome variable or variables

related to the outcome. If so, effects of the treatment could not be disentangle from effects of other variables and estimated treatment effects are likely to be biased. By exogenous variation in treatments we mean that the assignment to the treatment is not affected by any factor inside the analyzed system, that is, no important variable, from the perspective of the research question, influences the assignment. In this situation identification of treatment effects is possible because assignment is not related with the outcome variable.

Another possible opportunity for making causal inferences is in natural experiments, where assignment to the treatment is not generated by a purely random mechanism but is exogenous to the system of interest. In natural experiments, some external peculiarity, natural disaster, geography, or organizational irregularity assigns different people to different treatments or conditions.

We believe that in both randomized experiments and natural experiments (sometimes controlled for by observed variables), on average, people in different treatments or conditions induced by exogenous factors are similar on all variables of interest. That is, they only differ by the kinds of treatments or conditions. If this assumption is correct, the differences in outcomes between groups might reasonably be attributed to the causal effects of the treatment or conditions.

Both experimental studies and natural experiments could provide the required exogenous variation in treatment, which is necessary for making causal inferences. However, constructing experimental studies is expensive, difficult, and in many cases unenforceable because of ethical and practical reasons. Some experimental research can also be criticized because it is not always robust to factors that can potentially jeopardize internal or external validity like subject attrition, noncompliance, external events, maturation of subjects, diffusion of treatments, or overexposure to testing instruments (Campbell et al. 1963, pp. 13–20). Finding the appropriate natural experimental conditions for a particular problem could be troublesome and unsuccessful in many situations because circumstances that generate exogenous variation like certain institutional conditions, random events including natural disasters, and other situations that could bring randomness into investigated system are rather rare. There are also potential threats to the validity of quasi-experiments if a relationship between the random event and the outcome exists or if some subjects successfully counteract a random event (Murnane and Willett 2010, pp. 153–154).

With readily available data and the relative ease of collection over experimental designs, some social scientists in education use observational data, which include large-scale assessment studies, such as the Programme for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMSS), and the Programme for the International Assessment of Adult Competencies (PIAAC), as an attractive and viable alternative to experimental or quasi-experimental studies.

The main disadvantage of observational studies is that they usually lack sufficient evidence that the conditions of interest are exogenous to the investigated system. For instance, in examining observational data using simple methods like correlations or regression analysis, we are not able to credibly recognize whether schooling influences outcomes, such as earnings. Correlations between such a variables can never suggest a

causal relationship, they are only a necessary condition for causal relationships. We cannot exclude situations in which some unobservable variables are at the heart of correlations of the observed variables. For instance, without information about respondents' IQ we could not exclude the possibility that people with higher IQs spend more time in school than do others and that at the same time IQ is directly related to earnings. In this example the observed correlation between schooling and earning could emerge because of effects of IQ on both observational variables and not a causal relationship between the observed variables.

The problem of causal inference in observational studies is illustrated in Fig. 1. An unbiased estimate of causal effects using observational data is not possible when the predictor variable X is correlated with unobservable variables accumulated in U, which in regression analysis is described as the error term. If we utilize the example of the effect of schooling on subsequent earnings, the estimate of causal effects is problematic because other factors might affect (or at least are correlated with) both schooling (X) and earnings (Y), which makes an unbiased estimate of the relationship between X and Y impossible (Fig. 1a). The unbiased estimate of causal effects is only possible if there is no relationship between the predictor variable and other unobservable factors, as depicted in Fig. 1b. Only when the predictors and residuals are uncorrelated in the population, will ordinary last squares (OLS) regression give unbiased results. The strategy used to achieve this condition is to add additional exogenous covariates to the statistical model, under the condition that in the covariates in the population, there is no association between the predictor and the error term. Adding a set of convening covariates, such as IQ, social background, motivation, character, temperament traits, and others related to the earnings covariates might result in the estimation of the unbiased conditional causal relationship between schooling and earnings. This result will only materialize if all important variables (for the equation) are included. If not, the estimates will be confounded by omitted variables. Such a situation is also often called omitted variables problem or omitted variables bias (Wooldridge 2010, pp. 61–67). In many situations the omitted variables problem might be also interpreted as selection or a self-selection problem (Heckman 1979). In this interpretation bias occurs when assignment to the treatment or independent variable is endogenous and correlated with the selection mechanism. If the model lacks variables describing the existing selection process estimates of the parameters will be biased (or alternatively speaking, affected by omitted variables that describes selection process).

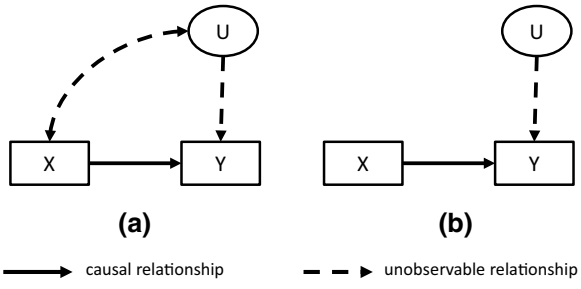

Fig. 1 Examples of a situation where the modeling of causal relationships using observational data will be biased (**a**) and a situation where it will be valid (**b**)

In most situations, social scientists are not equipped with all the necessary covariates to successfully model the conditional relationships between variables of interest. In such situations the instrumental variables (IVs) technique could be a helpful alternative. By using IVs, we can try to determine variation that is exogenous (i.e., related only to external causes) in treatment and use it in an estimate of causal effects.

What are instrumental variables?

The IVs technique is a statistical tool that could be applied to experimental data that fail to fulfill all assumptions necessary for unbiased inference,[1] data from natural experiments, and under several conditions, to observational data. Most common application of instrumental variables refers to quasi-experiments, but other applications of instrumental variables might be also found. In this paper, we focus on instruments that could be found in observational data.

In the IVs technique, variables called instruments are used to determine an exogenous part of the variability from the endogenous predictor. In other words this technique allows the use of only that part of the variation in the predictor that is "arguably random" i.e., is not related with unobservable factors affecting both predictor and outcome. Such a procedure allows researchers to effectively estimate the causal relationship between the outcome and the predictor.

Specifically, an instrumental variable Z is an additional variable used to estimate the causal effect of variable X on Y. *The traditional definition qualifies a variable Z as an instrumental (relative to the pair (X, Y)) if (i) Z is independent of all variables (including error terms) that have an influence on Y that is not mediated by X and (ii) Z is not independent of X* (Pearl 2000, p. 247). Therefore, the instrumental variable Z affects Y only through its effect on X. Consequently, variable Z is unrelated to the outcome (Y) but is related to the predictor (X) and is not causally affected (directly or indirectly) by X, Y, or the error term U. In this approach, not only one but also multiple IVs and/or causal paths could be used.

