Orbis scholae

VOL 12 / 2 / 2018

Using International Large-Scale Assessments for Informing National Policies

Guest editors
Paulína Koršňáková and Jana Straková

Charles University / Karolinum Press / 2018

ISSN 1802-4637 (Print)
ISSN 2336-3177 (Online)

Contents

Editorial

In recent decades, the implementation of international large-scale assessment surveys (ILSAs) has become an integral part of educational reality in many countries. The average achievements of individual countries in particular assessment domains and their developments over time have been reviewed and reflected on by educators and discussed in public and political arenas. At times, some additional findings occupy the spotligh such as social disparities, gender differences, student attitudes and motivation, teachers' roles, the use of technologies, and parental involvement.

Academic and research communities have been excitedly debating the limitations of these findings. While some focus on the negative impacts of ILSAs on teaching and learning, others engage in sophisticated international analyses that use more and more advanced statistical methods in the exploitation of the available datasets. Many of these analyses, as well as general discussions, however, do not give sufficient consideration to the unique settings of each of the educational systems involved, including culturally diverse contexts and historical experience. In addition, apart from the average scores, seeing and appreciating the findings is not always straightforward, and even when these are recognized and interpreted, it cannot be taken for granted that their causes are revealed or understood correctly.

While the frameworks and instruments of ILSAs, as well as the procedures, can be tuned to a wide range of respondents, the findings and especially their interpretations need to be reflected on and validated at a national level. The lack of national analyses creates a barrier to the valorisation of investments in ILSAs and also hinders progress in learning about educational policy issues faced by individual countries and the ways in which ILSAs could be used to inform policy making in individual jurisdictions.

The aim of this special issue is to showcase examples of the useful and interesting national utilization of ILSAs and the results they have been providing at a national level. We were interested in examples of national analyses that seek to provide answers to important questions of national education policies that may be difficult to answer by other means, in attempts to relate international assessments to national ones, and in national extensions that countries add to ILSAs. The issue contains six research papers that show a variety of possible uses of data from international large-scale assessments for national purposes.

https://doi.org/10.14712/23363177.2018.290
www.orbisscholae.cz

The first two papers provide analyses of national data obtained in PISA 2015. In the first paper, Francisco López Rupérez, Isabel García García, and Eva Expósito-Casas present a comparative efficiency analysis of public spending on education in 17 remarkably decentralized Spanish regions (Autonomous Communities). Their aim is to shed light on both the educational policies developed and the corrective state measures in favour of inter-territorial equity. The authors argue that the efficient use of resources is an essential factor in a good governance system, particularly in the area of public administration, where the needs are unlimited while the resources are always limited. In the analysis presented in the paper, educational outcomes are measured with an arithmetical average of scores obtained in the PISA 2015 tests in the main assessment domains, corrected for the socio-economic composition of the students in individual regions, and are related to educational expenditures. The authors categorize the regions according to their efficiency, effectiveness, and equity, and propose policy recommendations for both the regional and central government levels that are based on their findings.

The second paper exploiting the PISA 2015 data focuses on information and communication technologies in Czech schools. The authors, Libor Juhaňák, Jiří Zounek, Klára Záleská, Ondřej Bárta, and Kristýna Vlčková, begin with the notion that the implementation and the use of ICT in schools is one of the longstanding strategic objectives and priorities in education policy documents in the Czech Republic. Up to now, however, comparatively little attention has been paid to research on the relationship between the use of digital technologies and students' performance. The paper investigates the association of various ICT-related factors with the educational outcomes of students in Czech schools. It aims to determine the extent to which the availability and the use of ICT in school and at home affect students' educational achievements. The study shows that the relationships differ for different assessment domains and different student backgrounds and confirms the need for further exploration.

The following two papers relate IEA TIMSS studies to national assessments of the same age cohorts. In the Hungarian paper, Ildikó Balázsi and Ildikó Szepesi carry out a comparative analysis of TIMSS 2015 and the National Assessment of Basic Competencies (NABC) 2015, which assesses all students' reading and mathematics performance in Grades 6, 8, and 10. The authors utilized the fact that both studies assessed Hungarian Grade 8 students' mathematical abilities at the same time (spring 2015) and that the data collected in the two studies could be linked on the student level using Student Measurement IDs. Their aim was to compare the constructs measured by both studies and to validate the results of TIMSS, which assessed a sample of students by the data collected in the whole population. The analysis confirms that the estimations of population parameters based on TIMSS samples are of good quality and reveals that although the two tests use similar content and cognitive categorizations, there are crucial differences between the two constructs.

Barbara Japelj Pavešić and Gašper Cankar analysed the data from international and national surveys in order to study gender differences in the Slovenian education

system in Grades 8 and 12. In both age cohorts, they have three different assessments of mathematics at their disposal: the TIMSS assessment, national assessment, and teacher grades. The main reason for the study was unexplained gender differences in mathematics achievement, which are not consistent across all assessments. The authors utilized the fact that Grade 8 students who participated in TIMSS took the national assessment (NA) one year later and that TIMSS Advanced Maths students took the 'matura' examination in mathematics two months after the TIMSS Advanced assessment and it was possible to link the data at the student level. Moreover, both TIMSS assessments included questions about school grades from mathematics together with a series of questions about the effort put into solving the TIMSS test. The analyses focused on differences between boys and girls with respect to their assessment results, grades, attitudes towards mathematics, and future plans. It was found that the gender differences in national exams as well as in school grades differ from the gender differences in TIMSS and TIMSS Advanced. The analyses reveal some characteristics of the national exams and grading that would not be evident otherwise, and the results of the study provide fresh insights and explanations of different gender differences, providing some room for improvement in grading to teachers and policy makers.

The fifth paper, by Eva Potužníková, demonstrates the use of a national extension of an international study. To study the interest of Czech students in reading and mathematics, the author used the data from PIRLS and TIMSS 2011 together with the data obtained in the Czech Longitudinal Study of Education, which followed students participating in PIRLS and TIMSS 2011 at the time of their transition to lower secondary education. The study presented in the paper compares the effect of engaging instruction with the effect of student-related characteristics, such as gender, family background, leisure time preferences, and the perceived difficulty of the subject and investigates the development of interest over time. Implications for instructional practice are discussed, as are the advantages of the longitudinal nature of the follow-up survey.

In the last paper, Gašper Cankar uses data from PISA 2015 to demonstrate Simpson's paradox. Simpson's paradox, a case of contradictory interpretations when results are analysed by groups or aggregated as a whole, is very relevant for analyses of data from large-scale assessments as it can cause confusion and misunderstanding in the interpretation of the results. The author explores the occurrences of Simpson's paradox and conditions leading to them using PISA 2015 gender differences in achievement data in five Central European countries – Austria, Croatia, Czechia, Slovakia, and Slovenia. In countries where the occurrence of Simpson's paradox can be demonstrated, a correct interpretation of the results is discussed. The author also emphasizes the implications of his findings for educational governance and demonstrates it through the case of the Slovenian educational system.

The analyses presented in the issue demonstrate some benefits of combining international data with national resources and its potential contribution to education policy and practice. In countries without overarching national assessment systems,

international surveys are instrumental for studying regional or social disparities and providing opportunities to gain some insights into the relationships between student achievement, motivation and attitudes, and teaching practices within the structure and context of the respective educational system. A national extension of international studies increases the future value of data that has been collected, especially when it adds a longitudinal component that allows meaningful causal inferences to be drawn.

In countries with national assessments, a comparison between international and national assessments provides an opportunity for conceptual review and validation of both assessments and allows presumptions and biases hindering national practices to be disclosed that would otherwise remain as a blind spot of practitioners, administrators, and scholars at a national level. Interesting national features that deserve the attention of policy makers could also be explored by the comparative analysis of a smaller number of countries with similar cultural contexts, historical background, and educational traditions. The strong unifying aspect of all the articles presented here is their profound insider knowledge and detailed anchoring of the findings that are presented in the context of the education systems concerned and their current governance discourse.

Paulína Koršňáková, Jana Straková
Guest Editors

Educational Effectiveness, Efficiency, and Equity in Spanish Regions: What Does PISA 2015 Reveal?

Francisco López Rupérez, Isabel García García
School of Education, Camilo José Cela University, Madrid, Spain

Eva Expósito-Casas
School of Education at National University of Distance Education, Madrid, Spain

Abstract: The territorial organization of Spain into regions (autonomous communities) involves a remarkable decentralization. Therefore, it is interesting to make a comparative efficiency analysis of the public spending in education among regions that can to shed light on both the educational policies at the regional level, and the corrective state actions of inter-territorial imbalances. Furthermore, equity of the results of the education system is an indisputable goal of any society that aspires to justice and social cohesion. This research poses, firstly, an estimation of educational effectiveness and efficiency of public spending using a secondary analysis of PISA 2015 data that takes into account the value of ESCS. Subsequently, two educational equity parameters are estimated. The triple empirical categorization of autonomous communities, according to the efficiency and equity results, allows the derivation of policy recommendations of interest both at the regional and central government levels.

Keywords: efficiency; equity; public governance; evaluation; PISA 2015

The matter of efficiency in the administration of resources is a constant feature of public governance design and quality by international organisations. According to the definition of the World Bank, "Governance is the manner in which power is exercised in the management of a country's economic resources and social resources for development" (World Bank, 1992, p. 52). Kaufmann, Kraay, and Zoido-Lobatón (1999a, b), based on two research papers written for the above organisation, included in their public governance model, "the capacity of government to manage efficiently". The UN, the European Commission, and the OECD have assumed this doctrine and, often, governance itself has been considered a synonym of efficient management (López Rupérez, García & Expósito, 2017). The significance of the role of efficiency in public governance is particularly relevant in the case of education. As this is a service which, in developed countries, addresses a fundamental right – universal and free for major population age groups – education systems are public policy areas with a massive use of resources. Considering education and training as actual investments does not obviate the issue of efficiency of public expenditure but rather reinforces it (López Rupérez, 2001).

There is robust empirical evidence at an international level regarding the limitations of education expenditure as an unquestionable factor of continuous improvement

https://doi.org/10.14712/23363177.2018.291
www.orbisscholae.cz

of education outcomes (UNESCO, 2004; OECD, 2016). Recently, the OECD, after repeatedly ratifying in its various PISA reports what UNESCO has called the 'spending paradox', concluded that, "As expenditure on educational institutions per student increases, so does a country's mean performance; but the rate of increase diminishes fast, as indicated by the logarithmic scale on the horizontal axis" (OECD, 2016, p. 63).

When we focus on Spain, the question that arises is whether the variable of education expenditure in this country is still a significant factor upon which one should operate systematically in order to provide better education outcomes. While in Spain cumulative spending per student aged 6 to 15 is significant (US $74,947), regional distribution shows considerable differences between the various autonomous communities (Ministry of Education, Culture and Sports, 2017). This could lead to opportunities to improve outcomes through differential treatment of the regions regarding education spending if we take into account the non-linear relationship between spending per student and academic performance. Given this possibility, it would be essential to previously establish the most comparable picture possible, both of the public education spending of the various autonomous regions and its efficiency, without forgetting the conditions of equity that must have a bearing on the analysis and also on education policy and practice.

Throughout this study we will be addressing, first, a description of the corresponding conceptual and methodological framework. Next, we calculated spending per student in schools financed with public funds by autonomous regions harmonised through correction of the effect of rural schooling, a structural phenomenon which has a significant impact on expenditure. The above harmonisation of public expenditure (*inputs*) was followed by harmonisation of results (*outputs*), taking into account in this case regional differences in student socio-economic and cultural status (ESCS). Then, we calculated the efficiency of public expenditure on education in the autonomous communities. We analysed the relationship between wealth, public expenditure on education, and expenditure efficiency, to then proceed to an analysis and discussion of the consequences. This was followed by addressing the issue of educational equity within the autonomous communities, calculating characteristic parameters, diagnosing the situation in this regard in each community, and providing specific proposals for education policies other than those related to spending. Finally, we present a summary of the set of the most significant empirical conclusions and main recommendations in terms of policies for improvement of interest at both the regional and national levels.

1 The conceptual and methodological framework

1.1 A systemic approach

Starting from Ashby and his cybernetic paradigm (Ashby, 1956), the general functioning of an education system can be described as a combination of a set of inputs

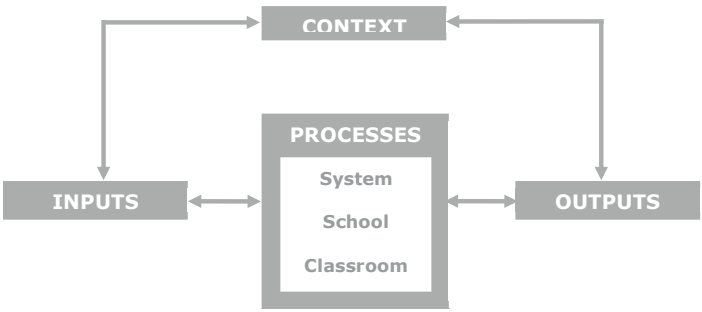


Figure 1 Systemic approach to the description of the education system.
Source: Authors' own work on the basis of an OECD scheme.

which, through a series of internal processes, turn into outputs. The context of education, with its various components, has an impact on the inputs, affects the processes (system, school, and classroom), and conditions achievement of results (Figure 1).

