Edukacja. An interdisciplinary approach 2017, 84–98
doi: 10.24131/3724.170506
issn: 0239-6858

The effects of the *Playing with pictograms* package

Mirosław Dąbrowski
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*

Bartosz Kondratek
Educational Research Institute

The article presents an analysis of data gathered to assess the package *Playing with pictograms*. The package is intended as support for the development of symbolic language skills by influencing the way teachers teach mathematics during the first years of primary school. The research was conducted using a cluster-randomised repeated measures experimental design with a control group. The main research problem investigated change in the level of relevant student skills, specifically associated with the package. IRT modeling and multilevel regression were employed in the analysis. Results demonstrated significant improvement in the use of symbolic language derived from the package.

Keywords: mathematical education; teaching style; IRT modeling; multilevel regression; cluster-randomised experiment.

Hugo Steinhaus (1887–1972), an outstanding Polish mathematician and co-author of the Polish school of mathematics, was well-known for his apt catchphrases and aphorisms. Long before the European Union launched the programme of building a knowledge-based society, he had warned that "A country without mathematics will not withstand competition with those who practise it". He also encouraged people to improve their mathematical skills and ensured: "After mathematical studies, you will do it better." Why? Most of all because mathematics can and should teach people to notice relations, dependencies and regularities, deduce and predict things, argue and convince others, so it can and should teach thinking. Current studies also prove that Polish early-school education is mainly focused on "training" children to solve typical problems according to the teacher's strict instructions, largely stressing the arithmetical correctness of the performed operations (for example Dąbrowski, 2013). It seems that such a method of teaching the youngest students does not help to optimally develop their mathematical skills, suppresses their creativity and courage to solve unfamiliar mathematical problems (Binkowska-Wójcik, Boroń, Brzyska, Cikorska and Fiertek, 2014). There are strong indicators that this situation is the consequence of the educational tradition of teaching mathematics, prevailing in

* Address: ul. Banacha 2, 02-097 Warszawa, Poland.
E-mail: m.dabrowski@mimuw.edu.pl

our society and the Polish school – not only in the initial stage of teaching (Karpiński, Grudniewska and Zambrowska, 2013) – a tradition maintaining that a child is unable to solve a problem autonomously and knows only what has been learnt from adults. The *Playing with pictograms* package was created, among others, to begin the process of leaving this tradition behind.

Playing with pictograms

The *Playing with pictograms* educational package was designed and constructed to encourage teachers to change their current teaching style, give more autonomy to students and encourage them to become more intellectually involved in the learning process – also to learn more often in cooperation – and improve their skill in using symbolic language effectively. The potential effectiveness of these changes is evidenced by contemporary studies of how a child's brain learns (Gopnik, Meltzoff and Kuhl, 2004), the proper rules of learning mathematics (Sfard, 2008), and also learning through cooperation (Kołodziejczyk, Salamon-Bobińska, Karaszewski and Bobula, 2014; Slavin, 2012). Moreover, the analyses carried out under the direction of John Hattie (2009; 2011) indicate that such educational interventions are highly effective.

To be able to learn mathematics effectively and understand it, children should be intellectually active, able to cooperate with each other and they should talk about what they do, because learning this subject is a social process. The educational package is meant to effectively encourage students in their early stages of education to think and act, so therefore, it takes their educational needs into consideration. This is one of the reasons why the *Playing with pictograms* package frequently uses enactive and iconic representation, which prepares students to understand mathematical symbols. On the other hand, the function of the package was to interest the children and enhance their motivation to learn. To this end, students were consistently encouraged to work in pairs and groups and do exercises at various levels of difficulty. Sometimes different ones were used from those usually offered by school. These exercises regularly included open-ended tasks, which naturally triggered explaining and arguing as well as problem-focused tasks – often of an interdisciplinary nature. Both the proposed organisation of work and types of exercises were also meant to change the communication style of the teacher and students and start a real process of talking about mathematics during lessons – students with each other, the teacher with students, as well as students with the teacher.

The objectives assumed at the tool's design stage required convincing teachers using the package that they should change their work style and enable children to act in a different way than during typical lessons. This is why training (the first session before starting to work with the package and the second in the middle of the school year) was conducted with teachers at the tool's testing stage and they were also offered ongoing consultations.

The specific style of communicating meanings and the symbolic "multi-level character" of the Asylco pictograms matched the presented assumptions well. The *Playing with pictograms* educational package was prepared in three variants: for classes 1–3 and 4–6 of primary school and for lower secondary school. It consists of a set of aids for students and materials for the teacher. The set of aids is designed for a four-person group of students (this is another way of promoting cooperation) and includes, among others: a variety of pictograms with diverse conventionality levels; stamps with pictograms for use, i.e. in solving and preparing tasks; games (boards, game pieces, dice), which develop, i.e. the understanding of the decimal system; dry-erase boards and felt-tipped pens for noting the solutions of exercises and designing

pictograms. The teacher's package includes: a guide for teachers; a set of proposed lesson scenarios; a set of worksheets with three difficulty levels used to individualise students' work; a set of aids for the teacher, which includes, demonstration pictograms, stickers with pictograms, models of weights and computer programs supporting the development of mathematical skills of students. In addition, an electronic version with downloadable materials and Internet training for teachers wanting to use the package (see www.piktografia.pl) were prepared.