Figure 2 shows three situations with potential IVs. Y and X are variables between which a causal relationship is to be estimated. Z is the potential instrumental variable, and the error term U stands for all factors that affect Y when X is held constant. In the first scenario (a) the variable Z is not a valid instrument because the instrument should be independent of U. The relationship between U and Z allows for an indirect association between Z and Y, which biases the estimation of the causal relationship between X and Y. In the second scenario (b) the so-called "exclusion restriction" is violated. That is, the instrument Z should not affect Y when X is held constant (Pearl 2009, p. 123). Only the third scenario (c) exemplifies a valid instrument for X. In this scenario, Z affects Y only indirectly through X and is not related to the unobservable variables accommodated in U.

In this example, the weakness of the IVs approach is that three of the situations shown in Fig. 2 are not empirically distinguishable. The use of observable variables does not allow for testing whether the chosen Z is a valid instrument. The first scenario is clear. It is not possible to test whether a relationship between Z and U exists because U is not

[1] For instance Angrist et al. (1996) used IV technique for randomized experiment where compliance with the assignment was not perfect so that the receipt of treatment was nonignorable.

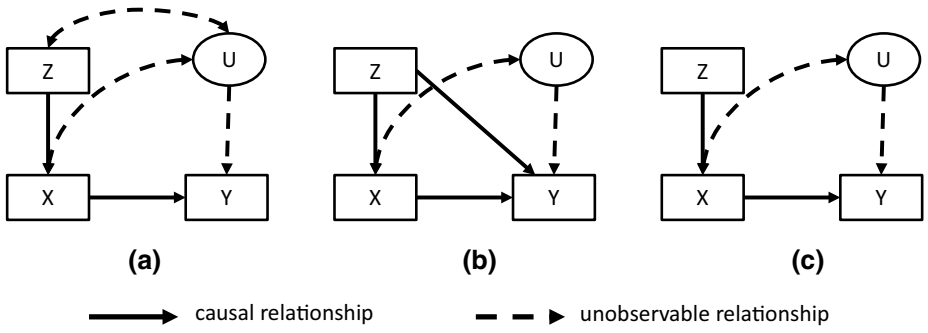

Fig. 2 Examples of situations where Z is an invalid instrument (**a**, **b**) and a valid instrument (**c**)

observable. In other words, the researcher never knows for sure whether the instrument that he or she used is correlated with unobserved variables that are causally related to the outcome.

The second situation is less obvious. It might appear that if an empirical association between the instrument and the outcome is detected after conditioning the predictor, then Z is an invalid instrument. However, this is not true. An empirical relationship might appear even though the IV identifying assumption is valid (Morgan and Winship 2014, p. 197) and there is no other empirical test that could reject improper instrument. The researcher must rely on current knowledge, theories, and intuition in deciding whether an instrument is valid. Only if the researcher is able to argue successfully that the modeling situation reflects the situation depicted in Fig. 2c, and not (a) or (b), then the validity of the instrument could be assumed.

Instrumental variables and causality in the educational context

Examples are provided to clarify the concept of instrumental variables. In the context of educational research, Murnane and Willett (2010) identified the three most popular sources of potential research instruments: (1) variables that describe the participants' proximity to relevant educational institutions; (2) institutional rules and personal characteristics; and (3) deviations from cohort trends.

As previously mentioned, some authors claim that that the very best instrument arises in a randomized experiment—namely the randomization mechanism (Angrist and Krueger 2001). However, the distinction between quasi-experimental situations and the sources categorized by Murnane and Willett (2010) seems blurred. In this study, we use this categorization as a useful tool for understanding and identifying potential instruments. Table 1 provides examples of instruments used in studies that analyzed several issues connected with education (Murnane and Willett 2010).

Institutional rules and personal characteristics

The variables of institutional rules and personal characteristics are the most often studied in the literature on the use of IVs. In the context of education, the popularity of this type of instrument emerged after the landmark study by Angrist and Krueger (1991). They studied the effects of educational attainment on the weekly earnings of males born in 1930s and 1940s in the United States. As previously discussed in this paper, the

Table 1 Examples of IVs in studies connected with education (sorted by year in categories)

Study	Outcome variable(s)	Explanatory variable	Instrumental variable
Institutional rules and personal characteristics			
Angrist and Krueger (1991)	Earnings	Education	Quarter of birth
Angrist and Lavy (1999)	Achievement test scores	Class size	Discontinuities in class size
Currie and Yelowitz (2000)	School quality, grade repetition (among others)	Public housing project participation	Sex composition in household
Angrist and Lavy (2002)	Achievement test scores	Use of computers	Funded program
Weber and Puhani (2006)	PIRLS scores among others	Age	Assigned relative age
Bedard and Dhuey (2006)	TIMSS scores	Age	Assigned relative age
Lee and Fish (2010)	TIMSS and NAEP scores	Age and grade	Assigned relative age and grade
Hanushek et al. (2015)	Earnings	Cognitive skills: literacy, numeracy, problem solving	Minimal school-leaving age
Deviations from cohort trends			
Hoxby (2000)	Achievement test scores	Class size	Cohort composition "surprise"
Hoxby (2002)	Achievement test scores	Classroom composition	Cohort composition "surprise"
Wößmann and West (2006)	TIMSS scores	Class size	School's average class size controlling for fixed effects
Proximity to relevant educational institutions			
Rouse (1995)	Educational attainment	Community colleges	Distance to school
Neal (1997)	Achievement test scores	Catholic school	Distance to school
Dee (2004)	Civic participation	College	Distance to college

estimation of the causal effects of this relationship might be biased by the endogenous character of the predictor (i.e., years of schooling). Therefore, they used the IVs technique. The authors noted that in most American schools, students born early in the year start school at a later age. This happens because students enter the first grade after they reach the age of six by the end of the calendar year in which they enter the school. Students entering school later can leave school after a shorter period of learning because compulsory schooling depends on the age of the student (16 or 17, according to the state). Hence, males born early in the year receive less schooling than males born later in the year do.[2] This situation reflects a good exogenous instrument in which schooling is correlated to institutional rules. It has no causal effects on earnings because the month of birth is random (Angrist and Krueger 2001, p. 74); therefore, there is no convincing, direct causal path between the month of birth and earnings. Therefore, Angrist and Kruger used the quarter of the birth year for estimating the unbiased effects of schooling on earnings. Interestingly, the effect did not differ significantly from estimates based on classical OLS models showing small but significant effects of schooling on earnings.