Based on a more complex view of this systemic pattern, which is simple and yet powerful, most of the relationships between components are bidirectional. Thus, processes act retroactively upon inputs, based on how priorities are established and the level of efficiency in their management; and outputs operate in the same way upon processes, in terms of validation or correction, and they do so with an intensity that depends on the level of intelligence of the system as a whole. Smart systems promote, deliberately, the type of feedback that generates positive returns and improves outcome quality. Finally, results have a retroactive effect on the context, at a social and economic level, with broad effects on the medium and long term, although certainly positive if the system is successful.

This study follows the intellectual tradition of a systemic approach, which is originally linked to a description of material systems and biological systems. In this tradition, the idea of efficiency is understood as the quotient between outputs and inputs or, in other words, the amount of outputs the system produces for each input unit. However, in the tradition linked to the economy, in particular – and, by extension, some social sciences – the idea of efficiency has taken on a rather more sophisticated theoretical and operational meaning.

1.2 Technical efficiency vs. productivity

The generic idea of efficiency as an output/input ratio is further refined, in the tradition of economics, in at least two other ways: first, introducing a conceptual distinction between 'technical efficiency' and productivity, and second, developing highly sophisticated calculation methods for the first (Mandl, Dierx, & Ilzkovitz, 2008; Coll Serrano & Blasco, 2006; Cordero, Salinas & Pedraja, 2005; Worthington, 2001). Nonetheless, what Cordero, Crespo, Pedraja, and Santín (2011) have pointed

out in relation to these procedures is that education is a highly complex process and there are problems of variable measurement errors, potential unobserved effects or omitted variables, together with the possibility of double causation between dependent and independent variables, all of which can generate endogeneity, which can affect the accuracy of the results.

In this paper we have preferred to use the term 'spending efficiency' to refer to what the scientific community of applied economics calls productivity (OECD, 2001), that is, the ratio between produced outputs and used inputs, so that the greater the output for the given input, or the lesser the input for the output given, the more productive the production unit.

1.3 The matter of equity

The matter of equity and social cohesion is a shared concern among developed countries which has led, among other things, to political statements in the European Union, first in connection with the Lisbon Strategy and, later, with the ET2020 Strategy[1]. In turn, the OECD has shown this same sensitivity across a broader geographical area and they have repeatedly expressed an interest in measuring the degree, or level, of equity in all the PISA editions published to date (OECD, 2010; 2014; 2016). The analysis of the relationship between the two variables of socio-economic and cultural status and academic performance helps to assess the level of equity of an education system. This was the line followed by PISA, which is based on the measurement of two characteristic parameters of this statistical relationship: the magnitude of the impact of the first of these two variables on the second one, and the intensity of those relationships.

As is known, the first is defined by the scale of the slope of the line that best fits the corresponding distribution of points upon a Cartesian graph, so that the greater the slope, the greater the difference in scores per socio-economic and cultural index unit (ESCS) (OECD, 2016). The second one measures the strength of the statistical relationship between the two variables, the percentage of performance variance explained by the ESCS variable, or, if we wish, the predictive power which ESCS has over school performance values (OECD, 2016, p. 216).

The education system of an advanced society must certainly aspire to being effective and efficient, but also fair and capable of diminishing the impact of socio-economic and cultural differences in the population on children and adolescent education outcomes, so that the liberal principle of true equal opportunity may become effective at the initial stage of human existence, at the starting line towards adulthood (Flamant, 1988).

[1] http://ec.europa.eu/education/policy/strategic-framework_en

This paper conducts a secondary analysis of the databases derived from the PISA assessment. The sample (39,066 students) comprises all of the Spanish autonomous communities that took part in the PISA 2015 assessment (all of them have a representative sample). The basic methodological framework of this study follows the systemic approach and, in particular, the pattern described in Figure 1. It is based on a single input, measured by the variable 'public spending per student in non-university educational institutions supported with public funds', and a single output, measured by the variable average score of the three PISA 2015 tests'. PISA provide 10 plausible values (used to measure the performance measurement average) and normalised student final weights (W_FSTUWT), which were used in the analyses carried out, thus providing more efficient estimates.

Moreover, the concern about equity, in the comparison between the various autonomous communities, leads to an analysis of this factor, specific of advanced educational systems, and to qualifying the resulting values of efficiency. In line with the above, the main steps to guide the corresponding calculation procedures will basically be the following:

a) Territorial harmonisation of the input variable for the seventeen Spanish autonomous communities taking into account the Rural Schooling Index (IER).

b) Territorial harmonisation of the output variable by correcting the effect of student Socio-economic and Cultural Status Index (ESCS) over the average PISA 2015 score in each one of the seventeen autonomous communities.

c) Calculation of efficiencies (outputs/inputs) and estimate of gain margins in relation to the average of the autonomous communities.

d) Calculation of parameters that confirm equity of the education systems in the autonomous communities: impact magnitude of socio-economic and cultural status on performance, and the strength or intensity of the statistical relationship between the two variables.

2 Harmonized public spending per student in non-university educational institutions supported with public funds, by autonomous community

Hereinafter, public spending per student in non-university educational institutions supported with public funds shall be considered a measurement of system inputs–treated here as a synonym for financial resources (Ministry of Education, Culture and Sports, 2017). The calculation method used follows the standards applied by the OECD in its international indicators of education systems (INES). Nonetheless, the notably different degree of rural schooling, as one of the characteristics of the unique context that exists in each autonomous community and whose influence on

education spending can be shown, requires harmonisation of the previous expenditure figures in order to improve homogeneity for comparison.

2.1 Public spending per student vs. Rural Schooling Index

As indicated elsewhere (Consejo Escolar del Estado, 2015), the factor that most explains the differences between autonomous communities, as far as the figures of public spending per student in non-university educational institutions supported with public funds are concerned, is the student/teacher ratio. A structural variable that strongly conditions ratio is the level of dispersal of the school population. This is a defining feature of rural areas that can be measured by the percentage of students enrolled in towns with less than 10,000 inhabitants. This percentage has been termed Rural Schooling Index (IER). Even when rurality can be defined by a broader set of traits, for the purposes of this study, this is the most pertinent approach and, furthermore, relatively easy to measure. A linear regression analysis between IER and public spending per student in non-university educational institutions supported with public funds confirms the existence of a direct relationship between the two variables and reveals the strength of such a relationship. This preliminary analysis indicates there is a contextual variable whose influence on expenditure should be harmonised in order to be able to make a comparison of autonomous communities under reasonably standard terms.

2.2 Comparison of harmonised public spending per student, by autonomous community

The results of the previous calculations warn about the advisability of considering this demographic contextual variable (IER). In other words, the aim is to calculate each value of public spending per student (y) resulting from standardising the degree of influence of the IER factor (x) in the various autonomous communities. To this end, we conducted the abovementioned regression analysis and then determined, by ordinary least squares (*OLS*), the best fit equation (1), presenting an R^2 coefficient of determination of 0.30 and statistical significance (0.02).

$$y = 33.73\,x + 4.0628 \quad (1)$$

Table 1 shows, in comparative terms for the various autonomous communities, gross values of public spending per student and net values resulting from applying the correction given by the model. As shown in this table, the harmonised values for public spending on education per student show notable differences between the autonomous communities exceeding, 1,400 € at the two ends: Galicia and the Community of Madrid.

Table 1 Values of the Rural Schooling Index (IER) in Spain and in each autonomous community and gross and corrected values for IER in public spending per student. Academic year: 2013–2014.

	Rural Schooling Index (IER)	Public spending per student	
		Gross values (Euro)	Corrected values (Euro)
Spain	18.7	4,537	4,693
Andalusia	15.8	4,042	4,596
Aragón	23.1	4,707	4,842
Asturias	32.8	5,530	5,169
Balearic Islands	13.0	4,808	4,501
Canary Islands	7.9	4,539	4,329
Cantabria	29.4	5,623	5,054
Castilla-León	27.1	5,109	4,977
Castilla-La Mancha	32.0	4,295	5,142
Catalonia	16.5	4,198	4,619
Community of Valencia	14.1	4,449	4,538
Extremadura	39.7	5,276	5,402
Galicia	47.6	5,404	5,668
La Rioja	27.7	4,827	4,997
Community of Madrid	5.5	3,857	4,248
Murcia	25.1	4,352	4,909
Navarre	35.0	5,692	5,243
Basque Country	20.8	6,448	4,764

Note: Public spending per student in non-university education, (occupational training is excluded). The student unit has been transformed into a full-time equivalent, according to the methodology used in international statistics aproaches. The 2013–2014 academic year is the last one for which consolidated data is available.
Source: Authors' own work using the data provided by *Las cifras de la educación en España. Curso 2014–2015 (Edición 2017)*. Ministerio de Educación, Cultura y Deporte.

3 Harmonised student academic performance, by autonomous community

For the purpose of measuring student academic performance, as the main output of the system, this study has considered the average scores obtained in PISA 2015 in the tests corresponding to the three traditional areas of literacy, mathematics and science (using the 10 plausible values in the estimation of the performance measurement average). The relatively strong link – depending on the countries – between students' socio-economic status and school outcomes forces us to subtract the influence of this variable on student results as an essential step in contextualisation

before conducting a reasonably standard comparison of country outputs and, in our case, autonomous communities.

3.1 Academic performance vs. ESCS by autonomous community

A linear regression analysis between academic performance measured with an arithmetic average of scores obtained in the PISA 2015 tests, in the three areas abovementioned, and ESCS in the various Spanish autonomous communities, showed the importance of this relationship in Spain when the autonomous community is used as the analysis unit. Of note is the considerable size of the R^2 determination coefficient (0.66) (statistical significance <0.01). That is, 66% of the variance of the results obtained in PISA 2015 by the autonomous communities can be explained by the socio-economic and cultural status index. This confirms the need, in this case, of correcting the influence of this variable on school performance when comparing the results of the autonomous communities in a reasonably standard manner.

Therefore, and in order to put the results in context, we corrected the PISA scores according to ESCS based on the 'gradient' method used by the OECD. Application of this methodology to the case in hand meant conducting 72 secondary analyses of student microdata: for all of Spain, for each one of the 17 autonomous communities, and for each score obtained in the 3 major areas of PISA 2015 (science, literacy, and mathematics), as well as for the overall score.

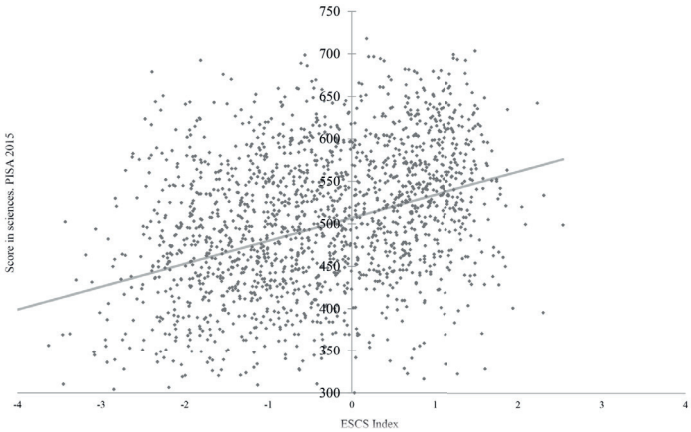
The scattered plots shown in Figure 2 indicate the overall behaviour of the two variables of interest (performance in each subject and ESCS) where each student is represented as a dot on the plane defined by their scores in both variables. This makes it possible to see the positive relationship that exists between them, which is emphasised by the best fit line for the point cloud, as well as determining the value of the ordinate at the origin indicating the corresponding corrected score for the ESCS effect.

3.2 Comparison of harmonised PISA results by autonomous community

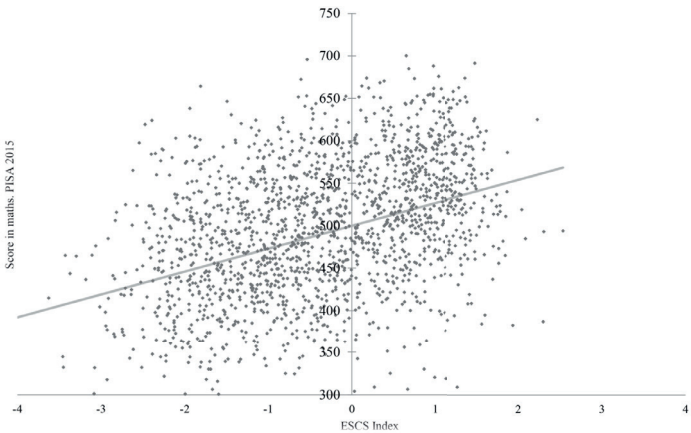
The analysis of the association between ESCS and overall performance in PISA 2015 for Spain and all of the autonomous communities shows a positive relationship and statistically significant in all cases, with an R^2 strength that ranges between 0.19 points in the case of Murcia and 0.07 in Galicia; Asturias and the Community of Madrid follow Murcia with 0.19 and 0.17, respectively. At the opposite end are Castile and Leon and the Basque Country, with values around 0.09. Table 2 shows, in comparative terms for the various autonomous communities, the gross values of overall average scores obtained in PISA 2015 and corrected values, following the PISA methodology for correcting the ESCS effect.