Research questions and hypotheses

To determine the effects of using this tool, we should analyse whether applying the package increased students' skills in solving mathematical exercises requiring the use of symbolic language and the development of these skills, in comparison to a situation when this package is not used and teachers applied traditional methods of teaching mathematics. This problem encouraged researchers to carry out an experimental study with a control group in which teachers did not use the *Playing with pictograms* educational package. On the basis of data collected in this study, efforts were made to test the hypothesis of the improvement of the skills of using symbolic language in the experimental classes resulting from the use of the *Playing with pictograms* package.

Due to the said characteristics of the package, two mechanisms of the potential influence of the package on students should be considered:

- direct, through the use of the specific educational aids during lessons and participation in educational situations created to support the development of the skill of symbolic thinking,
- indirect, through changes in teachers' attitudes and work styles, which translate into the introduction of more effective mathematics teaching methods and supporting the development of students' mathematical skills in a broad sense.

Although such a differentiation is not strictly disjunctive (since a teacher's decision to use the aids may be considered a change in his or her work style), it seems essential due to an understanding of the character of factors determining the possible value added resulting from using the package. In other words – the main hypothesis could be significantly complemented by an analysis of the extent to which the possible changes in students' skills are the direct result of using the package or of changes in teachers' attitudes resulting from their use of the package and participation in the training. Unfortunately, because of the small number of classes (and teachers) taking part in the experimental survey, a conclusive quantitative analysis of the hypothesis of the indirect influence of the package was impossible. The main research problem is therefore restricted only to an analysis of the value added resulting from the use of the *Playing with pictograms* package. However we will return later to the subject of the potential changes in the attitudes of teachers from the experimental group.

Design of the experimental study

The following organisational and economic factors influenced the experimental design:

- A class of students had to serve as the basic unit of experimental manipulation; the occurrence of a significant intra-class correlation (0.15) of the dependent variable was assumed;
- A plan of expensive qualitative studies accompanying the quantitative experimental study restricted the number of classes to be included in the experimental study to eight[1].

[1] According to the literature, this is the minimum number of clusters for a group-randomised control trial to achieve sufficient statistical power to detect effects that are significant from a practical viewpoint (Murray, Varnell and Blitstein, 2004).

- A significant intra-class correlation reduces the effective size of the sample. To estimate the degree to which the multi--level structure of data reducesthe strength of the statistical inference, the so-called design effect formula is used:

$$D = 1 + (m - 1)\rho, \qquad (1)$$

where: m is the number of persons in the group; ρ – the intra-class correlation coefficient. By assuming m in the range of 15–20 and ρ = 0.15, the effective sample size, $n_{ef} = {}^{n}/_{D}$, was estimated to be about 40 students per experimental condition.

In the face of the imposed restrictions, we considered that the most effective strategy (in terms of highest statistical power) of studying the effects would be to conduct a cluster-randomised experiment with a pair matching repeated measurement of the dependent variable. The general design of such a study is an experiment with a repeated measurement and a control group:

- experimental group: pretest → experimental impact → posttest,
- control group: pretest → no experimental impact → posttest,

where the experimental condition assignment is class-based and randomisation is performed within the school (which means that only schools with at least two classes could participate in the study). The inclusion of the control group in the study was necessary in order to account for the maturation effect, since a certain "natural" increase in the level of mathematical skills of students in the interval between the first and the second measurement was expected to occur. The specific experimental design implemented in the study is presented in Figure 1.

The variance between classes of the same school is significantly smaller than the general variance between classes. Two classes within a single school are expected to be more similar to each other than classes that would be randomly paired. This is the main rationale for performing the within-school randomisation. The use of pair matching reduces the probability of obtaining a control and experimental group, which would significantly differ from one another at the initial measurement. Moreover, the intra-school randomisation ensured the balancing of the experimental groups for potential confounders relating to the location of the school. This procedure is often recommended in experiments with randomised groups (Lipsey and Hurley, 2009). In addition, the within-school

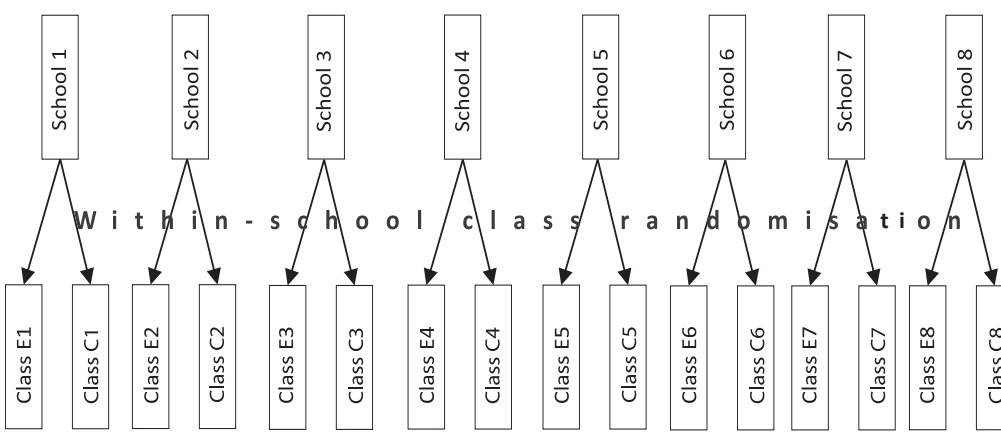


Figure 1. Assumed distribution of classes to the control group and the experimental group in the survey.

randomisation can also partially reduce the decline of effectiveness resulting from a significant intra-class correlation (Imai, King and Nall, 2009). It should be noted that randomisation within the same schools introduces certain contamination, due to the possibility of the teacher in the experimental group sharing information with the teacher from the control group. However, it was concluded that the benefits would outweigh the potential losses of such a solution.