The date of birth was also used as an IV for estimating the effect of age on cognitive proficiency in large-scale assessment data showing significant positive effects on

[2] It was true for men born in 1930s and 1940s. In more recent cohorts in the USA, the vast majority of students are not quitting school at 16 or 17. Potential reanalysis for more recent USA data would be consequently problematic however it would be interesting to see such an analysis for other cultures.

educational outcomes, such as PIRLS (Weber and Puhani 2006), TIMSS (Bedard and Dhuey 2006; Lee and Fish 2010), and NAEP (Lee and Fish 2010). The estimation of the causal path between age and cognitive proficiency is problematic because of several factors, such as grade retention, delayed entry, and entry grade acceleration. In such situations, the assigned relative age (i.e., birth month relative to the school cut-off date) or the assigned grade (i.e., grade in which the students would be expected to be enrolled, based on their birth date relative to the school's cut-off date) might be used as an IV for the observed age. On one hand, because most students enter school on time and are never retained, the relative age is correlated with the actual age. On the other hand, the random fact that a student was born closer to or further from the school's cut-off date should not affect his or her proficiency. There are no explanations for why students born at different times of the year are not more or less smart than others.

Birth dates and institutional rules (among others) were also used recently in other large-scale assessment studies. Using PIAAC data, Hanushek et al. (2015) used the IVs approach in a study of the returns to individual cognitive skills in the labor market. In estimating the relationship between cognitive skills and earnings, Hanushek et al. used two sets of instrumental variables. The first set, defined for all PIAAC countries, contained two variables: years of schooling and parental education. They argued that these variables could be used as instruments because they influenced skill development but were determined before the individual entered the labor market. However, the authors were skeptical about these instruments, pointing out that *family background may exert direct effects on earnings, and ability may show intergenerational persistence* (Hanushek et al. 2015, p. 119). Therefore, this set is a good example of the utilization of bad instruments, where assumptions necessary for IVs are not likely to hold. In such a situation causal inference is likely to be biased. What's more is that the direction of the bias is not easy to predict. The second approach is both interesting and credible. Because of the limitations of the data availability, the second approach was used in the US sample only. Hanushek et al. utilized the fact that different US states have changed their compulsory schooling requirements at different times, including the minimal school-leaving age. This strategy was also used in other contexts (e.g., Acemoglu and Angrist 2000). The individual minimum school-leaving age was used as an instrument for skills. Because this information was related with skill level but not earnings it seemed to provide a good instrument that confirms a positive relationship between cognitive skills and income.

Another interesting use of IVs that emerged from the variable of institutional rules is a study by Angrist and Lavy (1999). Estimating the causal effect of class size on scholastic achievement could be complex because classroom size is correlated with many unobservable characteristics, such as popularity of the school, quality of teaching, quality of management, available resources, etc. The use of this instrument is based on the bureaucratic rules that set the maximum number of students in a classroom. In this scenario, the instrument is the variable that concerns whether the school must create additional classrooms in order to avoid the maximum number of students in a classroom. This variable is negatively correlated with class size, but it is hard to anticipate the causal relationship between the instrument and student achievement as all schools were obligated to follow the rule regardless of size, quality of teaching, or location. The IVs estimates show that reducing class size induces a significant and substantial increase in test scores.

Currie and Yelowitz (2000) introduced a very interesting application of IV, which was based on institutional rules. The aim of their research was to compare public housing residents with households having similar characteristics on measures of satisfaction with housing neighborhood, children's scholastic achievements, school ratings, extra-curricular activities, and grade retention. Estimates using the OLS regression can be biased because the participants in the program might have unmeasured traits that contribute to poor housing outcomes and poor academic performance of their children. To overcome this problem, the authors used an IVs technique that was based on the institutional rules that were mandatory for the participants. According to these rules, a family with two same-sex children would be eligible for a two-bedroom apartment, whereas a family with opposite-sex children would be eligible for a three-bedroom apartment. Consequently, households with one boy and one girl are far more likely to be in public housing than are households with two boys or two girls because sex composition affects the size of the subsidy for which the family is eligible. The gender composition of the family appears to be a valid instrument because it correlates with the participation in program. However, as claimed by the authors, who reviewed the literature that investigated the relationship between sex composition and educational attainment, gender composition showed no causal links with outcomes. Using the IVs technique, Currie and Yelowitz (2000) found that the participants in the public housing program lived in better material conditions, and their children were less likely to be left behind than households that did not participate in the program, which was not consistent with probably biased OLS estimates.

The last example of an instrumental variable based on institutional rules refers to the work of Angrist and Lavy (2002). This study was subsequently replicated by Machin et al. (2006), who used different data. The IVs technique was applied to a situation similar to experimental settings. In their analysis, Angrist and Lavy aimed to estimate the effect of using technology in teaching on student achievement. The outcome variables were student scores on tests, and the predictor variable was based on a teachers' survey about the use of technology in the classroom. Although simple OLS estimates might not yield a causal effect because the scholastic achievements of students and the use of computer technology might correlate with unobservable factors (i.e., good teachers use technologies, which might not be causally related to the outcomes). The authors decided to use the IVs technique, which was possible because the Israeli state lottery funded a large-scale computerization program in elementary and middle schools. Participation in the program was not random but defined by a set of rules that allowed program participation. The authors claimed that the participation rules were not systematically associated with the student outcomes. This situation created an instrument that was not causally connected with the outcome, but it did relate to the usage of computer technologies. By using this instrument, the authors estimated that the causal effect of computer usage on student achievements did not significantly differ from zero.

Deviations from cohort trends

Another source of instrumental variables is the deviation from cohort trends. This type of instrument was first used by Hoxby (2002) to investigate the causal effect of class size and class composition on student achievement. As in Angrist and Lavy's (1999) study,

class size and composition were the most likely endogenous variables. As a consequence, students in classes of different sizes and compositions might differ in unobserved ways that affect students' test performance. Hoxby's (2002, p. 59) strategy for overcoming this problem was to use the "cohort composition surprise" as an instrument because year-to-year random fluctuations in local births can appreciably alter both class sizes and class compositions. If a cohort is larger than the previous cohort, the school must allocate the "extra" students to its classrooms. Similarly, if there are more females in the cohort than expected, some students in the cohort will have a peer group that has more females than is typical. Thus, "surprises" constitute an intriguing instrument. On one hand, because "surprises" are unexpected, they should not influence parents' and administrators' decisions that might be connected with student achievement and have no causal relationship to achievement. On the other hand, such surprises are correlated with class size and composition. As the IV, Hoxby used predicting models for annual enrollment and the prediction error for a given school in a given year. The results showed that class size did not have a statistically significant effect on student achievement; however, class composition did. For instance, both boys and girls performed better in reading when they were in classes that had larger proportions of girls.

Estimating the effect of class size was also explored by Wößmann and West (2006), who used TIMMS data. They used school fixed effects[3] to account for between-school sorting, and IVs for within-school sorting and estimating causal effects. As the instrument for the class size, they used the average class size at different grade levels in a particular school. Because they controlled for the fixed effect, their instrument deviated from the predicted size of the classrooms forecasted from different grades, which is similar in principle to Hoxby's idea. This IV was highly correlated with actual class size, but after controlling for school fixed effects, there was little evidence that grade and average class size would affect student performance.