From the analysis in Table 2, significant differences were found among the autonomous communities, with the highest value between Castile and Leon and the

Sciences



Math



Reading

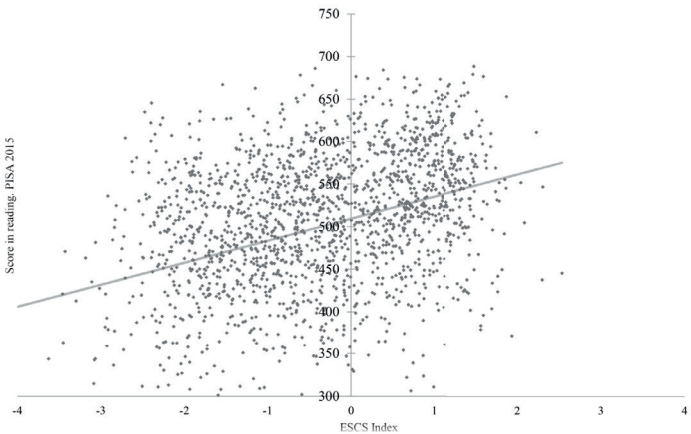


Figure 2 PISA 2015 results vs. student socio-economic and cultural status (ESCS) index in Spain. Source: Authors' own work based on PISA 2015 microdata.

Canary Islands: 32 PISA points, after correcting for the socio-economic and cultural status effect, which corresponds to approximately an average academic delay of one year between those autonomous communities.

Table 2 Gross values of overall average scores obtained in PISA 2015 and corrected values according to ESCS impact.

	Socio-economic and cultural status (ESCS) index	Average scores obtained in PISA 2015	
		Gross values	Corrected values
Spain	−0.51	491	505
Andalusia	−0.87	473	496
Aragon	−0.39	505	516
Asturias	−0.42	497	510
Balearic Islands	−0.65	482	498
Canary Islands	−0.8	470	492
Cantabria	−0.43	497	508
Castilla-León	−0.44	516	525
Castilla-La Mancha	−0.66	494	510
Catalonia	−0.35	501	511
Community of Valencia	−0.53	493	506
Extremadura	−0.79	474	494
Galicia	−0.52	505	515
La Rioja	−0.46	498	511
Community of Madrid	−0.1	513	516
Murcia	−0.82	480	503
Navarre	−0.32	515	523
Basque Country	−0.25	489	495

Source: Authors' own work based on PISA 2015 microdata.

4 Efficiency of public spending on education per student in Autonomous Communities

According to the notion of efficiency or productivity, Table 3 shows the values of this variable that are the result of considering harmonised input and output values; values which were previously shown in Tables 1 and 2, respectively. Each figure represents the euro cost of each PISA point in each autonomous community. Figure 3 is a graphic representation of the deviations of efficiency values in public spending on education per student compared to the Spanish average of the various Autonomous Communities.

Table 3 Efficiency values of public spending per student by autonomous community.

	Efficiency (PISA point /euro)
Spain	0.108
Andalusia	0.108
Aragon	0.107
Asturias	0.099
Balearic Islands	0.111
Canary Islands	0.114
Cantabria	0.101
Castilla-León	0.105
Castilla-La Mancha	0.099
Catalonia	0.111
Community of Valencia	0.112
Extremadura	0.091
Galicia	0.091
La Rioja	0.102
Community of Madrid	0.121
Murcia	0.102
Navarre	0.100
Basque Country	0.104

Source: Authors' own work.

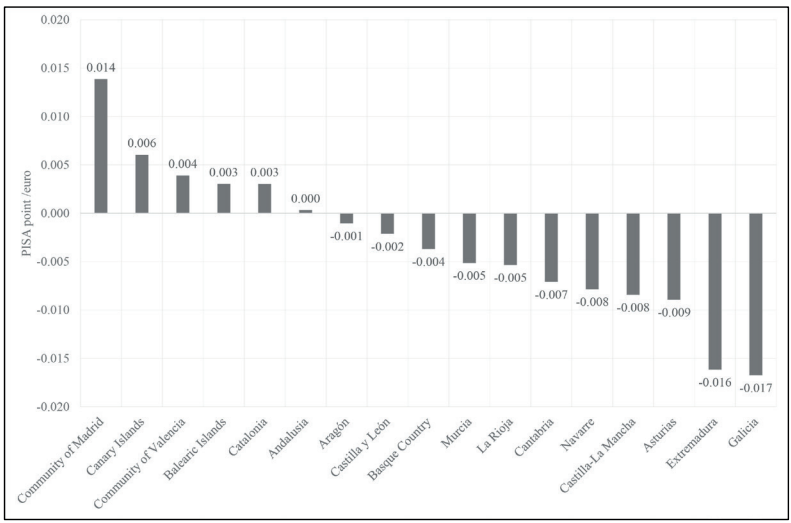


Figure 3 Deviations of efficiency values in public spending on education per student compared to the Spanish average by autonomous community.
Source: Authors' own work.

Negative deviations of the autonomous communities with efficiency values below the Spanish average show efficiency gain margins by most of them regarding the modest objective of being equal, at least, to that average. When results versus expenditure are shown on a chart and the corresponding regression analysis is carried out, the statistical relationship is very weak (R^2 = 0.09) and not significant (0.21), which indicates the heterogeneous impact of different factors on the efficiency of the various autonomous communities. Figure 3 shows this dispersion and recommends organising the autonomous community positions in quadrants in the inputs-outputs chart. Without prejudice to the subsequent analyses, it is worth examining the 'quadrant analysis' with special attention on the 'optimal quadrant' – low expenditure and high results – and the 'worst quadrant' – high expenditure and low results – in relation to the average values of the two variables considered. The first group would include, although in distant positions, the Community of Madrid, Catalonia, and the Community of Valencia; and in the second one, Murcia, the Basque Country, and Extremadura.

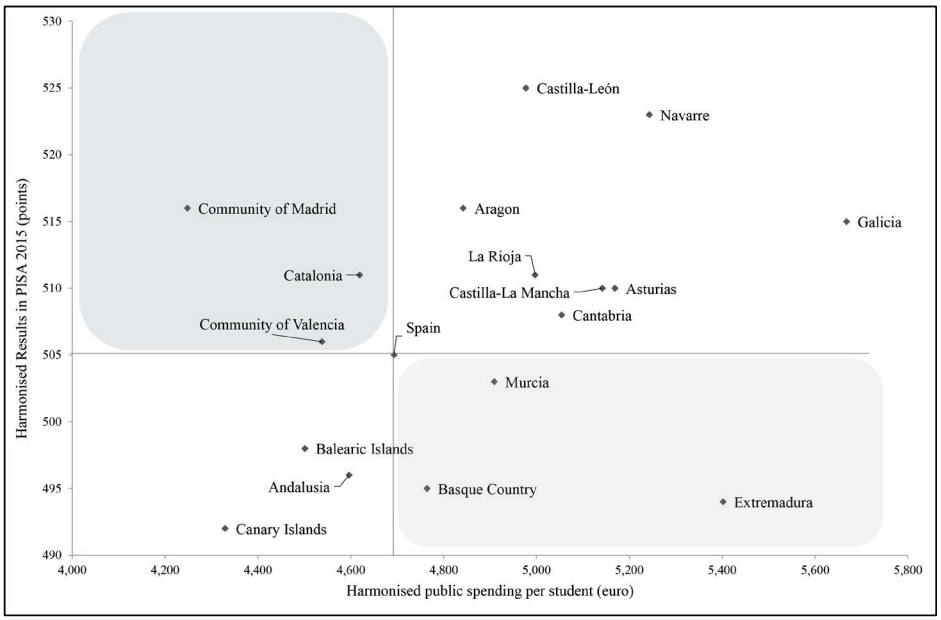


Figure 4 PISA 2015 Results vs. public spending per student and class, both harmonised, by autonomous community.
Source: Authors' own work.

5 Public spending on education vs. levels of wealth and effectiveness vs. spending efficiency

The previous analyses provide a very diverse picture of behaviours in the autonomous communities, both regarding expenditure and results, which recommends searching for clearer and more useful diagnoses in order to direct policy. The aim, after all, and in light of the resulting situation map, is to come up with recommendations for education spending in Spain, as well as for other policies.

5.1 Public spending on education vs. levels of wealth

The first step in this direction would be to include the level of wealth of the autonomous communities in the analyses. This would be justified for two reasons: first, because, as mentioned in the introduction, the impact of the expenditure variable on student outcomes depends on the degree of development of the countries – or economic units – which is reflected in the scale of their spending on education; and second, because, considering the widely recognised role of education and training as drivers of economic and social progress in the medium and long-term, the less wealthy autonomous communities should make an effort to spend more than the average on education per student.

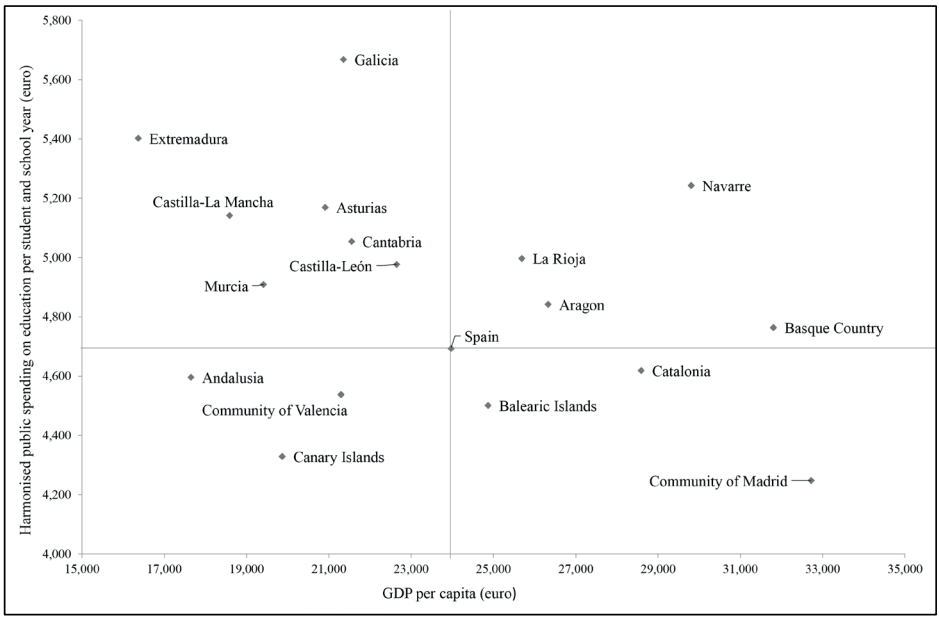


Figure 5 Harmonised public spending on education per student vs. level of wealth by autonomous community.
Source: Authors' own work.

Figure 5 shows the values of public spending on education per student versus wealth levels measured according to GDP per capita for all of the autonomous communities. The resulting point cloud evidences significant dispersion. This leads to the parameters resulting from the regression analysis and corresponding to ANOVA (R^2 = 0.10; *sig* 0.19). The above notwithstanding, the quadrant analysis provides information of great interest. When we look at the autonomous communities with below average wealth level, we find, again, heterogeneous performance. Thus, the Canary Islands, the Community of Valencia, and Andalusia, with a wealth level below average, spend less than the Spanish average; while Murcia, Castile-Leon, Cantabria, Castile-La Mancha, Asturias, Extremadura, and Galicia spend more than average. Therefore, 70% of the less rich autonomous communities are investing considerably in education through their spending policies.

5.2 Spending effectiveness vs. efficiency

Even when education spending aligned with population needs is a necessary condition to achieve good academic outcomes, it is not nearly enough. This is where the quality of the policies and their degree of efficiency to shape the well-known desideratum of 'spending better' comes in. It would therefore be advisable to follow this in the analyses and fill out the map above considering the effects of the other

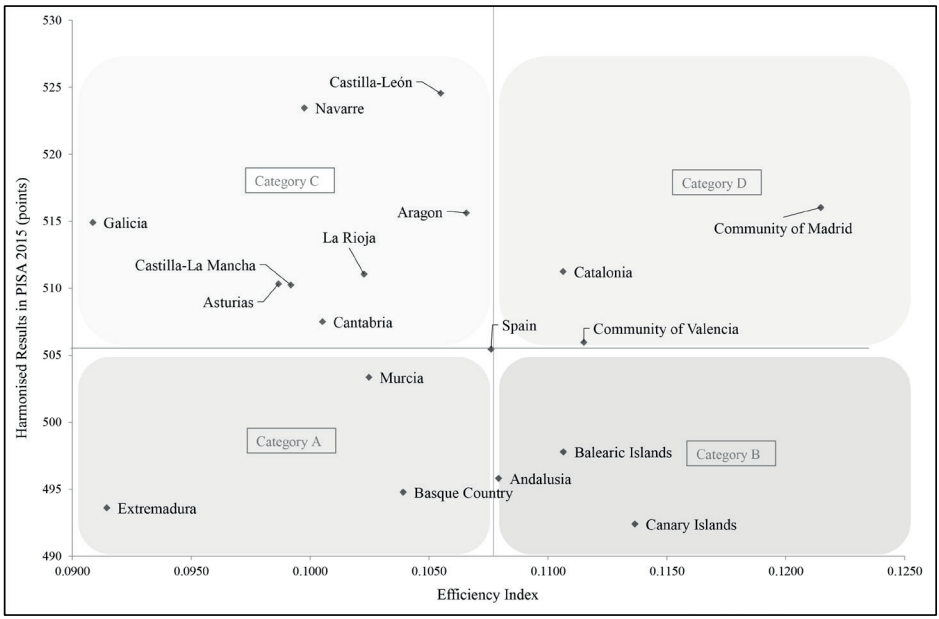


Figure 6 Effectiveness, measured by the average scores in PISA 2015, corrected for the ESCS effect, vs. the values of efficiency of public spending on education by autonomous communities.
Source: Authors' own work.

policies not related to expenditure. Figure 6 shows effectiveness values, measured by the PISA 2015 scores obtained in the various communities after correcting for the effect of ESCS versus the corresponding values of efficiency.