A repeated measurement of the same students (pretest and posttest) is another important feature of the proposed experimental design. Such a solution entails obvious methodological benefits (controlling the initial differences between the treatment and control group) and purely statistical ones, which consist of increasing the statistical power when making conclusions from the dependent measurements. However, when using a repeated measurement, the memory effect, which would occur if students solved the same test of mathematical skills twice, should be taken into account. This problem was solved by using tests consisting of different items in the first and the second measurement. However, this generated another challenge – the necessity to control the difficulties of using different tools. This is why a number of exercises from the Polish nationwide *Survey of key competences of third-grade students of primary schools* (*Badanie umiejętności podstawowych uczniów trzecich klas szkół podstawowych*) carried out in 2008 at the Central Examination Board (Centralna Komisja Egzaminacyjna, CKE; Dąbrowski, 2009) were included in the pretest and posttest. This survey was carried out on a large ($N = 3965$) representative sample of students and included a large number of mathematical exercises (8 different sheets). The results of this survey were calibrated in IRT models (Kondratek, 2009). Table 1 presents the general plan of the construction of both tools used in the pretest and the posttest. The tests do not include common items, but due to nesting within the CKE survey, it was possible to estimate the level of mathematical skills of students on a common scale.

The items used in the CKE survey in this design were divided into three disjunctive sets: "block A items", which were not used in the experiment but were used in the CKE survey; "block B items", which were included in the pretest; and "block C items", which were included in the posttest. To increase the reliability and accuracy of the tools used in the experimental study, the pretest included, in addition to block B items, an "additional item 1"; the same was done in the case of the posttest – in addition to items from block C, an "additional item 2" was included.

By adopting the conditions assumed in the experiment design (effective number of 40 students per experimental condition; significance level $\alpha = 0.05$; directional alternative hypothesis; ICC = 0.15 and, additionally, a correlation between two measurements of

Table 1
Plan of the distribution of exercises by tools used in the pretest and posttest, which makes it possible to link the results of students from different groups

Sample\Block of items	CKE block A items	CKE block B items	CKE block C items	Additional item no. 1	Additional item no. 2
CKE survey	✓	✓	✓		
Pretest (E and C groups)		✓		✓	
Posttest (E and C groups)			✓		✓

0.5) we estimated that in such an experiment, a range of statistical power of 0.8–0.95 corresponds to an ability to detect effects $f(V)$ in the range of 0.28–0.37. Such a sensitivity of the designed experiment was recognised as satisfactory. A module for analysing the F test power for the interaction effect in ANOVA with repeated measurements available in the G*power program (Faul, Erdfelder, Lang and Buchner, 2007) was used for the calculations. It should be noted that these estimations were approximate due to numerous simplifications (for example, the effective number of observations was used to simplify the problem of a decrease in statistical power due to clustering or the problem of the unreliability of measures was ignored) and a priori assumptions (value of ICC, correlation between measurements, number of students in classes). After completing the study, it turned out that the estimated values were very close to the ones adopted before the analysis: the ICC in the first measurement was 0.08 and in the second measurement, 0.17 (the increase in the ICC reflects an increase in the differentiation between the experimental group and control group); the correlation between the measurements amounted to 0.44; the number of students per class in the first measurement amounted to 10–26 (M = 20.6), whereas in the second measurement, it totalled 13–25 (M = 19.5).

Research tools used

Three areas, which were found the most important from the viewpoint of examining the effectiveness of using the package, were analysed during the testing: mathematical modelling, the understanding of notions and skill of using them, and the solution of problems with the use of cognitive processes important in mathematical thinking. The wide range of these areas necessitated their exemplification.

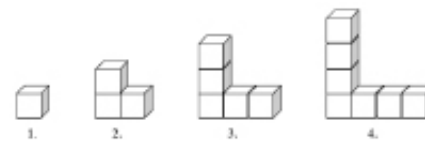
In the case of mathematic modelling, i.e. the use of mathematical tools, including symbolic language, to describe the examined phenomenon, we decided to focus on untypical word problem items. Word problem solving is the most advanced indication of mathematical modelling in the initial stage of education. The selection of untypical items, i.e. those usually not used during lessons, was meant to eliminate the possible effect of prior "training". One of these types of items used is presented here:

> Jacek and Wojtek had the same number of lollipops. Wojtek gave Jacek two lollipops. Now Jacek has more lollipops than Wojtek. How many more?

One of the most important purposes of mathematical education is developing an understanding of the decimal system. Its relational and uncorrupted[2] understanding is of key importance for all of the arithmetic taught at school, including the use of calculation algorithms. Moreover, the notation of numbers in the decimal system is the most popular school example of the practical use of symbolic language. To examine the extent of students' understanding of the decimal system, the following exercise was used:

> Some digits were blackened out in these two-digit numbers. Where possible, insert ">" or "<" in the field. Insert a question mark "?" in the other fields.
>
> a) 7 ■ □ 48 b) ■ 6 □ 33 c) 6 ■ □ ■ 2

Finally, we present one of the complex problem-focused tasks used:



[2] Understood as corrupted formalism, which is manifested by a situation when a student treats a symbolic notation, for example, of a two-digit number, as a "picture" and refers to its appearance and not the sense.