Proximity to relevant educational institutions

The third potential source of IVs is based on economic theory, which stipulates that, holding other things constant, the lower the cost of enrollment, the higher the ensuing attainment (Murnane and Willett 2010, p. 263). Here we need to assume that people are rational actors that calculate their costs and potential benefits. Costs might be defined by several variables, such as money, effort, and time spent commuting. However, in the context of educational research, the distance between the individual and educational institution has been used the most often as an instrumental variable.

Rouse (1995) used distance from the respondent's high school to the nearest junior college to estimate the effect of community colleges on educational attainment. Neal (1997) used proxies for geographic proximity to Catholic schools as an exogenous source of variation in attendance at these high schools by estimating its effect on test scores. Dee (2004) used the distance to college as the IV to determine the effect of college attainment on civic participation.

From a methodological point of view, these three studies are very similar. Let us focus on Dee's work as a representative example. Dee wanted to test whether college

[3] Fixed effects regression holds constant (fixes) the average effects of each school.

attainment affects civic participation. The main problem with his goal is that the explanatory variable—educational attainment—is not endogenous. Participants have selected their own levels of educational attainment rather than having it assigned randomly. Dee's causal estimation strategy consisted of using the distance to the college as an IV. In utilizing the IVs technique, it must be assumed that there is no causal relationship between distance to college and civic participation and that no unobservable factors are related to both civic participation and geographic placement of the colleges. Hence, the participants' geographic placement around the colleges must be distributed exogenously. This assumption was apparently too strong to hold. For instance civic minded parents might live in more developed areas with larger number of colleges. However, Dee's strategy was to include many control predictors at the individual, family, community, and county levels, such as age, gender, race, religious affiliation, prior academic achievements, parental education and income, as well as series of community level variables. After controlling for this rich set of control variables, the assumption that conditional geographic placement around the colleges was distributed exogenously appeared convincing. By using the IVs technique and a rich set of covariates, Dee showed that educational attainment had large and statistically significant effects on civic participation.

How does it work?

Classical regression analysis uses the OLS estimation relationship between the predictor and the outcome. The relationship is estimated using the covariance between the predictor and the outcome σ_{XY} and the variance of the predictor σ^2_X, that is, $\beta_{YX} = \sigma_{XY}/\sigma^2_{XY}$. In Fig. 3, the construction of the parameters are graphically represented by Venn diagrams. The construction of the parameter β_{YX}, which represents the association between X and Y, is presented on the left side of the graph. This parameter is simply computed as the ratio of the shared part of the variance of the variables X and Y to the total variance of the predictor, which yields the classical OLS linear estimate of the relationship that under the condition of exogeneity might be interpreted in terms of causality.

When the predictor is endogenous, but an exogenous instrument is found, the IVs technique could be applied. In the IVs estimation, the estimate is based only on the part of variance that we argue is exogenous, that is, as identified by our instrument. Therefore, to estimate the relationship between X and Y, we use only the part of the variation that is common to our variables of interest and the instrument. This is depicted on the right side of Fig. 2. In the IVs approach, β_{YX} is estimated as the ratio of the shared part of variance bounded by the variation of the instrument ($\sigma_{XY|Z}$) to the variance of the predictor that is shared with an instrument ($\sigma^2_{X|Z}$). The IVs estimate is similar to the OLS but is restricted by the variation of the instrument.

Conceptually, IVs estimation might be expressed as two independent stages of OLS estimations. The first stage is designed to "clean up" the endogenous predictor, leaving only the exogenous part of the variation that covaries with the instrument. This is achieved by regressing the endogenous predictor on the IVs and additional covariates:

$$\text{Stage 1:} \quad \mathbf{x}_i = \mathbf{I}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\upsilon} + \boldsymbol{\delta}_i \tag{1}$$

where $\mathbf{x}_i$ is a vector of the endogenous predictor i (where $i = 1,\ldots, N$-predictors), $\mathbf{I}$ is the design matrix for IVs, $\boldsymbol{\alpha}$ is the vector of slope parameters for IVs, $\mathbf{Z}$ is the design matrix

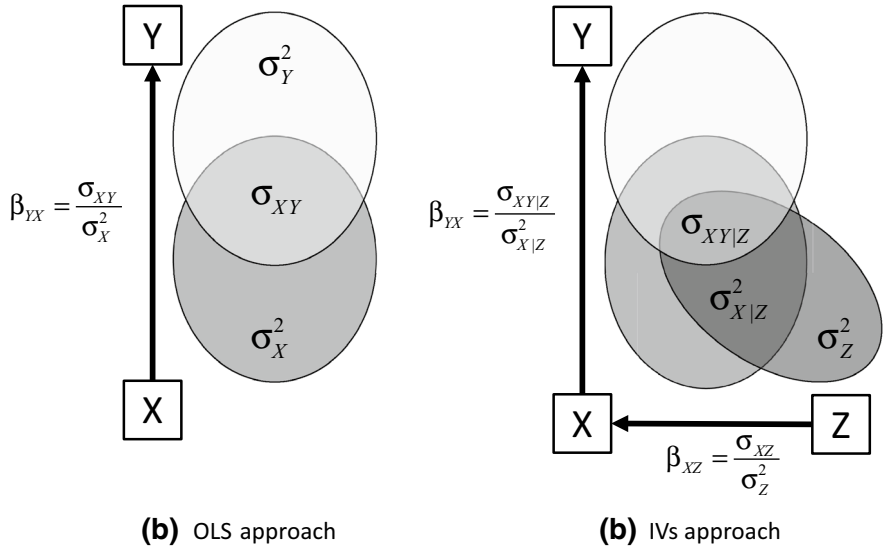

Fig. 3 Venn diagrams representing the OLS and IVs approaches used to estimate causal relationships between two variables (Murnane and Willett 2010, pp. 213, 232)

for the covariates, $\mathbf{v}$ is the vector of slope parameters for the covariates, and $\boldsymbol{\delta}_i$ is the error term.

In Stage 1, the model instruments and covariates must be exogenous (not correlated with first stage residuals). However, the covariates do not have to comply with the assumption that there is only an indirect influence on the outcome variable. In the estimation, the instruments and covariates are treated exactly the same. The difference is caused by the assumptions and treatment of these variables in Stage 2. The role of the instruments finishes at Stage 1 of 2SLS (two-stage least squares, see below). The estimation covariates should be added to the equations in Stage 2 as predictors because they are directly related to the outcome, adding them prevents the omitted variables problem.

The inclusion of covariates in the Stage 1 model helps to fulfill the assumption that there is no direct relationship of the instrument to the outcome. For instance, in the case of instruments based on the proximity to educational institutions, the distance to the institution could be considered a valid instrument only after controlling for observable exogenous characteristics and ruling out inequalities that might not cause the random allocation of institutions, as in Dee's (2004) work on civic participation.