Again, we see a notable dispersion in the point cloud (R^2 = 0.003; *sig* 0.82). Nonetheless, it is possible to divide the various autonomous communities into four classes: those not very effective and not very efficient (Category A); those not very effective but efficient (Category B); those effective but not very efficient (Category C); and finally, those effective and efficient (Category D).

Notwithstanding the above analyses, it should be noted that efficiency, as a feature of governance quality, is not a value in itself if not accompanied by the aspiration for equity. This dimension of our diagnosis, which is not minor, is examined in depth below.

6 Education equity in the autonomous communities

Without prejudice to that constitutional ideal of equal right to quality education across the country, which we shall refer to below, it is necessary to examine, empirically, the issue of equity within each autonomous community, as well as the existing differences between them and the corresponding consequences. All of this with the aim of implementing corrective policies, including spending policies, both at the state and autonomous community levels and in line with their respective responsibilities.

6.1 Two different and complementary approaches to the degree of equity in the education system

The two parameters of the statistical relationship between ESCS and PISA scores – impact and intensity – facilitate, as mentioned above, different and complementary approaches to the degree of equity in an education system. PISA 2015 provides direct data on these two variables – socio-economic and cultural status and academic performance – for the countries and economies participating in the programme. Furthermore, their rich micro database allows one to determine the two relationship parameters abovementioned, through secondary analyses for the regions of those countries that have participated with a broader, statistically representative sample of these geographical areas. This would be the case of Spain, as shown in Section 4. Secondary empirical analyses, described above, have helped us determine the two parameters related to equity in the education system: the m slope of the regression lines, shown in Figure 3, and the R^2 coefficient of determination of the corresponding analyses. The values of both parameters are shown in Table 4.

Table 4 Impact magnitude values (m) of ESCS on academic performance, based on average scores in the three PISA 2015 tests, and the intensity of the corresponding ratio (R^2) by autonomous community.

	Intensity (R^2)	Impact (m)
Spain	0.16	26.62
Andalusia	0.16	26.36
Aragon	0.14	26.56
Asturias	0.19	30.01
Balearic Islands	0.11	23.96
Canary Islands	0.15	27.05
Cantabria	0.11	23.88
Castilla-León	0.09	19.83
Castilla-La Mancha	0.14	23.59
Catalonia	0.16	27.52
Community of Valencia	0.14	23.96
Extremadura	0.13	24.01
Galicia	0.07	18.69
La Rioja	0.15	27.15
Community of Madrid	0.17	27.53
Murcia	0.19	28.07
Navarre	0.15	26.41
Basque Country	0.09	21.36

Source: Authors' own work.

As pointed out by the OECD, in relation to PISA 2015 (OECD, 2016), "While these two measures are positively correlated, they capture different aspects of the relationship between students' performance and socio-economic status, with potentially different policy" (p. 217). The preceding contributions, regarding implications (Willms, 2006; OECD, 2013), are in this case of utmost interest to prepare evidence-based recommendations on the most appropriate type of education policies for the various autonomous communities.

6.2 Analysis of the seventeen autonomous communities

In light of the above, we should identify the position of the various Spanish autonomous communities in an impact magnitude versus intensity of the relationship chart and in accordance with a quadrant chart defined according to the national averages of these two parameters. Figure 7shows the distribution of the seventeen

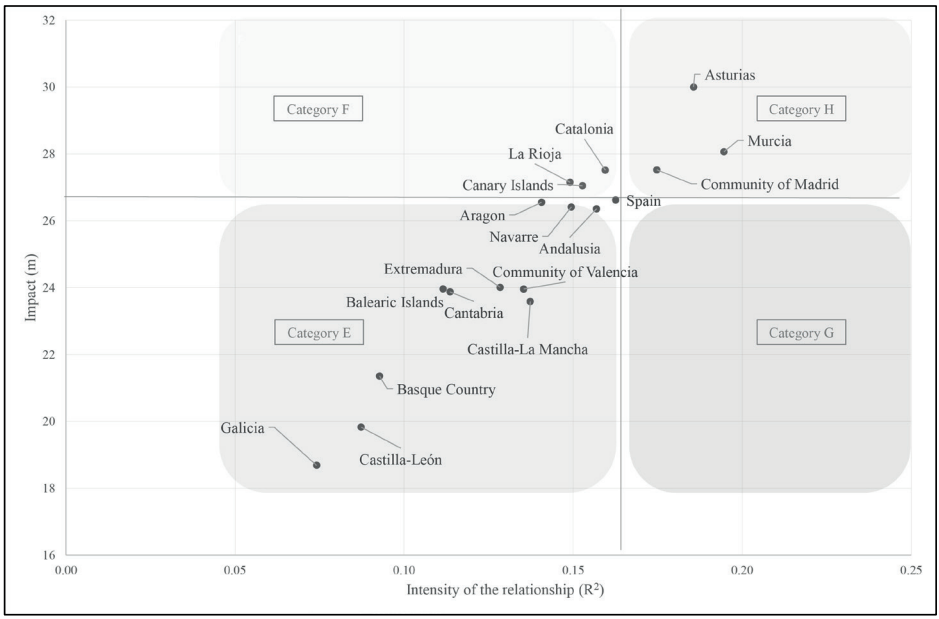


Figure 7 Distribution of the seventeen autonomous communities in the four categories according to the values of the two parameters – impact (*m*) and intensity of the relationship (R^2) of education equity.
Source: Authors' own work.

autonomous communities in the four quadrants of the chart, each one identified with the corresponding category: Category E (weak impact, weak intensity), Category F (strong impact, weak intensity), Category G (weak impact, strong intensity) and Category H.

6.3 Efficiency and equity

There is broad consensus among developed societies that, while efficient management of public resources is a characteristic of good governance, the administration of public spending cannot turn its back on the need for equity. For this reason, it was pertinent to supplement the above bivariate analyses on Spanish autonomous communities with another similar one that considered their positioning in an efficiency versus equity chart. In this case, we measured efficiency of spending on education as the efficiency index calculated previously, and as a reverse indicator of the degree of education equity, the impact magnitude (m) of the socio-economic and cultural (ESCS) status on school performance based on the average score obtained in the three core PISA tests.

Table 5 Impact values (*m*) of ESCS on PISA 2015 performance, and the efficiency index by autonomous community.

	Impact (*m*)	Efficiency (PISA point/euro)
Spain	26.62	0.1076
Andalusia	26.36	0.1079
Aragon	26.56	0.1066
Asturias	30.01	0.0987
Balearic Islands	23.96	0.1106
Canary Islands	27.05	0.1137
Cantabria	23.88	0.1005
Castilla-León	19.83	0.1055
Castilla-La Mancha	23.59	0.0992
Catalonia	27.52	0.1106
Community of Valencia	23.96	0.1115
Extremadura	24.01	0.0914
Galicia	18.69	0.0909
La Rioja	27.15	0.1023
Community of Madrid	27.53	0.1215
Murcia	28.07	0.1025
Navarre	26.41	0.0998
Basque Country	21.36	0.1039

Source: Authors' own work.

Table 5 shows the results corresponding to each autonomous community. A linear regression analysis of the two variables indicates a statistically weak and not significant relationship between them (R^2 = 0.13; *sig* 0.15). The notable dispersion of the point cloud is not compatible with a sufficiently established causation ratio between the two variables, so there is no type of determinism that makes efficiency and equity two irreconcilable factors. The challenge, both for autonomous communities and for the state, is to make the two factors compatible and not opposing. This shall undoubtedly depend on the appropriateness of the definition and implementation of education policies, including those related to spending.

Four categories corresponding to the respective quadrants of the chart in Figure 8 can be identified: Category I (low impact, low efficiency), Category J (high impact, low efficiency), comprising the so-called 'worst quadrant' as it groups inefficient and low equity behaviours, Category K (low impact, high efficiency), corresponding to the 'optimal quadrant', Category L (high impact, low efficiency) with high levels of efficiency and their good, or relatively good, performance outcomes, and finally Category I (low impact, low efficiency), which is the most populated one, including 8 autonomous communities. This is the predominant category in the country as it

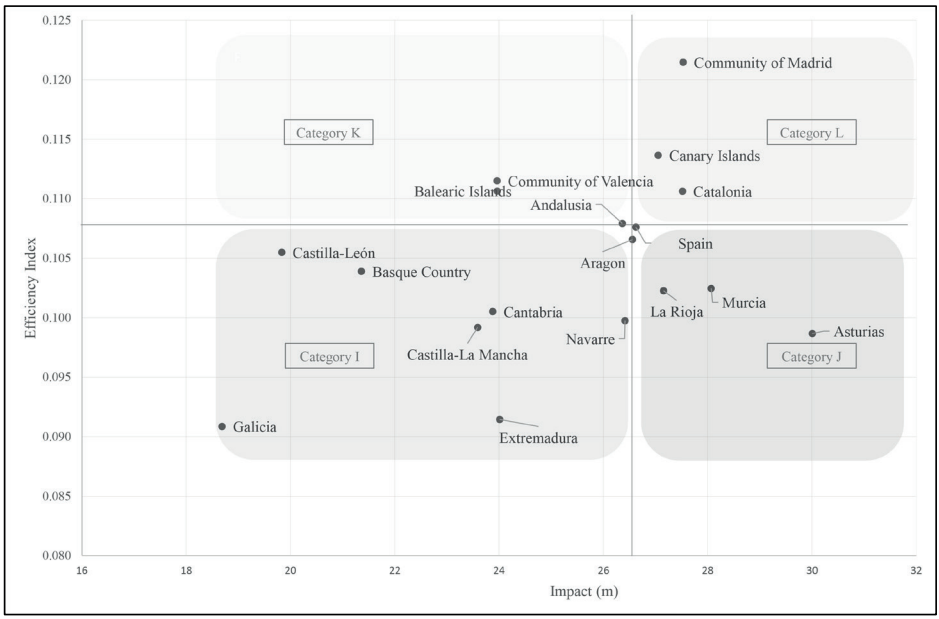


Figure 8 Efficiency vs. equity for the seventeen Spanish autonomous communities. Source: Authors' own work.

groups almost half of the 17 autonomous communities in which there is a level of equity that is higher than average, accompanied, nonetheless, by a level of spending efficiency that is lower than average.

7 Discussion

One of the issues that has emerged from the analyses of the data in this study is the considerable dispersion of the point cloud in the inputs versus outputs charts. Unlike the acknowledgement made by UNESCO in 2004, or the comments repeated by the OECD, in this same sense on the PISA data and described above, this notable dispersion of the point cloud in the case of the seventeen Spanish autonomous communities poses a problem when it comes to identify, even approximately, the spending threshold below which the magnitude of resources could have a significant impact on outcomes in Spain. Determining this threshold would, to a certain extent, have allowed clarifying actions of the public administrations in this regard and, in particular, the actions by the state to effectively ensure real equal opportunities among Spanish students, regardless of the autonomous community in which they live.

The OECD, using cumulative spending per student aged 6 to 15 as an input indicator, set this threshold at US $50,000 (PPP). In comparison, Spain, as a whole, with an amount of US $74,947 (PPP), is significantly above the threshold (OECD, 2016). This leads one to think that, in spite of the existing differences between the autonomous

communities regarding spending on education, they are all above the threshold figure. However, from the standpoint of assurance of the constitutional principle of equal opportunities, the problem of the source of the differences in school outcomes among the autonomous communities still stands, without being able to completely rule out the possibility that funding differences is one of the variables, internal to the Spanish education system, which could be having, among others, a statistically significant impact on these differences in academic performance. Various studies have previously addressed the problem of the determining factors of the differences in education performance in Spain (Villar, 2012). Beyond the influence of certainly different regional socio-economic levels, which could be harmonised with statistical procedures as we did in this study, those papers gave rise to certain factors related to policies – preschool education, school characteristics, etc. However, a substantial part of the differences found cannot be attributed to any of the explanatory variables considered.