These structures were made from identical wooden blocks. They were built according to a certain rule. Guess the rule.
(a) How many blocks should the next such structure consist of?
(b) How many blocks would you need to build the tenth such structure?
(c) And how many blocks would you need to build the twentieth such structure?
(d) Describe how to quickly calculate the number of blocks needed to build the twentieth such structure.

Such a problem requires noticing the regularity of defining a sequence of solids, using it in a simple situation and in situations that become gradually more complicated, which necessitates, for example, generalisation and – finally – providing a possible clear explanation or even argument, also with the use of symbolic language.

Table 2 presents items from the 2008 CKE representative survey of third-graders' skills, which were used to prepare the tests for the first and second measurements of students' skills, divided into the three skill areas referred to above. These items correspond to "B" and "C" blocks in the plan presented in Table 1. The examples presented above of items in Table 1 have the following symbols: M1B_6, M2B_5a–c and M2B_7s respectively.

Sampling and time of performing the study

The study was carried out at the level of 3rd grade and lasted one year. The selection of the age level was dictated by the scope of using the *Playing with pictograms* educational package in the teaching process – the potential of using it is greatest in the 3rdgrade. The first measurement of students' skills was performed in September 2012 and the second one at the end of May and beginning of June 2013.

Many organisational, economic and substantive factors restricted the possibility of randomly selecting schools for the experimental survey. The organisational and economic factors include the necessity of restricting the study to three voivodships (Małopolskie, Mazowieckie and Pomorskie) and schools with at least two classes of students in 3rd grade. The selection of schools for the study was also further restricted by the assumption that because

Table 2
Use of items from the representative CKE survey in preparing the pretest and posttest

Item information			No. of observations in pretests and posttests				
Item code in the CKE research	Type of skill*	Max. no. of points	2008 research edition	Pretest (C group)	Pretest (E group)	Posttest (C group)	Posttest (E group)
M2C_7b	NRE	1	940	170	160		
M2C_7as	NRE	4	940	170	160		
M2B_7s	NRE	4	1 046			163	149
M2B_5a	DS	1	1 046	170	160		
M2B_5b	DS	1	1 046			163	149
M2B_5c	DS	1	1 046			163	149
M2B_5d	DS	1	1 046	170	160	163	149
M1B_6	PS	1	1 046	170	160		
M1A_6	PS	1	1 078			163	149
M2A_6	PS	1	1 077			163	149
M2C_6	PS	1	940	170	160		

* NRE – noticing regularities and explaining; DS – decimal system; PS – problem solving.

the schools will have to be significantly involved in the research process, they will need to be selected from a list of schools that declare their willingness to participate in the study. Moreover, given the very small number of schools to be included in the study (8 institutions), it was decided that carrying out a simple random selection of schools for the research sample would entail a serious risk of obtaining a sample of schools significantly different from the characteristics of the total population of schools with third-grade classes in Poland. Having the above boundary conditions in mind and to reduce the risk to the external validity of the study, we assumed that eight of the schools declaring participation in the research program would be selected in such a way to accomplish each combination of the following two variables:

- school location, which has two values: (a) villages and towns below 10 000 residents; (b) cities over 10 000 residents,
- average mathematics performance of a school from the *Survey of key competences of third-grade students of primary schools,* which has four values resulting from the division of average school results into equivalent quarters.

The decision to divide school locations into two categories is due to the very similar results of rural schools and those in towns with less than 10 000 inhabitants (as compared to larger cities), which was observed, among others, in the *Polish nationwide survey of key competences of third-grade students* (*Ogólnopolskie badanie umiejętności trzecioklasistów* OBUT; Pregler and Wiatrak, 2011). Moreover, dividing school location into two categories divides the population of students in Poland into more or less half, and each of the 8 fields resulting from the combination of both defined variables corresponds approximately to 1/8 of the population of third-grade students in Poland.

In the context of the discussed experimental study, the average result of a school on the scale of mathematical skills presented in the OBUT 2011 survey was adopted as the best available measure of the "average level of school mathematics performance", as it refers to the mathematical skills of third-grade students. Using the results of the sixth-grade primary school completion test for stratification was analysed as an alternative option. The compulsory nature of this exam would be an advantage (schools participate voluntarily in the OBUT surveys – this survey does not cover the entire population).

However, stratification was chosen due to the results of the mathematical part of OBUT, while the 6th grade [compulsory] test was a cross-subject test and related to the skills of sixth-graders. Table 3 presents the intervals of the average results of schools from the OBUT 2011 survey, which were obtained by dividing the schools into eight values.

Out of 34 schools that declared their willingness to participate in the study with at least two classes of third-graders, it was

Table 3

Ranges of the average results of primary schools on the general scale of mathematical skills from OBUT 2011 surveys (scale: 100; 15) in quarters by location[*]

	Range of the average results of schools			
School location	I quartile	II quartile	III quartile	IV quartile
Rural areas and towns under 10 000 inhabitants	< 93.8	[93.8;98.2)	[98.2;103.2]	> 103.2
Cities over 10 000 inhabitants	< 97.5	[97.5;101.1)	[101.1;105.0]	> 105.0

[*] Schools with fewer than five students were excluded from the calculation of the distribution of schools' average results.

possible to complete only seven cells of Table 3. No school applied to participate in the study that met the condition of 1st quartile + village and cities below 10 000 inhabitants. For this reason, a decision was taken to recruit two different one-class schools, which jointly met this condition, instead of one school. One school was drawn for the control group and the other for the experimental group.