Stage 1 of the model could include as many instruments as possible in order to isolate as much exogenous variation as possible from the endogenous predictor and to increase the precision of the estimates. However, this guideline refers only to strong instruments. If the instruments are weak, that is, the instruments are weakly correlated with the endogenous predictors, it might be beneficial to use only the strongest instruments. Weak instruments might introduce a serious bias into estimates, particularly when the number of instruments is large (Bound et al. 1995; Staiger and Stock 1997).

In Stage 1, the model might include not only one instrumented predictor and the instrument(s) referring to it but a set of instruments and instrumented predictors,

depending on the formulation required in Stage 2. However, for clarity, one endogenous instrumented predictor is formulated in Stage 2 in this example:

$$\text{Stage 2:} \quad \mathbf{y} = \hat{\mathbf{x}}_i\beta_i + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e} \tag{2}$$

where $\mathbf{y}$ is the vector associated with the outcome variable, $\hat{\mathbf{x}}_i$ is the vector of predicted values of $\hat{\mathbf{x}}$ based on the first stage estimation, β_i is the parameter reflecting the causal effect from X to Y, $\mathbf{Z}$ is the design matrix for the covariates also used in the Stage 1 estimation, $\boldsymbol{\beta}$ is the vector of slope parameters for the covariates from $\mathbf{Z}$, and $\mathbf{e}$ is the error term.

The Stage 2 equation is simply the OLS regression, where the potential endogenous predictor $\mathbf{x}_i$ is replaced by predictions estimated as $\hat{\mathbf{x}}_i$ in the Stage 1 model. Sets of covariates ($\mathbf{Z}$) on both stages of estimation should be the same. On the one hand covariates not included in the Stage 1 but included in Stage 2 are likely to be correlated with Stage 1 residuals, which will bias all estimates in Stage 2. On the other hand covariates not included in the Stage 2 but included in Stage 1 might bring omitted variables problem to Stage 2 estimation (Baltagi 2002, p. 277).

Operationally, estimates for the IVs regression are not obtained by the two OLS regressions in Stages 1 and 2 because this procedure results in the biased estimation of standard errors. Instead, statistical software (e.g., Stata ivregress and ivreg2; SAS PROC SYSLIN) providing several methods of estimation that can be used in the IVs estimation. The most often used estimators are the following: two-stage least squares (2SLS),[4] limited-information maximum likelihood (LIML), generalized method of moments (GMM), and the Bayesian approach (Kleibergen and Zivot 2003).

Asymptotically, all listed estimators have the same properties. They are asymptotically unbiased (i.e., consistent), that is, the bias is zero when the sample size is very large. Because a portion of the variation used in estimating of causal effect in IV is substantially smaller than in OLS, only the exogenous part of the variation identified by the instrument is used to determine the precision of the estimators. Consequently, in the IVs approach, standard errors are larger than those generated in OLS estimations.

Interpretation of IVs estimates

IVs estimates do not provide the average treatment effect (ATE), which could be interpreted as the expected average causal effect of the treatment or condition. However, they do provide the local average treatment effect (LATE), which means that the IVs technique estimates the average causal effect on those affected by the instrument (Kleibergen and Zivot 2003, pp. 173–188). The LATE therefore is informative about subjects "who will take the treatment if assigned to the treatment group, but otherwise not take the treatment" (Angrist and Krueger 2001, p. 77).

For example, in Angrist and Krueger's (1991) study on returns to education, the estimated effect of schooling was informative for those deciding to quit school when they have the opportunity to do so. Research that uses the "surprise" cohort composition (Hoxby 2000, 2002) found effects that were informative for schools that were affected by

[4] The name of the 2SLS estimator is derived by the fact that the same point estimates could be obtained using OLS regressions in two steps. However 2SLS estimator is conducted in one stage and only with this procedure it provides correct standard errors.

unexpected changes in cohort size but not for schools that would have strict regulations on class size and/or student allocation. In studies that used proximity to relevant educational institution, the LATE effect was only informative for participants whose decision to enroll was influenced by the distance to the institution, but it was not informative for those that were insensitive to distance while deciding whether to enroll.

Regarding the validity of an IV, the question is whether estimates based on the participants' responses to the instrument might be generalized to the entire population. Several questions could be posed to address this issue. The first question concerns whether there is reason to believe that the effects estimated using IVs might be heterogeneous, that is, differ among the groups of respondents. If the effects are assumed to be homogenous, LATE equals ATE, and generalizations based on IVs estimations are legitimate. If the heterogeneity of the effect could not be clearly rejected, additional questions could be asked. How large is the group that is affected by the instrument? If the group covers the majority of the population, the problem of generalizability might not to be important. Another question about representatives might be asked. If there is no reason to believe that the individuals affected by the instrument are substantially different from the entire population (i.e., biased the sensitivity to the instrument), the estimated LATE effect might be considered valid for the entire population.

Empirical example: friends in classroom

This section presents an example of the IVs technique using large-scale assessment data. We focus on the question of whether having neighborhood friends in the same classroom might affect learning. In this paper, this effect is termed the "friendship effect."

Most previous studies showed that children who have friends perform better at school than those who do not have friends (Bandura et al. 1996; Frentz et al. 1991; Wentzel and Asher 1995). The results showed that effects of friends were mostly indirect but substantial. These effects were shown to determine motivation (Nelson and DeBacker 2008) and behavior (Wentzel and Caldwell 1997; Wilkinson et al. 2000), thus leading to positive learning outcomes.

In this paper, we refer to Polish data and students in the first grade of upper-secondary school (grade 10). In Poland, most students finish non-selective lower-secondary school at 16 years of age and then must choose an upper-secondary school. The upper-secondary school system is selective, and students are streamed into different courses of study. The move from lower-secondary to upper-secondary school is accompanied by a drastic change in the learning environment, including classroom peers, teachers, and the location of the school. It seems rational to assume that having friends in the new classroom might be beneficial for the adaptation process and that friends can indirectly affect learning outcomes. In other words, we want to check whether learning gains of students are determined by the fact of having neighborhood friends in the same classroom.

Data

The data used in the present analysis were collected from a national extension of the PISA (2009) study. In Poland, parallel to the international study, an independent sample of first-grade students in upper-secondary schools (grade 10) was selected and transformed into a panel study, which is currently named From School to Work (FS2W)

(Domański et al. 2012). A two-stage stratified sample was employed. In the first stage and the second stage, the school and the class (grade 10), respectively, were selected at random with equal probabilities. In March 2009, all cognitive and questionnaire instruments used in the international version of the study (OECD 2012) were applied. Six months later, in October 2009, the second wave of the study was conducted, which focused on several psychological measures. In April 2010, using instruments from PISA (2009), the cognitive measurements were repeated and an additional survey questionnaire was applied. In the first, second, and third waves, 4951, 4041, and 3989 participants, respectively, took part in the survey. The same 3472 students participated in all three rounds.