At this point, we should note the limitations there may be in the spending efficiency values calculated in this paper, which are partly due to the relatively diffuse variables involved, that is, not apparently linked to the policies, such as cultural guidelines or level of social involvement, but which, nonetheless, have an impact on outcomes. We are considering here education policy in a broad sense, which includes the explicit definition of the goals of the reforms and their priorities (the actual 'policies'), formulation of strategies to achieve those goals, and specific plans for their implementation (Mingat, Tan, & Sosale, 2003). These cultural guidelines are based on family and social values that have an impact not only on the confined family setting, but also on school culture, on peer interactions and on school climate. This contributes to creating a social atmosphere, in general, that favours academic success while at the same time generated by it in a kind of virtuous cycle. This social mechanism is not necessarily linked to the level of wealth of the corresponding autonomous community but rather to the nature and strength of its alignment with the shared values that are conducive to academic achievement (Méndez, Zamarro, García, & Hitt, 2015). The role of the so-called 'non-cognitive skills' which, as pointed out elsewhere, are strongly linked to the world of attitudes and the area of values (López Rupérez & García, 2017), has proven to be pertinent in order to explain differences in academic performance between autonomous communities. Thus, the study by Méndez et al. (2015) estimated that the reduction of the standard deviation in the differences found in non-cognitive skills linked to academic performance leads to a reduction of approximately 25% of the differences found among autonomous communities regarding their average scores in the PISA tests. Another factor related to school climate and culture, as a set of shared standards and values, is the interaction among students (peer effects), for which Hattie (2003), based on meta-analytical syntheses, estimated it explained between 5% and 10% of the performance differences among students.

According to the above, those autonomous communities that have this valuable collective capital, with equal spending on education, will be more effective and

probably more efficient. Upon careful consideration of the remaining policies other than those of expenditure, we should now consider whether it would be possible to operate in that area of classic virtues, values, and attitudes at the level of the autonomous communities and also at the state level regarding their responsibility for providing equal and basic conditions for school achievement. The answer is definitely in the affirmative (López Rupérez & García, 2017), therefore it cannot be discarded that significant explanation for interregional variance in school outcomes is associated with these policies unrelated to spending. For example, the introduction of the so-called 'character education' in the school syllabus, as proposed by the Center for Curriculum Redesign (Fadel, Bialik, & Triling, 2015) and contemplated later by the OECD's BIAC (BIAC, 2016), ratifies the above. Stressing these types of policies must be one of the goals of quality education governance. In other words, the absence or omission of these policies is an intrinsic source of inefficiency whose impact is probably embedded in the data of the quadrant chart in Figure 4 and the subsequent analyses. At an international level, a relatively intense relationship has been identified between resilience – as a recognised non-cognitive skill – and performance in PISA 2015 in the set of participating countries (López Rupérez & García, 2017). With an R^2 determination coefficient of 0.76, the study reveals both the strength of this relationship as well as the privileged position of some Eastern countries, even those with a lower level of development, a position that is most likely linked to the education philosophy of those societies and the shared code of values in their schools (Stevenson & Stigler, 1992).

One of the global results revealed in this study is the remarkable territorial inequality which, both in inputs or resources, and outputs or outcomes, comprises the Spanish territorial landscape. This inequality implicitly alludes to the conditions under which citizens enjoy the fundamental right to education, and its correction concerns the state *prima facie*, one of its exclusive responsibilities pursuant to Article 149.1.1 of the Spanish Constitution, being, "Regulation of the basic conditions that guarantee equality among all Spanish citizens to exercise their rights and compliance with constitutional obligations."

The reasoning and evidence provided by this study give rise to the appropriateness of a twofold action in this area: one on the side of inputs that allows increasing resources wherever it is objectively necessary, due to the low level of spending on education, due to the lower level of wealth, or due to the notably lower results; and another one on the side of outputs, acting upon procedures in order to help increase school results through improvement of education governance quality. With this two-pronged approach, actions both by the central government and the regional governments must be coordinated, loyal, and smart. A provision of extraordinary funds by the state must be earmarked, as a priority, for those autonomous communities which, in spite of their lower level of wealth, devote above average amounts of resources and yet obtain lower results. Establishing a strong bond between those extraordinary funds from the state and an improvement in processes, monitoring policies, their evaluation and results, is an essential procedure to ensure efficient

use of these additional resources. Equally important is support from the Ministry of Education, through plans agreed upon with the regional education authorities, in the form of assistance for diagnosis, orientation, international consulting, etc., in those autonomous communities that, in light of their results, require and request them, which will provide knowledge and competences and will contribute to the success of education improvement plans which can no longer be delayed.

8 Conclusions and recommendations

Effectiveness, efficiency, and equity are three factors of education systems that can be measured with a secondary analysis of PISA 2015, alluding to other characteristic features of governance quality. This study provides new data and original diagnosis analysis related to each one of the seventeen Spanish autonomous communities. Below is a set of conclusions summarising the essential findings of the study and providing recommendations, in line with those findings, for each one of the autonomous communities based on their position in relation to national averages in each one of the three factors that are characteristic of advanced education systems. The aim is to provide public authorities with grounded guidance aimed at facilitating their intervention in the Spanish education system upon an empirical base.

8.1 Conclusions

From the empirical data and the analyses generated in this study, we can draw the following conclusions, summarised as follows:

a) Regional distribution of spending on education per student shows considerable differences among autonomous communities, which together with the proven non-linear relationship between spending per student and academic performance established in international analyses, opens the door to the possibility of improving outcomes through a different treatment of the autonomous communities regarding spending on education with efficiency criteria.

b) The existence of significant differences among autonomous communities regarding the level of rurality of their school systems, with actual impact on the average cost of the school place supported with public funds, requires, for the purpose of comparison, empirical territorial harmonisation actions on public spending on education per student.

c) The known influence of the socio-economic and cultural status (ESCS) of students on academic performance makes it necessary to control said influence in order to ensure that the comparison of autonomous communities is relatively homogeneous.

d) In line with the systemic approach, efficiency of public spending on education of the various autonomous communities can be calculated as the quotient between the system output, measured with the average score in PISA 2015 corrected by ESCS,

and its input, measured with public spending on education per student harmonised in relation to school rurality.

e) Representation of the positions of the seventeen autonomous communities in the chart of harmonised inputs-outputs leads to a point cloud with a considerable degree of dispersion, which indicates the existence of factors of a different nature that have an impact, heterogeneously, on the efficiency of the various autonomous communities. Unlike the case of international analyses conducted in this regard by UNESCO and the OECD, it is impossible to draw an efficiency curve that matches, in a statistically significant manner, those point cloud and, therefore, it is not possible to empirically determine, with this methodology, the threshold under which an increase in spending per student could result in an evident improvement in academic performance.

f) Notwithstanding the above conclusion, we conducted a quadrant analysis, based on national averages, with the following key results. In the 'optimal quadrant', related to efficiency (low expenditure and high outcomes), are Catalonia, the Community of Madrid, and the Community of Valencia. As for the 'worst quadrant' (high expenditure and low outcomes), it includes Extremadura, Murcia, and the Basque Country.

g) The empirical relationship between public spending on education per student and the level of wealth of an autonomous community, measured by GDP per capita, is not very clear. There are autonomous communities with a lower level of wealth that spend more than average and other richer ones that spend less than average.

h) When we look at the autonomous communities with a level of wealth that is below average, we find that Murcia, Castile-Leon, Cantabria, Castile-La Mancha, Asturias, Extremadura, and Galicia spend more than average after harmonising expenditure in line with the Rural Schooling Index. Therefore, 70% of the less rich autonomous communities are investing considerably in education through their spending policies.

i) The representation of the effectiveness values – measured by the PISA 2015 scores obtained in the various autonomous communities after correcting for the effect of ESCS – versus the corresponding values of efficiency allows us to group the 17 autonomous communities into four categories:

- *Category A* (low efficiency, low effectiveness) includes Murcia, Extremadura and the Basque Country. All of them are making a financial effort in favour of education which is greater than average, but this is not being reflected, at least at present, in the outcomes. This situation indicates a problem with processes and policies, that is, with governance.
- *Category B* (high efficiency, low effectiveness) includes Andalusia, Balearic Islands and the Canary Islands. The situation of these autonomous communities indicates a problem of insufficient funding that should be corrected either by the state or the community itself.
- *Category C* (low efficiency, high effectiveness) includes Galicia, Asturias, Castile-La Mancha, Navarre, Cantabria, La Rioja, Castile-Leon, and Aragon. These

communities spend more than average to obtain higher than average results but with an efficiency that is lower than average. In these cases, efficiency should be improved with a 'focus on outputs', which means operating on the processes in order to spend available resources better.

- *Category D* (high efficiency, high effectiveness) includes the autonomous communities that are effective and efficient. This is the case of Catalonia, the Community of Valencia and the Community of Madrid. The above notwithstanding, it is necessary to remember that efficiency is not an acceptable value if it occurs at the expense of equity.

j) Representation of the values of the autonomous communities in an efficiency versus equity chart allowed grouping the 17 autonomous communities into four categories:

- *Category I* (low equity, low efficiency), groups the autonomous communities of Asturias, La Rioja and Murcia. Priority actions should focus both on improving the outcomes of all and on serving socially disadvantaged environments.
- *Category J* (high equity, low efficiency) is the most populated one, including 8 autonomous communities (Aragon, Cantabria, Castile and Leon, Castile-La Mancha, Galicia, Extremadura, Navarra, and Basque Country). The challenge for this category is then to improve its efficiency without reducing its equity level, which must be done by either improving outcomes without reducing spending or increasing both but in a way that the rise in outputs is greater than that in inputs.
- *Category K* (low equity, high efficiency) is occupied by the Canary Islands, Catalonia and the Community of Madrid which, in spite of their high, or relatively high, efficiency in public spending on education, show equity level levels lower than average. Given their high levels of efficiency and their good, or relatively good, performance outcomes, Catalonia and the Community of Madrid have the necessary conditions to prioritise equity policies. This is not the case of the Canary Islands, whose high levels of efficiency stem from very limited spending on education combined with low academic outcomes, as in the case of the Balearic Islands and Andalusia.
- *Category L* (high equity, high efficiency), includes the Balearic Islands, the Community of Valencia and Andalusia. The presence of these three communities in this category does not imply ignoring any improvement initiative given that their PISA results show noticeable progress, particularly in the Balearic Islands and Andalusia.

8.2 Recommendations

Based on the empirical evidence summarised in the conclusions section, below are a series of recommendations on how to direct policies for improving education in each one of the seventeen autonomous communities included in Table 6, with details on their characteristic traits regarding effectiveness, efficiency, and equity.

Table 6 A synopsis of the recommended educational policies based on the empirical diagnosis of work.

Autonomous Communities	Effectiveness	Efficiency	Equity	Recommendations
Community of Valencia	+	+	+	Despite the positive assessment in the three factors considered, the results obtained in PISA show that this Autonomous Community still has room to improve efficiency compared to other. For this reason it would be advisable to pay attention to the educational policies of a general nature, described above and aimed at raising the level of performance of all students.
Catalonia Community of Madrid	+	+	−	Because of its high efficiency levels and good, or relatively good, performance results, these Communities have the necessary conditions to prioritize compensatory policies such as those described above, focusing particularly on helping students of modest status to obtain better results.
Aragón Cantabria Castilla-León Castilla-La Mancha Galicia Navarre	+	−	+	Without reducing its level of equity, it is about improving its efficiency, either by improving the results without reducing the cost, or by increasing both, but in such a way that the increase in outputs to be greater than that of the inputs. For this, it would be recommended to influence in the policies that have the greatest impact on the results, particularly those based on the teaching centers as preferential units of action.
Andalusia Balearic Islands	−	+	+	Focus on educational policies of a general nature designed to raise the level of performance of all students, through interventions from the State (model of the teaching profession, general management of the curriculum, conception of school management, etc.), and from the Autonomous Community (school management, school climate, permanent teacher training, stimulus system, complementary academic organization, family-school relations, etc.). Develop actions aimed at improving the non-cognitive abilities of students.

Autonomous Communities	Effectiveness	Efficiency	Equity	Recommendations
Asturias La Rioja	+	−	−	Prioritize compensatory policies focused on socially disadvantaged students and evaluate systematically their degree of effectiveness. Develop specific plans on centers that, according to objective indicators defined for that purpose and referring to socioeconomic aspects, require a priority intervention by public authorities. Set intervention plans on those centers that show lower performance than what would be expected of them due to the socio-economic and cultural level of the population that they are schooling. Evaluate the impact of such policies.
Canary Islands	−	+	−	Undertake coordinated policies by the State and Autonomous Community, both of a general nature and specifically focused on disadvantaged sectors. Mobilizing the material, human and knowledge resources necessary to save these situations of regional disadvantage.
Extremadura Basque Country	−	−	+	Focus on educational policies of a general nature aimed at raising the level of performance of all students, with criteria of efficiency, through interventions of the State (model of the teaching profession, general management of the curriculum, conception of school management, etc.), and of the Autonomous Community (management of the centers, school climate, permanent teacher training, stimulus system, complementary academic organization, family-school relations, etc.) with greater impact on the results. Develop actions aimed at improving the non-cognitive abilities of students.
Murcia	−	−	−	Undertake coordinated policies by the State and the Autonomous Community, of a general nature and specifically focused on disadvantaged sectors. Mobilizing, with criteria of efficiency, the material, human and knowledge resources necessary to save these situations of frank regional disadvantage. Promote actions aimed at improving the non-cognitive abilities of students.

Note: The + or − signs of the table indicate values of each of the three variables considered higher or lower respectively to the corresponding national averages.
Source: Authors' own work.

References

Ashby, W. R. (1956). *An introduction to cybernetics*. London: Chapman & Hall.