Methods of statistical data analysis used

The performed analysis of experimental data can be divided into three stages:

- fitting a multi-group IRT model to the data,
- generating a set of plausible values (PV) for each student in each of the measurements to be subsequently used as an indication of mathematical ability,
- estimating the parameters of the model of the multi-level linear regression in which the dependent variable was mathematical ability and the independent variables were: assignment to individual experimental condition and nesting of students within measurements and schools.

Each of the listed stages will be briefly described. The analysis of Tables 1 and 2 indicates that the full data matrix includes responses to items by students from five different groups: students from the representative 2008 CKE survey and from the four groups created by combining the dichotomous conditions: "control and experimental group" and "before and after experimental impact". However, the items used from the 2008 survey in the pretest and posttest differed. In order to account for the possible differences in ability distribution between the groups, we used methods characteristic for linking test results (Kolen and Brennan, 2004; Pokropek and Kondratek, 2012; Szaleniec, Grudniewska, Kondratek, Kulon and Pokropek, 2012).

A multi-group one-dimensional IRT model was fitted to the data. The model has the following form:

$$P(\boldsymbol{U}=\boldsymbol{u}|\mathcal{P})=\int f(\boldsymbol{u},\theta,\beta)\ \psi_{\mathcal{P}}(\theta)\ d\theta, \qquad (2)$$

where: θ is a hidden random variable describing the level of students' mathematical ability; $\psi_{\mathcal{P}}(\theta)$ is a probability density function, which specifies the distribution of θ variable in $\mathcal{P}$ population; $f(\boldsymbol{u},\theta,\beta)$ is a function which specifies the probability of observing a specific $\boldsymbol{u}$ value of the $\boldsymbol{U}$ response vector, depending on the ability θ and the vector of item parameters $\beta = (\beta_1, \beta_2, ..., \beta_n)$, where the parameters of β_i item may also be vectors. The two-parameter logistic model (2PLM) was fitted to dichotomous items, whereas the other items were modelled by the graded response model (GRM).

Model fit was assessed by analysing the arrangement of the empirical proportions of the responses from each centile of ability in relationship to the characteristic curves calculated with the estimated parameters of the IRT model. The IRT model fit was considered satisfactory (Figure 2). A successful fit of a one-dimensional IRT model to the data can be regarded as an element confirming the theoretical validity of the tool – we obtain confirmation that a single main factor of mathematical skill is responsible for the observed covariance between the responses to the items used in the study.

The multi-group IRT model was fitted in such a way so that the mean and the standard deviation of the distribution of the ability of students taking part in the CKE survey were fixed at 0 and 1 respectively.

This enabled a common scale of the hidden variable, on which the results of the pretest and posttest are presented, to be related to the standard distribution of results of the large and representative sample of students who participated in the CKE survey. This allows the results of the students taking part in the experiment to be referenced to the total population and the external validity of the study to be assessed.

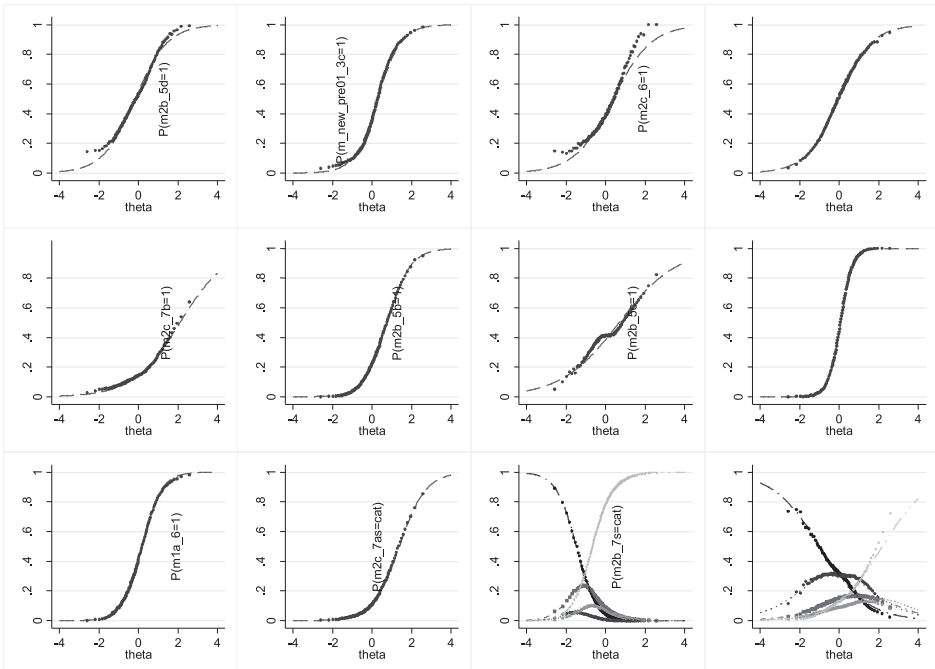


Figure 2. IRT model-fit for items measuring the mathematical skills of students.

Because of the use of psychometric tools to measure mathematical skills, we needed to account for their unreliability in the analysis of results. Otherwise, estimations of the effects and their standard errors would be biased. Use of the IRT model made this possible. Instead of using point estimations of students' results for testing the formulated research hypothesis, the analysis was performed using sets of two hundred PVs generated on the basis of the vector of responses given by each student. These are realisations from the a posteriori distribution of a student's ability, conditioned by the independent variables used in further analyses. An introduction to statistical analyses with the use of PVs is provided by Margaret Wu (2005), whereas a more in-depth treatment of this approach can be found in the monograph of Roderick Little and Donald Rubin (2002). During the generation of PVs, in addition to the vector of student responses, the fixed effects for experimental conditions and nesting of students within measurements and schools were taken into account as additional variables for conditioning the ability variable. To estimate the reliability of the measurements of the dependent variable in the first and second measurements, the average correlation between all PVs of students in these measurements was measured. The values of 0.72 for the first and 0.67 for the second measurement were obtained.