The assessment component of the PISA survey evaluates the students' ability to apply their knowledge and skills to real-life situations. It covers three main domains that are part of all PISA cycles: reading, mathematics, and science. The PISA (2009) survey focused on reading (131 items), mathematics (35 items), and science (53 items), in addition to minor areas of assessment. The test items were a mixture of open-ended questions that required the students to construct their own responses and multiple-choice items (OECD 2012). PISA uses a balanced incomplete block test design, in which students answer a sample instead of all test questions. Hence, in this study, in two cognitive measurements, each student answered different, but linked, sets of questions.

The data analysis performed in this study used the PISA scores of reading, mathematics, and science measured in 2009 and 2010, in addition to the PISA Index of Economic Social and Cultural Status (ESCS; OECD 2012). The indicator of friends in the classroom was based on the students' responses to the question, "Whether friends from your neighborhood learn with you in the classroom," which was asked in the second wave of the panel. The distance to school was expressed in the number of minutes required to travel from home to school, which was reported by each participant. Gender of participants and type of school (General comprehensive, Vocational with comprehensive program, and Basic vocational school) were used as covariates. The participants also indicated whether they lived in a village or in a city, which was used in all models.

Missing data were imputed using stochastic regression imputation (Enders 2010, pp. 46–49). For missing data imputation all variables used in modeling phase were used. In all analysis sampling weights were used and robust standard errors were computed accounting for clustering of students in classrooms. In all analyses one plausible was used. Although this is not an optimal use of the data and using only one plausible value produces estimates that do not contain between-imputation variance (Rutkowski et al. 2010), it is sufficient as a pedagogical example. Such treatment will results in underestimation of standard errors but was chosen because plausible values and multiple imputation techniques are not yet well established in context of IVs estimation and statistical tests applied routinely in IVs analysis.

OLS results

In this section the OLS estimates of the friendship effect are presented. Later in the paper we will compare these results with results obtained with IVs estimation. In the OLS approach the problem of the endogenous independent variable "friends" is inadequately solved by specifying multiple regression models with covariates that are

assumed to capture all potential selection (and omitted variables) problems. The OLS model might be expressed as:

$$\begin{aligned} \text{PISA2010} = \beta_0 + \beta_1 \text{FRIENDS} + \beta_2 \text{PISA2009} + \beta_3 \text{ESCS} + \\ \beta_4 \text{COUNT} + \beta_5 \text{VOC} + \beta_6 \text{BVOC} + \beta_7 \text{FEMALE} + e \end{aligned} \quad (3)$$

In the presented models, the outcome variables were defined as PISA scores in 2010 (PISA2010). Three models were estimated separately for reading, mathematics and sciences. The predictors were as follows: friends in classroom, PISA score in 2009 (PISA2009), economic social and cultural status of family (ESCS), living in the country (COUNT), type of school (VOC—vocational with comprehensive program; BVOC—basic vocational school) and gender (FEMALE).

Regarding the outcome variable, the PISA score in 2010 might be interpreted as the post-test. The first predictor was used to identify the causal effect of friends. PISA scores in 2009 served as the pre-test. As we wanted to estimate the effect of having friends on learning gains, prior scores were used in estimation. Without prior scores the effect of having friends on cognitive proficiency (not learning gains) would be estimated. Although this is an interesting problem it is a separate question.

Optimally, we would prefer that the pre-test be conducted before the beginning of the school year in Poland (i.e., September 2009) but such a measure was not available. However, available measurement allowed us to assess changes in proficiency over 1 year of learning between the middle of 10th and 11th grade. It was also a good proxy for achievement possessed by the participant before entering the new school. The remaining predictors were used to control potentially important predictors that might affect the outcome and possession of friends such as socio-economic status, ubanicity, type of school, and gender. Table 2 shows the results of OLS estimates for three types of achievement test scores.

Let us focus on the estimates of the effects of having at least one friend in the classroom. Surprisingly, the effect was negative. For science, the effect was statistically significant. This effect may be interpreted in two ways. The first is a causal interpretation. For instance, having a known peer in a high-school classroom might be connected with

Table 2 OLS estimates and dependent variables (PISA scores in 2010)

Predictors	Reading	Mathematics	Science
Friends in classroom	−0.733 (1.410)	−0.340 (1.384)	−2.983** (1.435)
PISA 2009 score	0.779*** (0.0124)	0.882*** (0.0140)	0.842*** (0.0136)
ESCS	2.274** (1.101)	2.440** (1.005)	1.443 (1.175)
Live in the countryside	0.186 (1.940)	0.230 (2.026)	−1.170 (2.168)
Type of school			
General comprehensive	(Reference category)		
Vocational with comprehensive program	−14.69*** (3.854)	−2.025 (4.075)	−9.460** (4.076)
Basic vocational school	−39.53*** (4.071)	−30.22*** (4.317)	−34.16*** (4.642)
Female	5.942*** (1.934)	−0.821 (2.072)	−0.312 (2.168)
N	4951	4951	4951
Adj. R^2	0.793	0.833	0.777

Standard errors in parentheses

* p < 0.1; ** p < 0.05; *** p < 0.01

antisocial behavior if friendship was based on anti-school subculture and thus indirectly negatively affects learning. Having a known peer in the classroom might open alternative ways of spending time outside class, thus limiting learning time. However, this interpretation does not align with the findings of previous studies that investigated the effects of friends (Wentzel and Asher 1995; Wentzel and Caldwell 1997). Another possibility is that the results were biased because some unobservable variables not included in model were related to both the number of friends in school and learning outcomes. For instance, smaller classes might be more effective because the probability of finding friends from the neighborhood would be lower in small classes compared to large classes.

IVs results

The IVs technique was used to check weather OLS results are unbiased and unobservable variables related to both the number of friends in school and learning outcomes were not standing behind obtained results. Distance to school from home in minutes was an instrument for the independent variable, friends in classroom. This variable was expected to be connected to the predictor because the closer the school is to the student's home, the higher the probability that someone in the same neighborhood would enroll there. Another assumption in IVs is that the instrument should not be causally related with the outcome. We argue that in this analysis there is no direct path between distance to school and learning outcomes. By doing this we are also assuming that time spent on traveling to school does not substantially affect learning, rest, or sleeping time. This is likely to be true as distance to school is measured in minutes rather than hours (it takes 29.6 min to get to school for an average respondent; with a range of 7.5–75 min).

The model for IVs estimation might be expressed by two equations:

$$\begin{aligned}\text{Stage 1:}\quad \text{FRIENDS} = {} & \alpha_0 + \alpha_1 \text{DISTANCE} + \alpha_2 \text{PISA2009} + \alpha_3 \text{ESCS} \\ & + \alpha_4 \text{COUNT} + \alpha_5 \text{VOC} + \alpha_6 \text{BVOC} + \alpha_7 \text{FEMALE} + \delta \end{aligned} \tag{4}$$

$$\begin{aligned}\text{Stage 2:}\quad \text{PISA2010} = {} & \beta_0 + \beta_1 \text{FRIENDS}^* + \beta_2 \text{PISA2009} + \beta_3 \text{ESCS} \\ & + \beta_4 \text{COUNT} + \beta_5 \text{VOC} + \beta_6 \text{BVOC} + \beta_7 \text{FEMALE} + e \end{aligned} \tag{5}$$

where DISTANCE is the number of minutes required to travel from home to school, which was reported by each participant and all other variables stay unchanged from Eq. (3). Along with the recommendation of Baltagi (2002, p. 277; see also Sect. 4) the same set of covariates were used in both stages of the IVs estimation.