BIAC (2016). *Business Priorities for Education. A BIAC Discussion Paper.* Retrieved from http://biac.org/wp-content/uploads/2016/06/16-06-BIAC-Business-Priorities-for-Education1.pdf

Coll Serrano, V., & Blasco Blasco, O. M. (2006). *Evaluación de la eficiencia mediante el análisis envolvente de datos. Introducción a los modelos básicos*. Valencia: B-EUMED 2000. Retrieved from www.eumed.net/libros/2006c/197/

Consejo Escolar del Estado (2015). *Informe 2015 sobre el estado del sistema educativo. Curso 2012–2013*. Madrid: Ministerio de Educación, Cultura y Deporte.

Cordero Ferrera, J. M., Crespo Cebada, E., Pedraja Chaparro, F., & Santín González, D. (2011). Exploring educational efficiency divergences across spanish regions in PISA 2006. *Revista de Economía Aplicada*, *57*(XIX), 117–145.

Cordero Ferrera, J. M., Pedraja Chaparro, F., & Salinas Jiménez, J. (2005). Eficiencia en educación secundaria e *inputs* no controlables: sensibilidad de los resultados ante modelos alternativos. *Hacienda Pública Española/ Revista de Economía Pública*, *173*(2/2005), 61–83.

Fadel, C., Bialik, M., & Triling, B. (2015). *Four-dimensional education. The competence learners need succeed*. Center for Curriculum Redesign. Retrieved from http://curriculumredesign.org/our-work/four-dimensional-21st-century-education-learning-competencies-future-2030/

Flamant, M. (1988). *L'Histoire du liberalism*. Paris: P.U.F.

Hattie, J. (2003). Teachers make a difference: What is the research evidence? *Australian Council for Educational Research Annual Conference on Building Teacher Quality*. October, 1–17.

Kaufmann, D., Kraay, A., &Zoido-Lobatón, P. (1999a). Aggregating governance indicators. *World Bank Policy Research*, Working Paper no. 2195. Retrieved from www.worldbank.org/wbi/governance.

Kaufmann, D., Kraay, A., & Zoido-Lobatón, P. (1999b). Governance matters. *World Bank Policy Research*, Working Paper no. 2196. Retrieved from www.worldbank.org/wbi/governance.

López Rupérez, F. (2001). *Preparar el futuro. La educación ante los desafíos de la globalización*. Madrid: La Muralla.

López Rupérez, F., García García, I., & Expósito Casas, E. (2017). *La calidad de la gobernanza del sistema educativo español. Un estudio empírico*. Madrid: Universidad Camilo José Cela.

López Rupérez, F., & García García, I. (2017). *Valores y éxito escolar. ¿Qué nos dice PISA 2015?* Universidad Camilo José Cela. Retrieved from https://www.ucjc.edu/la-universidad/estructura-academica/catedras/catedra-politicas-educativas/#pane-0-3.

Mandl, U., Dierx, A., & Ilzkovitz, F. (2008). The effectiveness and efficiency of public spending. *Economic Papers*, *301*. European Commission. Brussels: Directorate-General for Economic and Financial Affairs Publications. Retrieved from http://ec.europa.eu/economy_finance/publications.

Méndez, I., Zamarro, G., García Clavel, J., & Hitt, C. (2015). Habilidades no cognitivas y diferencias de rendimiento en PISA 2009 entre las Comunidades Autónomas españolas. *Participación Educativa*, 2ª época, 4, 51–61. Retrieved from http://ntic.educacion.es/cee/revista/n6.

Mingat, A., Tan, J. P., & Sosale, S. (2003). *Tools for education policy analysis*. Washington, DC: World Bank.

Ministry of Education, Culture and Sports (2017). *Las cifras de la educación en España. Curso 2014–2015*. Retrieved from https://www.mecd.gob.es/servicios-al-ciudadano-mecd/estadisticas/educacion/indicadores-publicaciones-sintesis/cifras-educacion-espana.html.

OECD (2001). *Measuring productivity*. OECD Manual. Paris: OECD.

OECD (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics and science (Volume I)*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264091450-en.

OECD (2013). *Résultats du PISA 2012: L'équité au service de l'excellence (Volume II): Offrir à chaque élève la possibilité de réussir.* Paris: Éditions OCDE. Retrieved from http://dx.doi.org/10.1787/9789264205321fr.

OECD (2014). *PISA 2012 results: What students know and can do. Student performance in mathematics, reading and science (Volume I, Revised edition, February 2014)*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264201118-en.

OECD (2016). *Résultats du PISA 2015 (Volume I): L'excellence et l'équité dans l'éducation*. Paris: Éditions OCDE. Retrieved from http://dx.doi.org/10.1787/9789264267534-fr.

Stevenson, H. W., & Stigler, J. W. (1992). *The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education*. New York: Touchstone.

UNESCO (2004). *Education for all. The quality imperative. EFA global monitoring report 2005*. Paris: UNESCO.

Villar, A. (coord.) (2012). *Educación y desarrollo, PISA 2009 y el Sistema Educativo Español*. Madrid: Fundación BBVA. Retrieved from https://www.fbbva.es/wp-content/uploads/2017/05/dat/DE_2012_IVIE_educacion_desarrollo.pdf.

Willms, J. D. (2006). *Learning divides: Ten policy questions about the performance and equity of schools and schooling systems* (Working Paper no. 5). Instituto de Estadística de la UNESCO, Montréal, Canada.

World Bank (1992). *World development report 1992. Development and the Environment*. New York: Oxford University Press. Retrieved from https://openknowledge.worldbank.org/handle/10986/5975.

Worthington, A. C. (2001). An empirical survey of frontier efficiency measurement techniques in education. *Education Economics*, *9*(3).

Dr. Francisco López Rupérez
Facultad de Educación. Cátedra de Políticas Educativas
Universidad Camilo José Cela
Castillo de Alarcón, 49
28692 Madrid, Spain
franciscolopezruperez@gmail.com

Prof. Isabel García García
Facultad de Educación
Universidad Camilo José Cela
Castillo de Alarcón, 49
28692 Madrid, Spain
isabel.garciagarcia@gmail.com

Dr. Eva Expósito-Casas
Facultad de Educación
Universidad Nacional de Educación a Distancia
Calle Juán del Rosal, Nº14
28040 Madrid, Spain
evaexpositocasas@edu.uned.es

The Relationship between Students' ICT Use and Their School Performance: Evidence from PISA 2015 in the Czech Republic

Libor Juhaňák, Jiří Zounek, Klára Záleská, Ondřej Bárta, Kristýna Vlčková

Masaryk University, Faculty of Arts, Department of Education Sciences

Abstract: In the last decades, Information and Communication Technologies (ICT) have become recognized as an important and integral part of life as well as education. At the same time, the implementation and use of ICT in schools is one of the longstanding strategic objectives and priorities in education policy in the Czech Republic. However, up to now, rather little attention has been paid to the research in the use of digital technologies in Czech schools with regard to students' performance. The purpose of the present study is therefore to investigate various ICT-related factors associated with school performance of students in the Czech Republic. Specifically, this study takes data from the latest Programme for International Student Assessment (PISA 2015) to determine the extent to which availability and use of ICT in school and at home is related to students' educational achievements. Results of this study can provide substantial implications and suggestions for national ICT policies (especially the Strategy for Digital Education until 2020).

Keywords: ICT availability; ICT use; school performance; PISA 2015

Information and communication technologies (ICT)[1] are undoubtedly one of the key elements in current education. The importance of ICT for education in the Czech Republic (CR) has been declared in current strategic documents, not only within the *Strategy for digital education until 2020* (MŠMT, 2014) but also for instance in the concept *Digital Czechia v. 2.0, a pathway to digital economy* (MPO, undated document). The latter document declares that the state perceives the inevitability of ICT becoming integrated in the whole process of learning at primary schools and in all individual subjects. In an overwhelming majority of schools in CR, however, digital technologies are already playing an important role not only in teaching and learning but also in everyday school bureaucracy (ČŠI, 2017b). Last but not least, ICT in education is perceived as important also in current pedagogical research. Evidence of this was provided for example by the 2017 conference of the Czech Education Research Association, whose theme was the influence of technologies in education and educational research (Michek, Vondroušová, & Vítová, 2017). Another example

[1] The term is used here to refer to any technologies and technological tools enabling communication and working with information in an electronic form, see e.g. Zounek and Šeďová (2009). In the context of this study, similar terms such as *information technologies*, *digital technologies* or *modern technologies* are treated as synonyms.

https://doi.org/10.14712/23363177.2018.292
www.orbisscholae.cz

is the 2015 specialized issue of *Pedagogika* focusing on ICT in education and the 2018 specialized issues of *Pedagogika* and *Studia paedagogica*.[2]

Despite this, digital technologies in education have not been considered as a factor enough in CR in the long run, neither in educational research nor in educational policies. As Zounek and Tůma (2014) have shown in their analysis focusing on four main Czech journals of educational research,[3] these journals published only nine empirical studies dealing with ICT between 1990 and 2012. The same is true of educational policies as here, too, systematic monitoring of the operation of the Czech educational system – to be undertaken through research and/or evaluation activities at the national level – is missing (see Potužníková, Lokajíčková, & Janík, 2014; Straková, 2009). The situation in ICT is similar, as has been remarked even in the *Strategy for digital education* (MŠMT, 2014), where insufficient research in and monitoring of the implementation of digital technologies in education represents one of the key topics. Therefore, although there is no doubt that digital technologies have been influencing Czech education for a rather long period of time, their influence on education and the educational system has paradoxically been under-researched.

This study therefore aims to contribute to a better understanding of the role of ICT in student learning and education; more specifically, we are focusing on understanding the relationship between availability and use of digital technologies (both in school and at home) and students' school performance. Results of such analysis may yield important scholarly knowledge as a contribution to the professional debate on the influence of digital technologies on student learning. Also, our findings can partly be regarded as feedback or evidence for educational policies to rely on in making decisions concerning future directions in ICT use and implementation in education as well as in planning research and/or evaluation activities at the national or international level.

1 ICT in Czech educational policies and research

This part of the paper will first map ICT in Czech educational policies and then move on to the current situation in ICT availability and use by students at school and at home as seen through the lenses of national and international surveys. The last part of the chapter will provide an outline of research focusing directly on the link between ICT availability and use and students' school achievements.

1.1 Czech educational policies and ICT

Educational policies paid attention to integration of "modern" technologies in education as early as in the 1980s (in the socialist Czechoslovakia of then), when in 1985 the strategic document titled *Long-term comprehensive programme of elec-*

[2] For more details see pages.pedf.cuni.cz/pedagogika, www.studiapaedagogica.cz.

[3] These were *Orbis scholae*, *Pedagogická orientace*, *Pedagogika* and *Studia paedagogica*.

tronization in education and upbringing in the educational system was adopted by the government (Caha, 1986; Zounek & Šeďová, 2009). The Velvet Revolution of 1989 put an end to any activities associated with this document. In the 1990s a clearly formulated national educational policy or vision for future development to accentuate the issue of modern technologies was missing, despite the fact that the 1990s were a period when ICT were becoming important not only in education.

Educational policies turned their attention to ICT as late as at the turn of the millennium, with the government approving the *Concept of National Information Policy in Education* in April 2000. The *Concept's* goals, beyond equipping schools with computers and connecting schools to the internet, were educating teachers in using ICT and developing digital (electronic) educational programmes and information resources. The process of implementation of the *Concept* was however lagging behind the plan from its very first year, indicating that the worry that the project was focusing too much on technologies themselves and neglecting their integration in classroom activities or student learning was justified (Punar, 2008). In spite of that, this was a period when information technologies became reality in most Czech schools. In 2008, a strategic document called *Developmental Strategy on ICT in Education for 2009–2013* was created, with the goal of initiating again and setting up centralized support to the implementation of digital technologies in education. As early as in 2009 it however turned out that due to the financial possibilities of and situation in the Ministry of Education the proposed programme could not be implemented as planned. Introducing digital technologies to schools nevertheless went on to some extent. Evaluation of the implementation of digital technologies in schools and the efficiency of means expended however remained entirely unsatisfactory (see MŠMT, 2014).

The latest educational policies document dealing with ICT so far has been the *Strategy for digital education until 2020*, which sets out three principal goals: 1) open up education to new methods and ways of learning mediated by digital technologies; 2) improve students' competencies in using information and digital technologies; and 3) develop computational thinking in students (see MŠMT, 2014). These goals should be achieved through a set of measures structured within seven intervention directions, including the following ones: setting up a non-discriminating approach to digital educational resources, guaranteeing conditions for developing digital literacy in students and teachers, building up educational infrastructure and supporting innovative approaches and increasing public informedness regarding educational technologies. What can be regarded as a crucial statement is the explicit acknowledgment of systematic data collection and monitoring of the current state of implementation of digital technologies in education (including educational research).[4] The existing evaluations of the strategy in progress (see MŠMT, 2017, 2018)

[4] This has been addressed by several measures, primarily Measure 5.2 (*Support to educational research of the use of digital technologies*), Measure 5.3 (*Support to regular data collection, situation monitoring and use of digital technologies in education*), and Measure 5.4 (*Improving*

unfortunately suggest that the implementation of some measures is, again, lagging behind the plan.