To test the formulated hypothesis about the positive influence of the *Playing with pictograms* educational package on the studied skills of students, a three-level regression model was used:

$$Y_{ijk} = \gamma_{000} + \gamma_1 {}^1X_{ijk} + \gamma_2 {}^2X_{ijk} + \\ + \gamma_{1x2} {}^1X_{ijk} {}^2X_{ijk} + \varepsilon_{00k} + \varepsilon_{0jk} + \varepsilon_{ijk}, \quad (3)$$

where:

- Y_{ijk} – value of the dependent variable, i.e. the level of mathematical ability of i^{th} student nested within *j-th* measurement (pretest – posttest), nested within k^{th} class;
  - $^{1}Xijk$, $^{2}Xijk$, $^{1}Xijk$, $^{2}Xijk$ – values of independent variables specifying the experimental conditions, respectively:
  - $^{1}X_{ijk}$ – variable indicating the group (0 = control; 1 = experimental);
  - $^{2}X_{ijk}$ – variable indicating the second measurement (0 = pretest; 1 = posttest);
  - $^{1}X_{ijk}\,^{2}X_{ijk}$ – product of the above variables, i.e. the interaction variable of the group indicator and the measurement indicator (0 = other cases than "1"; 1 = posttest and the experimental group);
- γ_{000}, γ_{1}, γ_{2}, γ_{1x2} – fixed regression coefficients (fixed effects), the values of which we estimated from the data; respectively: intercept; effect for $^{1}X_{ijk}$ variable; effect for $^{2}X_{ijk}$ variable; and effect for interaction of both $^{1}X_{ijk}$ and $^{2}X_{ijk}$ variables;
- ε_{ijk} – value of the random error from student's level for i^{th} student nested within j^{th} measurement, nested within k^{th} class; we assume that this random component has $N(0, \rho_{1})$ distribution;
- ε_{0jk} – value of the random error from the measurement's level, for j^{th} measurement nested within k^{th} class; we assume that this random component has $N(0, \rho_{2})$ distribution;
- ε_{00k} – value of the random error from the level of class, for k^{th} class; we assume that this random component has $N(0, \rho_{3})$ distribution;

The above three-level regression is based on a division of the residual component into: ε_{ijk} error responsible for the unexplained variance of students' results nested within measurements (i.e. within a single student) ρ_{1}; ε_{0jk} error responsible for the unexplained variance between measurements within class ρ_{2} and ε_{00k} error responsible for the unexplained variance between classes ρ_{3}. This is the model for repeated measurements and as such, it increases the strength of statistical inference by including the correlation between the results of the same students from the first and the second measurement and it also includes a significant (non-zero) intra-class correlation. Examples of using a multi-level regression for analysing data from repeated measurements can be found in the studies of Sophia Rabe-Hesketh and Anders Skrondal (2008). It is noteworthy that the same regression model was also used for conditioning during the above-described PV generation.

The inclusion of three independent variables in the regression model as in equation (3) is a classic method of analysing data collected in an experimental design using repeated measurements with a control group. Regression coefficients in the case of the above-described coding of variables are interpreted as follows:

- γ_{000} intercept corresponds to the average level of skills in the control group at the time of the first measurement,
- γ_{1} specifies how much the level of skills of the experimental group is greater than that of the control group, regardless of whether we take the first or the second measurement into account,
- γ_{2} specifies how the level of students' skills in the second measurement is greater compared to the first measurement, regardless of whether we take the control group or the experimental group into account,
- γ_{1x2} specifies how much the level of skills of the experimental group in the second measurement is greater, if we already take into account the information from the earlier two coefficients, i.e. after including the average difference in the skill level between measurements and after including the average difference in the level of skills between groups.

The last parameter γ_{1x2} is thus the most important parameter in the context of

assessing the effects of using the *Playing with pictograms* package. It specifies the increase of skills of the experimental group in the second measurement – which is specifically connected with the experimental influence.

Results

The results of fitting the regression model to the collected data are presented in Table 4. Let us recall that according to the manner of anchoring the IRT multi-group model (which was used to generate PVs), the distribution of mathematical skills has an average of 0 and standard deviation of 1 for the results of students who participated in the 2008 representative survey. Therefore, the effects presented in Table 4 are expressed on the standard deviation scale in the representative surveys.

The estimated value for the intercept (γ_{000}) is negative, which indicates a lower level of skills measured during the first measurement in the control group than among students who participated in the 2008 representative survey. This direction of the difference is not surprising, since the pretest was performed at the beginning of the school year in the 3rd grade, whereas the CKE representative survey was performed at the end of this grade. Due to the large error of parameter estimation, there are no grounds for stating that this effect is statistically significant.

Estimation of the γ_1 parameter in the case of the variable indicating the experimental group is very close to zero and not statistically significant, which means that there were practically no differences between the experimental group and the control group in skill level during the first survey (although the 95% confidence interval of the parameter is quite broad: from -0.46 to 0.43 of the standard deviation). This is the expected effect of the randomisation of classes to the experimental group and the control group.