The Stage 1 equation is designed to "clean up" the endogenous predictor, leaving only the exogenous part of the variation that covaries with the instrument. The inclusion of covariates in the Stage 1 model helps to fulfill the assumption that there is no direct relationship of the instrument to the outcome. Learning outcomes might be connected with distance by school selection (i.e., better students are more selective and better schools are more selective) and the potential non-random distribution of good schools. Because we controlled for prior achievement, the problem of selection seemed irrelevant. There was also little evidence for the non-random distribution of schools after conditioning on prior achievement, socio-economic status, type of residence, and type of school. If better

schools are located closer to the students that learn better, we should be able to rule out this by conditioning on the covariates.

Stage 2 is identical to an OLS equation except that variable FRIENDS* is "cleaned up" by the instrument (DISTANCE in Stage 1). Under the expressed assumptions such treatment allows us to estimate the causal effect of having friends on learning outcomes. As OLS regression Stage 2 is specified to show learning gains that requires inclusion of prior achievements (PISA 2009) and all relevant-for-learning outcome variables.

Table 3 shows the results of the first stage 2SLS described by Eq. (4). The PISA (2009) score, ESCS indicator and distance were significantly related to the instrumental variable. It is especially important that instrumental variable (DISTANCE) is significantly related with instrumented variable (FRIENDS) because this relationship is required by the IVs technique. Though R-square measures are not very high, the strength of the instrument is appropriate. The *F* test is commonly used to test the strength of the instruments and is provided in most statistical packages. The suggested threshold for identifying strong instruments is an *F* value of 10 (Stock et al. 2002; Stock and Yogo 2005). Lower values suggest that the instruments are weak, and the estimates might be prone to biases. This is an important limitation in models that comprise multiple instruments but not so serious in single instrument studies as in the present study (Bound et al. 1995). High values of the *F* test for this study (4 times higher than the threshold) suggest good properties of the instruments in the presented analysis.

Table 4 presents the results of the second stage of the IV 2SLS estimation. Let us focus on the effects of friends. Estimates based on the IVs reveal results that contrast those found by the OLS estimates. The estimates of the effects of friends were positive although not significant. The large standard errors are an artifact of the IVs estimation method and results in reduced power in comparison to classical OLS estimation.

The results of IVs estimation reverse the sign of the effect, which in turn changes the interpretation of the results. However there are several reasons for challenging this result.

Table 3 IV 2SLS estimates of the first stage model, dependent variable: friends in the classroom

Predictors	Friends (for reading)	Friends (for mathematics)	Friends (for science)
Distance	−0.0027*** (0.0004)	−0.0027*** (0.0004)	−0.0028*** (0.0004)
PISA 2009 score	−0.0005*** (0.0001)	−0.0004** (0.0001)	−0.0004** (0.0001)
ESCS	−0.0240** (0.0101)	−0.0240** (0.0102)	−0.0249** (0.0102)
Live in the countryside	0.0047 (0.0173)	0.0043 (0.0174)	0.0062 (0.0173)
Type of school			
General comprehensive	(Reference category)		
Vocational with comprehensive program	−0.0421* (0.0241)	−0.0415* (0.0237)	−0.0383 (0.0239)
Basic vocational school	0.0421 (0.0346)	0.0543 (0.0332)	0.0564* (0.0332)
Female	−0.0119 (0.0170)	−0.0338* (.0177)	−0.0278 (0.0172)
N	4951	4951	4951
F (excluded instrument)	41.7518	39.9005	41.821
Adj. R^2	0.0272	0.0266	0.0266

Standard errors in parentheses

* p < 0.1; ** p < 0.05; *** p < 0.01

Table 4 IV 2SLS estimates of the second stage model, instrument: distance to school (PISA score in 2010)

Predictors	Reading	Mathematics	Science
Friends in classroom	6.885 (16.18)	17.24 (15.58)	6.171 (17.07)
PISA 2009 score	0.782*** (0.0147)	0.889*** (0.0154)	0.845*** (0.0152)
ESCS	2.466** (1.135)	2.878*** (1.041)	1.682 (1.190)
Live in the countryside	0.317 (2.047)	0.533 (2.141)	−1.026 (2.246)
Type of school			
General comprehensive	(Reference category)		
Vocational with comprehensive program	−14.26*** (3.858)	−1.088 (4.111)	−9.010** (4.027)
Basic vocational school	−39.74*** (3.997)	−30.86*** (4.424)	−34.54*** (4.579)
Female	6.074*** (1.971)	−0.133 (2.238)	−0.0123 (2.332)
N	4951	4951	4951
Adj. R^2	0.791	0.825	0.775

Standard errors in parentheses

* p < 0.1; ** p < 0.05; *** p < 0.01

First, it might be an effect of the violated assumptions in this instrument. However, as previously discussed, all assumptions should hold in this setting. Second, because the IV provided only LATE, it is possible that the effects of having friends were heterogonous. In the presented situations, LATE refers to students for whom having a friend in the classroom was affected by the distance to school. However, this group probably included most of the population as it is very hard to imagine a situation where distance to school from home would not affect the number of friends from the neighborhood (expect students that do not have friends in neighborhood), thereby reducing the importance of the problem of the LATE. Finally, because the precision of the estimates were poor, the focus should be on coefficient intervals instead of the point estimates, which are subject to random error and one could rather advocate no effect of friends at all.

The results were somewhat disappointing. Because of the high uncertainty of the estimated parameters of the effects of friends, definite conclusions about the true size of the effect cannot be drawn. However, the results showed that the OLS results might be biased, and the negative direction of the effects of friends is uncertain.

Conclusions

This paper reported serval examples of the IVs technique in educational studies as well as included a pedagogical illustration using large-scale assessment data. Under several conditions, the IV technique has the potential to provide causal estimates even when non-experimental data are used. The necessary conditions consist of finding a theoretically appropriate instrument or instruments related to the research questions. One should remember that even the most convincing arguments and theoretical claims will not change an invalid instrument into valid one. In spite of arguments, an instrument might appear to be invalid and in this situation, the estimated effects would still be biased. Consequently, results of IVs estimation, like any other results obtained from observational studies, should be taken with scientific caution and critical examination of the instrument.

Moreover, the instrument must not only be valid but also sufficiently strong to determine the necessary amount of exogenous variation to allow for precise estimations.