It can be summarized that the implementation of ICT-supporting activities in the Czech environment has been considerably non-systemic and irregular (Schoolnet, 2015) and evaluation activities and monitoring so far have been entirely unsatisfactory. Positive trends however also merit a mention: Czech educational policies have gradually transitioned in terms of their priorities from emphasis on providing the largely technological infrastructure for schools to developing teacher education and student competencies within ICT and, above all, support to evaluation and research of ICT use in education (see the first explicit mention in the *Strategy for digital education until 2020*).

1.2 An overview of current situation in ICT in Czech schools and lives of students

As has been suggested in the introduction, this study focuses on ICT availability and use by students in the school environment as well as at home. Focusing our attention first on schools, the most up-to-date information on equipment and use of digital technologies in Czech schools can be drawn from the specialized report of the Czech School Inspectorate (CSI) from September 2017 (ČŠI, 2017b). It presents results of inspection activities of the Inspectorate focusing on identifying conditions for using digital technologies and it has also been included in the Inspectorate's annual report published at the end of 2017 (ČŠI, 2017a). The inspection activities were carried out throughout the school year 2016/2017 through an on-line questionnaire filled in by headmasters of all kindergartens, all primary schools, all secondary schools and higher vocational schools. The specialized report dealt, among other things, with physical and personnel prerequisites for working with digital technologies in schools and with teaching with the support of ICT. In connection with the *Strategy for digital education until 2020*, part of the data on the use of digital technologies in Czech schools has been made available in the form of so-called open data.[5]

The data presented in the specialized report is, unfortunately, rather brief,[6] not providing a comprehensive overview of the current state in the Czech Republic. It nevertheless offers some interesting, even though basic, information. It for instance turns out that ICT have been an everyday part of the life of virtually all schools in CR, with 99% of big primary schools,[7] secondary schools and higher vocational schools

information and knowledge-base in the use of digital technologies, developing digital literacy and thinking in line with information science). For more details, see MŠMT (2014).

5 See website *Statistická data o ICT ve školách v podobě otevřených dat* [Statistical data on ICT in schools in the open-data form] accessible through the website of the Ministry of Education (www.msmt.cz/vzdelavani/skolstvi-v-cr/statistika-skolstvi/otevrena-data).

6 Even when compared with the Czech School Inspectorate's annual report for 2011/2012 (ČŠI, 2013), where ICT are paid considerably more attention. It is also necessary to bear in mind that CSI does not conduct scientific research.

7 I.e. primary schools with over 150 students.

using an information system of some kind to deal with the administrative agenda and having a webpage of their own. Small primary schools and kindergartens are less well equipped with and less good at using technologies but nevertheless around 90% of kindergartens and smaller primary schools have a website of their own and an information system. Another rather positive fact is that most schools have an ICT development plan which they subject to updates (ČŠI, 2013, 2017b), which might play a positive role in how often and how teachers use ICT during classroom exposure.

Although the equipment and infrastructure may seem sufficient for the present time, the relatively fast ageing proves to be a problem. The Czech School Inspectorate even warns in its specialized report from 2017 quoted above that the share of schools with dated technologies has been growing considerably, which may consequently mean a deterioration of the prerequisites for ICT-supported teaching.[8] Another problem is insufficient personnel basis for ICT-supported teaching, with the position of an ICT administrator often missing (it exists in 17.8% kindergartens and 35.1% of primary schools). This means, among other things, that digital technologies administration in schools is often the responsibility of the ICT teacher or ICT coordinator, leaving them with less time for their own work – coordinating ICT in the school and providing methodological support for teachers. The Czech School Inspectorate thus concludes its report with a rather alarming statement that the minimal standards of quality of conditions for using digital technologies have been met by 5% of small primary schools, nearly 10% of big primary schools and approximately 20% of secondary schools and higher vocational schools (ČŠI, 2017b).

Turning our attention to the accessibility of ICT in home environment for Czech 15-year-old students, outputs by the Czech Statistical Office (CSO) may be used, focusing generally on how Czech households are equipped with digital technologies, or ILSA (International Large-Scale Assessments) findings – covering also the Czech Republic and providing necessary information[9] – may also be used. According to the CSO, in 2017 there was a PC in 94.6% of households with children younger than 15 and 95.9% of households with children could access the internet. The PISA 2012 survey (OECD, 2015) arrived at similar results. The ensuing report states that in CR, more than 98% of students could use a PC at home in 2012 and slightly less (97.4%) could access the internet from home. It may therefore be inferred that the basic level of ICT availability in home environment is more or less universal for Czech students. Differences in ICT availability based on socioeconomic status do still exist but the problem seems to be less pronounced than in the past (Basl, 2010; OECD, 2005).

[8] E.g. according to results from 2009, more than a half of computers (56%) for classroom use were younger than 5 years; now less than 10% meet this condition (ČŠI, 2009, 2017b).

[9] These were primarily ICLIS and PISA surveys. ICILS survey is coordinated by IEA, focusing on computer and information literacy. Its most recent run was in 2013, with students of Grade 8 of primary schools and corresponding grades of 8-year and 6-year gymnasiums as the tested group. The PISA survey is implemented by OECD and besides the main area of focus, measurement of educational outcomes, it partly also focuses on issues such as ICT in education. The most recent survey was conducted in 2015, with 15-year-old students as the tested group. For more information on international research in education see e.g. Soukup (2012), Basl (2014) or Straková (2016).

To conclude this section, we will focus on how 15-year-old students in CR use digital technologies, not only at school but also beyond school. There is, however, a shortage of data for CR coming from national (topic-related) research and only results from international ICILS and PISA surveys can thus be used. Those suggest that within an international comparison, Czech students are generally among the frequent ICT users, both at school and at home within their leisure-time activities. The time spent online outside of school does not widely differ depending on students' socioeconomic status (ČŠI, 2016).[10] At the same time, ICT use in school has been growing over the recent years. For all types of activities falling within the index of ICT use at school,[11] the OECD averages for 2009 to 2012 have shown a growing trend in terms of students reporting they were involved in the activity at least once per week (OECD, 2011, 2015). PISA 2012 survey was the first to focus on how much time 15-year-old students spend online, differentiating between school and home and between using ICT on weekdays and over the weekend. According to PISA 2012, approximately 36% of Czech students spend four or more hours online, which is a value exceeding the international average. Even during the working days, Czech students spend more time online in their homes than the average OECD value is; a comparison with PISA 2015 data shows that the amount of time spent online has recently been growing (OECD, 2015, 2017a). In contrast, Czech students spend less time online at school compared with the OECD average.

1.3 ICT in connection to students' school performance

To shift attention to ICT in connection to students' educational outcomes, in CR only some information is available, namely only information based on PISA survey data. It is worth mentioning for instance PISA 2006 secondary analysis (Kubiatko & Vlčková, 2010), dealing with the relationship between ICT use by students and their performance in science. The authors of the analysis have found a positive relationship between ICT use and knowledge of Czech 15-year-olds in science. For instance, students who used PCs for educational activities more often, performed better. Similarly, the longer the experience of using a PC, the better the students performed. By contrast, the results of the secondary analysis by the Czech School Inspectorate using PISA 2012 data (ČŠI, 2016) were less optimistic. The analysis focused on ICT use in school and it turned out that both for primary schools and for 8-year and 6-year gymnasiums (i.e. general secondary schools of the lyceum type) it was true that students in less successful schools used ICT more widely, and, conversely, that students from very successful schools used ICT the least. The international PISA 2012 survey (OECD, 2015) reached similar conclusions, also finding rather a negative relationship

[10] Certain differences however concern activities pursued. Students with lower socioeconomic status report activities such as acquisition of practical information online or reading online less often while the frequency of activities such as gaming does not seem to be influenced by the socioeconomic status of the student's family.

[11] For more information on the index see the section on methodology (below).

between ICT use (at school) and students' performance. Mathematical and reading skills are generally lower in countries with higher shares of students using ICT in school. PISA results also suggest that students in countries focusing on introducing PCs into schools between 2003 and 2012 more than in other countries performed less well than students elsewhere. Overall the results of PISA 2012 (OECD, 2015) may be summarized by saying that as for the effect of ICT on student performance, there is a negative rather than positive (even though rather weak in some cases) relationship between the use of ICT and learning outcomes.

In the Czech Republic, unfortunately, more detailed research of the effect of ICT use on students' outcomes is missing. One may resort to results of international research dealing with the issue but it never fully considers the context of the Czech Republic, and cannot therefore replace missing national research fully. Research focusing on students' school performance has, for instance, shown repeatedly that the Czech educational system is characterized by a relatively strong link between the socioeconomic status and cultural capital of the family on the one hand and the student's school performance on the other (Matějů & Straková, 2006; Matějů, Straková, & Veselý, 2010; Potužníková et al., 2014; Straková, 2009). This means that children from families with higher cultural and socioeconomic status perform better. The Czech educational system is thus not sufficiently capable of levelling out the differing input potential students are carrying over from their home environments. This is also connected to the big gap in student performance between different types of school. While students studying in Czech gymnasiums score among the best in the international comparison, students in vocational training score among the worst in international comparisons and often are unable to achieve even the basic qualification level (Matějů et al., 2010; Palečková, Tomášek, & Basl, 2010; Sucháček, 2014). Moreover, this gap seems to deepen (Straková, 2010). Gender also proves an important factor influencing learning outcomes of students in CR. Girls in CR are generally better at reading while boys are better in maths and sciences (Potužníková et al., 2014). At the same time, as Matějů and Simonová (2013) show, girls are at an advantage in CR to some extent as, for instance, they achieve better grades in maths despite having less good mathematical skills than boys according to PISA results. This and other Czech specifics related to students' school performance may also be reflected in whether and how the availability and use of ICT relate to student achievements.

International research addressing ICT in education and specifically in connection to school performance is relatively abundant. Clear answers concerning the effect of ICT on learning outcomes (whether positive or negative) are however rather scarce. The great heterogeneity and even contradictory nature of the results may be ascribed to the varying focus of the studies, the considerable complexity of the topic of ICT in education or the application of a wide range of methodologies (Biagi & Loi, 2013; Cox & Marshall, 2007). Fundamental lack of clarity besets even very basic questions concerning the effect of ICT on learning outcomes. The research includes studies finding positive effect of ICT on learning outcomes (Erdogdu & Erdogdu,

2015; Spiezia, 2010) as well as studies finding the effect to be negative (Leuven, Lindahl, Oosterbeek, & Webbink, 2007). Other studies find the effect to be non-existent (Falck, Mang, & Woessmann, 2017; Wittwer & Senkbeil, 2008) or present mixed results (Biagi & Loi, 2013; Comi, Argentin, Gui, Origo, & Pagani, 2017; Luu & Freeman, 2011; Ponzo, 2011; Skryabin, Zhang, Liu, & Zhang, 2015).

Focusing on ICT availability first, for instance Erdogdu & Erdogdu (2015) may be mentioned, who come up with the finding that the availability of the internet, whether at home or in school, has a positive effect on learning outcomes of students. However already an older study by Woessmann & Fuchs (2014) has shown that the relation between ICT availability and school performance may turn the other way round as soon as other relevant factors are taken into consideration. In their study, including variables concerning family background and school characteristics resulted in turning the originally positive correlation into a negative one for PC availability at home and into a non-significant correlation for PC availability at school. As for ICT use by students, Ponzo (2011) has identified a significant positive relationship between students' learning outcomes and the frequency of using the PC as an educational tool at home but mentions also the negative effect of PC use in school on learning outcomes. Biagi & Loi (2013) also present mixed results, finding a positive relationship of some ICT-based activities with learning outcomes but a negative relationship for other activities. All of the above stated shows that the correlation between ICT availability and use, whether at school or at home, is not straightforward and may be influenced by a number of other factors. This presents the obvious requirement for researchers to be aware of the complexity of this relationship and choose appropriate analytical procedures. This also indicates with increasing urgency the need to rely not only on quantitative indicators of ICT use (see Lei, 2010) but monitor and consider in analyses other relevant ICT-related factors and variables.

2 Research problem

As has already been stated, modern technologies are a topical issue in the context of Czech educational policies. Despite this, many questions in ICT-in-education research and monitoring remain unanswered. Even data concerning the use of ICT obtained within International Large-Scale Assessments remain largely unaddressed, the reports and secondary analyses published by the Czech School Inspectorate being a rather rare exception. Lack of clarity also characterizes the international research in digital technologies in education, where the findings of research in the effect of technologies on students' learning outcomes tend to be mixed and contradictory. This study therefore aims to reach a better understanding of how ICT availability and use by Czech students in school and outside of school is reflected in their school performance. We would like to contribute to a better understanding of these issues, lessening the "white spots" in the map of this topic.

Availability of ICT to students is naturally influenced both by the family environment and the environment at school. Therefore, we want to focus on access to modern technological equipment in schools as well and we want to find out whether the level of this equipment plays a role in students' learning outcomes. One of the key elements of the use of digital technologies in contemporary society is using the internet. The analysis will therefore specifically focus also on time spent by 15-year-olds online. Finally, we seek to determine how students' learning outcomes reflect their interest in modern technologies and their perceived competence and autonomy in using them.

We have therefore formulated the individual research questions as follows:

1) To what extent is the availability of ICT to students in school and at home related to their school performance?