Estimation of the γ_2 parameter for the variable indicating the second measurement (posttest) totalled +0.657 and the standard error was several times smaller (0.120). This means that the level of students' skills measured between the first and second measurement, irrespective of the experimental impact, increased by more than half of the standard deviation, and is statistically significant. The increase in the level of students' skills between the first and second measurement is also an expected result, which proves the validity of performing the survey in an experimental scheme with a control group. Without including the control group in the survey, it would have been impossible to separate the effect of using the *Playing with pictograms* educational package from the significant increase in students' mathematical skills, which naturally occurs in the 3rd grade.

Table 4
Estimations of regression coefficients of a multi-level model testing the statistical significance of the results of the study on the effectiveness of the Playing with pictograms package[*]

			95% confidence interval			
Variables in the regression model	Effect (*γ*)	*se*	Lower limit	Upper limit	*z*	*p*
intercept	-0.158	0.159	-0.470	0.154	-0.991	0.322
Group indicator E ()	-0.016	0.226	-0.459	0.428	-0.069	0.945
Second study indicator ()	0.657	0.120	0.422	0.891	5.494	0.000
Interaction ()	0.314	0.178	-0.035	0.662	1.762	0.078

[*] *p*-values for non-directional alternative hypothesis H_1: $\gamma \neq 0$.

Estimation of the γ_{1x2} parameter, namely the interaction parameter measuring the increase in mathematical skills specifically related to the experimental impact, totalled 0.314. The increase in the level of mathematical skills was therefore estimated to be over 0.3 of the standard deviation, amounting to about 48% of the increase taking place due to the lapse of time between the first and second measurement. With a fairly large standard error, the 95% confidence interval of this estimation ranged from -0.03 to 0.66 of the standard deviation, so it even includes negative values. In the case of the non-directional alternative hypothesis, this would mean, then, the lack of a statistically significant effect (p-value = 0.078).

To assess the effectiveness of the experimental impact in the case of a research hypothesis assuming a certain direction of influence, the null hypothesis for the interaction parameter is tested against an alternative directional hypothesis, which assumes that the change will occur according to the planned direction. In our situation, this is denoted by the pair:

$$H_0 : \gamma_{1x2} = 0$$

$$H_1 : \gamma_{1x2} > 0$$

Given such an alternative hypothesis and a positive value of the estimated effect, this means that the p-value is equal to half of the p-value for the non-directional alternative hypothesis, i.e. 0.039. In the case of a significance level of $\alpha = 0.05$, this means that there is a statistically significant effect. Thus, finally, the collected evidence makes it possible to reject the null hypothesis about the lack of an effect of the experimental impact on the level of students' skills in favour of the alternative hypothesis, which assumes an increase in the level of skills specifically related to using the *Playing with pictograms* educational package.

Discussion

The survey showed that the increase in the skills of using symbolic language among students from classes using the educational package for a year was statistically significantly greater than in the control classes. The *Playing with pictograms* package was constructed in such a way as to persuade teachers working with it to think about their working methods and encourage them to modify their style of daily work. This is why we can assume that the observed significant effect is the direct consequence of changes in the teacher's working style.

Although the described experimental survey was not designed to test such a hypothesis, the measurement of students' skills was also accompanied by a study of the educational views of teachers in the control and experimental classes. Among teachers from the experimental group, significant changes, in terms of value, in the level of the three measured dimensions of educational views were observed: their educational pessimism and educational formalism declined, whereas the results on a scale of promoting self-reliance grew (Dąbrowski and Żytko, 2013). The direction of changes corresponds to the expected one, since other studies (for example Kondratek, 2011) have shown that the aforementioned three dimensions are significantly correlated with the mathematical skills of students (the first two negatively and the last one positively). Unfortunately, due to the very small sample of teachers, we lacked the statistical power to recognise the said changes as significant. It should be also taken into account that the relationship between the educational views of teachers and the results of teachers discovered in the earlier survey are correlative, and on their basis, it is not possible to conclusively deduce the causal direction, for example, between a decrease of educational pessimism and an increase in students' skills.

Additional information about the potential influence of the change of the teacher's working style on the level of students' skills was derived from the qualitative data collected during systematic observations conducted throughout the experiment, from reports written by teachers while working with the package (Dąbrowski and Żytko, 2013), and interviews carried out with teachers and students at the end of the annual testing of the package. It might be worth quoting several statements in this context[3]:

> I viewed students from a different perspective, for example, the weaker students surprised me.
>
> I think I didn't know earlier that asking questions was so valuable. Now I've gotten used to it and if I introduce a topic, I ask them many questions that they have to answer and present their views.
>
> As a teacher with many years' teaching experience, I have learnt a lot. I stopped suggesting answers to children and I patiently wait for their answers. I do not assume that a child will not cope with an exercise as sometimes happened before. I know that the students, and not the teacher, must be active during the teaching process.

The analysis described in this article was performed for a general mathematical ability scale in order to obtain the most reliable measure of skills as possible. An analysis performed at the level of single items (cf. Dąbrowski and Żytko, 2013) indicates that the increase of the results in the experimental group was recorded in all assumed areas, i.e. in the field of solving untypical word problems, understanding the decimal system, using it and solving problems. The said results divided into subareas of skills are, however, biased due to the small sample of items per scale. For this reason, they should be treated as a guide for possible directions of future research.

[3] The statements are taken from the *Report from the evaluation of innovative teaching aids: the Playing with pictograms educational package and the effects of using it at the testing stage* (www.projekt-piktografia.pl). The authors of this report stress that they are a good reflection of the views of the entire group of teachers testing the aids.