The estimated effects should be homogenous rather than heterogeneous, which would allow for the generalization of the results. Nonetheless, the fulfillment of these conditions would be the exception rather than the rule. However, the potential benefits are that future research could produce results that are significant.

The IVs technique is an intriguing option for researchers interested in estimating causal relationships using observational data, especially researchers working with large-scale assessment data. However, the main obstacle for using this technique on large-scale assessment data is the difficulty in finding proper IVs. Of course, finding IVs that suit a particular problem and fulfill all required assumptions is not easy. It depends mostly on the knowledge, experience, and imagination of the researcher. But knowledge, experience and imagination may be supported by the design of research instruments.

Large-scale assessments bring very detailed information about student's background characteristics, detailed information on different attitudes to learning, teaching and different aspects of knowledge are collected and analyzed. However, there is still no systematic thinking about preparing variables that might be considered to be used as IVs. For instance, information about distance to school is not available in the PISA international database and was added only to the Polish extension of this project. IVs estimation could be facilitated with the addition of the distance variable to the PISA main survey. The IVs classification system developed by Murnane and Willett (2010) seems to be a useful conceptual tool for developing questionnaires to provide potential IVs. If research is designed for answering some educationally relevant question using international assessment data, the researcher could begin this process by determining whether the database contains variables that describe (1) the participants' proximity to relevant educational institutions; (2) institutional rules and personal characteristics; (3) or deviations from cohort trends. If these variables exist, theoretical assumptions should be inspected.

Even without coordinated efforts to include reasonable instruments in large-scale assessment questionnaires the richness of data like PISA and the variety of educational systems participating in such programs makes large-scale assessment data a promising area for conducting analysis based on IVs.

Authors' information

Artur Pokropek is an assistant professor at the Polish Academy of Sciences and at the Educational Research Institute in Warsaw. He holds a PhD in education sciences, an MA in sociology and an MA in education sciences from the University of Warsaw. He is a researcher in psychometrics and applied statistics in the field of education. He specializes in causal inference, multilevel analysis, structural equation modeling and IRT modeling. In addition to the methodological and statistical fields, he works on topics such as: determinants of school effectiveness, social structure, gender segregation and behavioral genetics. He has authored or co-authored several books and academic articles on these topics.

Acknowledgements

This article has been prepared under the Project from School to Work: Individual and Institutional Determinants of Educational and Occupational Career Trajectories of Young Poles which is funded by the Polish National Science Centre, as part of the grant competition Maestro 3 (UMO-2012/06/A/HS6/00323).

Received: 12 January 2016 Accepted: 20 January 2016
Published online: 24 February 2016

References

Acemoglu, D., & Angrist, J. (2000). How large are human-capital externalities? Evidence from compulsory-schooling laws. *NBER Macroeconomics Annual, 15*(1), 9–59.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics, 106*(4), 979–1014.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *The Journal of Economic Perspectives, 15*(4), 69–85.
Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on student achievement. *Quarterly Journal of Economics, 114*(2), 533–576.
Angrist, J. D., & Lavy, V. (2002). New evidence on classroom computers and pupil learning. *The Economic Journal, 112*(482), 735–765.
Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumentalvariables. *Journal of the American statistical Association*, *91*(434), 444–455.
Baltagi, B. H. (2002). *Econometrics*. New York: Springer.
Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Development, 67*(3), 1206–1222.
Bedard, K., & Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics, 121*(4), 1437–1472.
Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association, 90*(430), 443–450.
Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.
Currie, J., & Yelowitz, A. (2000). Are public housing projects good for kids? *Journal of Public Economics, 75*(1), 99–124.
Dee, T. S. (2004). Are there civic returns to education? *Journal of Public Economics, 88*(9), 1697–1720.
Domański, H., Federowicz, M., Pokropek, A., Przybysz, D., Sitek, M., Smulczyk, M., & Żółtak, T. (2012). From school to work: Individual and institutional determinants of educational and occupational career trajectories of young Poles. *ASK. Research and Methods, 21*(1), 123–141.
Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Publications.
Frentz, C., Gresham, F. M., & Elliott, S. N. (1991). Popular, controversial, neglected, and rejected adolescent: Contrasts of social competence and achievement differences. *Journal of School Psychology, 29*(2), 109–120.
Hanushek, E. A., Schwerdt, G., Wiederhold, S., & Woessmann, L. (2015). Returns to skills around the world: Evidence from PIAAC. *European Economic Review, 73*(C), 103–130.
Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.
Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics, 115*(4), 1239–1285.
Hoxby, C. M. (2002). How does the makeup of a classroom influence achievement? *Education Next, 2*(2), 56–63.
Kleibergen, F., & Zivot, E. (2003). Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics, 114*(1), 29–72.
Lee, J., & Fish, R. M. (2010). International and interstate gaps in value-added math achievement: Multilevel instrumental variable analysis of age effect and grade effect. *American Journal of Education, 117*(1), 109–137.
Machin, S. J., McNally, S., & Silva, O. (2006). New technology in schools: Is there a payoff? IZA Discussion Paper No. 2234.
Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference*. New York: Cambridge University Press.
Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford: Oxford University Press.
Neal, D. (1997). The effects of Catholic secondary schooling on educational achievement. *Journal of Labor Economics, 15*(1), 98–123.
Nelson, R. M., & DeBacker, T. K. (2008). Achievement motivation in adolescents: The role of peer climate and best friends. *The Journal of Experimental Education, 76*(2), 170–189.
OECD. (2012). *PISA 2009 technical report*. Paris: OECD Publishing.
Pearl, J. (2000). *Causality: Models, reasoning and inference*. New York: Cambridge University Press.
Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*, 96–146.
Rouse, C. E. (1995). Democratization or diversion? The effect of community colleges on educational attainment. *Journal of Business and Economic Statistics, 13*(2), 217–224.
Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151.
Staiger, D. O., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica, 65*(3), 557–586.
Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics, 20*(4), 518–529.
Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*. New York: Cambridge University Press.
Weber, A. M., & Puhani, P. A. (2006). Does the early bird catch the worm? Instrumental variable estimates of educational effects of age of school entry in Germany: Diskussionspapiere des Fachbereichs Wirtschaftswissenschaften, Universität Hannover.
Wentzel, K. R., & Asher, S. R. (1995). The academic lives of neglected, rejected, popular, and controversial children. *Child Development, 66*(3), 754–763.
Wentzel, K. R., & Caldwell, K. (1997). Friendships, peer acceptance, and group membership: Relations to academic achievement in middle school. *Child Development, 68*(6), 1198–1209.
Wilkinson, I. A., Hattie, J. A., Parr, J. M., Townsend, M. A., Fung, I., Ussher, C., & Robinson, T. (2000). *Influence of peer effects on learning outcomes: A review of the literature*. Wellington: Ministry of Education.
Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.
Wößmann, L., & West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review, 50*(3), 695–736.