Literature

Binkowska-Wójcik, W., Boroń, I., Brzyska, S., Cikorska, M., Fiertek R. et al. (2014). *Bydgoski bąbel matematyczny. O wprowadzaniu zmian w nauczaniu matematyki w klasach I–III*. Warszawa: Instytut Badań Edukacyjnych.

Dąbrowski, M. (2013). *(Za)trudne, bo trzeba myśleć*. Warszawa: Instytut Badań Edukacyjnych.

Dąbrowski, M. (ed.). (2009). *Badanie umiejętności podstawowych uczniów trzecich klas szkoły podstawowej. Raport z badań ilościowych 2008*. Warszawa: Centralna Komisja Egzaminacyjna.

Dąbrowski, M. and Żytko M. (eds.). (2013). *Raport z testowania innowacyjnej pomocy dydaktycznej: pakiet edukacyjny Gramy w piktogramy*. Warszawa: Wydawnictwo Bohdan Orłowski.

Faul, F., Erdfelder, E., Lang, A.-G. and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Gopnik, A., Meltzoff, A. N. and Kuhl, P. K. (2004). *Naukowiec w kołysce. Czego o umyśle uczą nas małe dzieci*. Poznań: Media Rodzina.

Hattie, J. (2009), *Visible learning. A synthesis of over 800 meta-analyses relating to achievement.* Routledge: London–New York.

Hattie, J. (2011). *Visible learning for teachers: maximizing impact on learning.* Routledge: London–New York.

Imai, K., King, G. and Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science*, *24*(1), 29–53.

Instytut Analiz Europejskich (n.d.). *Raport z ewaluacji innowacyjnej pomocy dydaktycznej: pakiet edukacyjny Gramy w piktogramy i efektów jego stosowania na etapie testowania*. Retrieved from http://projekt-piktografia.pl/wp-content/uploads/2013/12/zal.-21-Raport-z-ewaluacji.pdf

Karpiński, M., Grudniewska, M and Zambrowska, M. (2013). *Nauczanie matematyki w gimnazjum. Raport z badania*. Warszawa: Instytut Badań Edukacyjnych.

Kolen, M. J. and Brennan R. L. (2004). *Test equating, scaling, and linking: methods and practice* (2nd ed.). New York: Springer.

Kołodziejczyk, J., Salamon-Bobińska, K., Karaszewski, N. and Bobula, S. (2014). Nauczanie kooperatywne (uczenie się we współpracy). In G. Mazurkiewicz (ed.), *Edukacja jako odpowiedź. Odpowiedzialni nauczyciele w zmieniającym się świecie* (pp. 163–177). Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Kondratek, B. (2009). Konstrukcja skal mierzących umiejętności językowe i matematyczne uczniów oraz poglądy edukacyjne nauczycieli. In M. Dąbrowski (ed.), *Badanie umiejętności podstawowych uczniów trzecich klas szkoły podstawowej. Część III: trzecioklasista i jego nauczyciel* (pp. 186––215). Warszawa: Centralna Komisja Edukacyjna.

Kondratek, B. (2011). Poglądy edukacyjne nauczycieli klas 1–3. In M. Dąbrowski (ed.), *Trzecioklasiści 2010. Badanie umiejętności podstawowych uczniów trzecich klas szkoły podstawowej. Raport z badań ilościowych* (pp. 230–241.) Warszawa: Centralna Komisja Edukacyjna.

Lipsey, M. W. and Hurley S. M. (2009). Design sensitivity: statistical power for applied experimental research. In L. Bickman and D. J. Rog (eds.), *The SAGE handbook of applied social research methods* (2nd ed., pp. 44–76). Thousand Oaks: Sage Publications.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Murray, D. M., Varnell, S. P. and Blitstein, J. L. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health*, *94*(3), 423–432.

Pokropek, A. and Kondratek, B. (2012). Zrównywanie wyników testowania. Definicje i przykłady zastosowania. *Edukacja*, *120*(4), 52–71.

Pregler, A. and Wiatrak, E. (eds.). (2011). *Ogólnopolskie badanie umiejętności trzecioklasistów. Raport OBUT 2011*. Warszawa: Centralna Komisja Edukacyjna.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and longitudinal modelling using Stata*. College Station: Stata Press.

Sfard, A. (2008). *Thinking as communicating: human development, the growth of discourses, and mathematizing.* Cambridge: Cambridge University Press.

Slavin, R. E. (2012). Uczenie się oparte na współpracy. Dlaczego praca w grupach jest skuteczna. In F. Benavides, H. Dumont and D. Istance (eds.). *Istota uczenia się. Wykorzystanie wyników badań w praktyce* (pp. 248–276). Warszawa: Wolters Kluwer.

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F. and Pokropek, A. (2012). Wyniki egzaminu gimnazjalnego 2002–2010 na wspólnej skali. *Edukacja*, *119*(3), 9–30.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation 31*(2–3), 114–128.

The survey was performed as part of the project "Pictography. Developing the skill of using symbolic language in mathematical education with the use of Asylco pictograms" („Piktografia. Rozwijanie umiejętności posługiwania się językiem symbolicznym w edukacji z zakresu nauk matematycznych z zastosowaniem piktogramów Asylco") project carried out in cooperation with the University of Warsaw and co-financed from the European Social Fund (Human Capital Operational Programme 2007–2013, Priority III High quality of the education system). A preliminary version of this article was published in Polish in *Edukacja*, *134*(3), 2015.