1a) To what extent is the level of ICT equipment in schools related to students' school performance?

2) To what extent is the use of ICT by students in school and at home related to their school performance?

2a) To what extent is the frequency of using the internet by students in school and at home related to their school performance?

3) To what extent is students' interest in using ICT related to their school performance?

4) To what extent is students' perceived autonomy and competence in ICT usage related to their school performance?

3 Data and methodology

3.1 Sample and procedure

The analyses are based on data from PISA 2015 (Czech dataset), specifically data concerning students' results in tests of mathematical (MATH), reading (READ) and science (SCIE) literacy, data from the student questionnaire (primarily the ICT Familiarity Questionnaire) and data from the school questionnaire.

The final sample contains data from 6812 students aged 15 to 16 (range = 15.3–16.3, M = 15.8, SD = 0.28, 49.7% of girls) from 333 schools.[12] In terms of schools, the data was collected in 144 primary schools (PS), 53 secondary vocational schools without maturate (SVS), 56 secondary technical schools with maturate (STS), 44 8-year and 6-year gymnasiums (G8–6) and 44 4-year gymnasiums (G4). The school data included 88.6% of state-funded schools and 9.6% of private or church-funded schools. For 6 schools in the dataset this piece of information was missing.

[12] The analysis excluded 82 students from 11 practical and special schools.

3.2 Measures

To answer the research questions formulated above, the research based the proxy indicator of school performance of students (i.e. the dependent variable) on students' performance in tests of mathematical, reading and science literacy. The analysis was conducted separately for mathematical, reading and science literacy, as the analyses have shown that there are certain differences as to the influence of the ICT-related variables on student performance between the individual areas.

Let us also remark that in PISA 2015 data, student performance in these areas is represented by 10 plausible values each (compared with only 5 plausible values in the previous runs). This fact was taken into account in an appropriate way by the analyses.[13] Weight coefficients were also applied so that the calculations are correct with respect to the nature of PISA 2015 data.

3.2.1 ICT availability

As for ICT availability from student perspective, students' answers in the ICT Familiarity Questionnaire provide two indexes: one reflecting ICT availability in school (ICTSCH) and the other measuring ICT availability at home (ICTHOME).[14] Questions of both kinds were formulated asking about selected devices and their availability at home or in school. The list of devices in both kinds of questions included options such as PC, laptop, mobile phone, or USB flash disk. Some devices were only listed in connection with school (such as the interactive whiteboard), others only in connection with home (such as the PlayStation). For each of the selected devices, students were choosing from among these options: *Yes and I use it; Yes but I don't use it* and *No*. The resulting index was calculated as a sum of the component items.

The questionnaire administered to school headmasters (school questionnaire) within PISA 2015 survey included several questions concerning ICT equipment available in the school (questions SC004Q01 through SC004Q07). Our research in ICT availability to students therefore also focused on the level of ICT equipment in school. We monitored five variables:

- Number of PCs per student. The variable was calculated using answers to question *How many computers are approximately available to these students in your school for their learning?*; the expression 'these students' refers to the previous question, focusing on the number of students in the grade under analysis, i.e. grade 9 in primary schools, grade 1 in secondary schools and 4-year gymnasiums and the corresponding grade in 8-year and 6-year gymnasiums. The number of computers listed thus did not have to correspond to the overall number of computers in the school; the question targeted the number of computers available to

[13] As Soukup (2016) or Straková (2016) claim, the analyses have to be made for each of the plausible values separately, then calculating a mean value (the standard error is calculated from the variance of the individual values).

[14] The abbreviation in the brackets is the reference to the given variable in PISA 2015 dataset. For more information on how individual indexes or scales were constructed see the OECD Technical Report for 2015.

students in the grade under analysis. The construction of the variable excluded several instances of extreme values where the assumption was that the headmaster filling in the questionnaire considered the whole school instead of the grade in question.

- Number of portable PCs per student. This variable was similar to the one described above, the difference being that only portable computers (laptops) were to be considered. In this case, too, several extreme values were excluded from the dataset.
- Number of PCs per teacher. The variable was constructed using answers to the question *How many internet-connected PCs are available in your school to teachers?* The answers were related to the total number of teachers with a full-load employed by the school (question SC018Q01TA01 in the questionnaire).
- Number of interactive whiteboards in the school. The variable was based on the question *How many interactive whiteboards are available in your school?* (without further adjustments).
- Number of data projectors in the school. The variable was based on the question *How many data projectors are available in your school?*, again without any further adjustments.

Let us remark that due to the distribution of these five variables not being normal, each of the five variables was subjected to logarithmic transformation before being included in the model.

3.2.2 ICT use

PISA 2015 also measured ICT use by students both in school and at home. In domestic environment, it was further differentiated between ICT use in connection to school (i.e. primarily to prepare for classes) and ICT use for enjoyment and/or in one's leisure-time.[15] This provided us with three indexes. The corresponding questions in the questionnaire focus on the frequency of using electronic devices, students choosing from among the following options: *Never or hardly ever; Once or twice in a month; Once or twice a week; Almost every day; Every day.* The individual indexes have been constructed using IRT modelling (OECD, 2017d).

- Students' use of ICT at school (USESCH). Examples of activities students responded to are, for instance: *Chatting online at school*; *Playing simulations at school* or *Using school computers for group work and communication with other students*. The item reliability of the index for CR is 0.887 (Cronbach's alpha).
- Students' use of ICT outside of school for school work (HOMESCH). Examples of activities are: *Browsing the Internet for schoolwork*; *Using email for communication with other students about schoolwork* or *Doing homework on a computer.* The item reliability of the index for CR is 0.901.

[15] We believe that this differentiation should be applied also when ICT use in school is studied, as the use by students in school itself does not guarantee that ICT are used primarily for school purposes.

- Students' use of ICT outside of school for leisure activities (ENTUSE). Examples of activities are: *Chatting online*; *Browsing the Internet for fun*; *Playing online games via social networks* etc. The item reliability of the index for CR is 0.810.

Besides the above-listed indexes, we have analysed data from questions focusing specifically on the frequency of using the internet. These are three questions in the questionnaire asking about the time usually spent by students online (IC005Q01TA, IC006Q01TA, IC007Q01TA). The first one focuses on the school environment, the second on the home environment during an average weekday, and the third one on the home environment during the weekend (average Saturdays and Sundays). With all three questions, students select from among seven options: *No time*, *1–30 minutes per day*; *31–60 minutes per day; Between 1 hour and 2 hours per day; Between 2 hours and 4 hours per day; Between 4 hours and 6 hours per day* and *More than 6 hours per day*.

For the purposes of the analysis the number of categories for each variable was reduced to 5, considering the number of cases within each category (and joining primarily those categories which contained few cases). For the variable characterizing the use of internet in school, the following five levels were distinguished in the final analysis: *No time*; *1–30 minutes per day; 31–60 minutes per day; Between 1 hour and 4 hours per day* and *More than 4 hours per day* while with variables concerning time spent online at home, the following levels were distinguished: *Between no time and 30 minutes per day*; *Between 31 minutes and 2 hours per day*; *Between 2 hours and 4 hours per day*; *Between 4 hours and 6 hours per day* and *More than 6 hours per day*.

3.2.3 ICT interest and ICT in students' social life

PISA 2015 survey measured interest in ICT in general as well as to what extent ICT was integrated into the lives and social interactions of 15-year-old students. In both cases, the question in the questionnaire was exploring the degree of agreement or disagreement with selected statements concerning students' interest in ICT. Students scored each statement on a four-point Likert scale, the options ranging from *Strongly disagree* to *Strongly agree*. Let's add that PISA 2015 was the first occasion for these questions to be used in a PISA survey.[16]

- Students' ICT interest (INTICT). Examples of statements students responded to are: *I like using digital devices*; *I am really excited discovering new digital devices or applications* or *I really feel bad if no internet connection is possible*. The item reliability of the index for CR is 0.775.
- The degree to which ICT is a part of students' daily social life (SOIAICT). Examples of statements students responded to are: *I like to share information about digital devices with my friends* or *To learn something new about digital devices, I like to talk about them with my friends*. The item reliability of the index for CR is 0.880.

[16] The previous PISA run, i.e. PISA 2012, in contrast, included two sets of questions (and two indexes) focusing on attitudes towards computers (computer as a tool for school learning) (OECD, 2014).

3.2.4 ICT competence and autonomy in ICT usage

Perceived competence and autonomy in ICT use was the focus of the another two questions newly used in PISA 2015. Both questions had the same form as the above described questions focusing on interest in ICT (i.e. respondents expressed their interest or lack of interest using a four-point Likert scale).

- Students' perceived competence in ICT usage (COMPICT). Examples of statements students responded to are: *I feel comfortable using digital devices that I am less familiar with* or *If my friends and relatives have a problem with digital devices, I can help them.* The item reliability of the index for CR is 0.858.
- Students' perceived autonomy related to ICT usage (AUTICT). Examples of statements students responded to are: *If I need new software, I install it by myself* or *If I have a problem with digital devices I start to solve it on my own.* The item reliability of the index for CR is 0.821.

Table 1 Basic descriptive statistics for continuous variables at the student level.

	N	Min	Max	*M*	*SD*	Skewness	Kurtosis
ICTSCH	5800	0.00	10.00	5.68	2.07	0.02	0.17
ICTHOME	5973	0.00	11.00	8.28	1.72	−0.52	0.52
USESCH[1]	6330	−1.67	3.63	0.27	1.03	0.46	1.84
HOMESCH[1]	6239	−2.69	3.60	0.13	0.95	0.39	4.35
ENTUSE[1]	6382	−3.71	4.84	0.17	1.03	1.14	7.60
INTICT[1]	6304	−2.99	2.82	−0.14	0.94	0.66	1.98
SOIAICT[1]	6260	−2.14	2.43	−0.09	1.01	0.32	0.55
COMPICT[1]	6262	−2.66	1.97	−0.1	0.96	0.25	0.28
AUTICT[1]	6295	−2.50	2.10	−0.09	0.95	0.50	0.56
ESCS[1]	6716	−3.01	3.49	−0.19	0.79	0.11	−0.23
MATH[2]	6812	153.74	801.74	495.93	88.21	−0.03	−0.18
READ[2]	6812	91.48	879.05	490.94	98.12	−0.17	−0.25
SCIE[2]	6812	139	823.97	495.84	93.87	0.04	−0.39

[1] The variable is conceived in such a way that the mean value across OECD countries is 0 and the standard deviation is 1.
[2] The variable is conceived in such a way that the mean value across OECD countries is 500 and the standard deviation is 100.

Let us conclude this section by noting that the analyses and modelling used also some other variables, mainly as control variables. These included especially the index of economic, social and cultural status (ESCS),[17] which was used both at the student level and at the school level (ESCS – L2). At the student level the analyses

[17] The construction of the index of economic, social and cultural status in PISA surveys is a rather complex issue; for a better insight into how the index is constructed see OECD (2017) or Appendix 4, Indexes and scales, in Blažek & Boudová (2017).

considered also gender of students (coded by effect coding) and at school level the analyses considered the type of school and an indicator differentiating between state-funded and private schools (PRIVATE). The school type has been coded by dummy coding, with primary school serving as a reference category. The variable PRIVATE has been coded by dummy coding, with state-funded schools serving as a reference category.

3.3 Data analysis

Since this study is working with hierarchical data (i.e. students are nested within schools), the analyses are based on multilevel modelling[18]. This is a method increasingly used in recent years not only in educational sciences and in connection with data from ILSA, but also in other disciplines such as sociology, psychology and others (see Hox, 2010; Snijders & Bosker, 2012; Heck & Thomas, 2015). The method has not yet found significant application in Czech educational research; a more detailed introduction has been provided especially by Soukup (2006).

Our modelling followed the recommended general strategy proposed by Heck and Thomas (2015) and mentioned by Soukup (2006) while the analysis itself was carried out in statistical environment R (R Core Team, 2017), especially using the BIFIEsurvey package (BIFIE, 2017). As has already been mentioned, the analyses were conducted for each of the areas separately (mathematical, reading and science literacy). The first step involved creating a so-called nullmodel and then a so-called baseline model including only fundamental variables commonly used to explain differences in school performance of students. Only then the models were enriched by ICT-related variables relevant to the research questions, each modelling step only preserving those variables that proved to be statistically significant.

4 Results

First for each area of analysis a nullmodel was created as a basis for calculating the intra-class correlation coefficient (ICC). ICC for mathematical literacy was 0.401, meaning that approximately 40.1% of the variance of students' school performance can be attributed to differences between schools.[19] This ICC can be regarded as relatively high, which is in line with the well-known fact that the educational system in the Czech Republic is rather strongly stratified and inter-school differences are relatively high. This has been evidenced by the conclusions of the Czech national report from the most recent survey (Blažek & Příhodová, 2016), which also shows

[18] International research also refers to this type of analysis as *multilevel regression models*, *hierarchical linear models*, *mixed-effects models* or *random-coefficient models* (see Heck & Thomas, 2015).

[19] The remaining variance of student performance, i.e. 59.9%, is due to inter-student differences.

differences in school performance between schools in CR as above-average compared with other OECD countries.[20] Let us say for the sake of completeness that the remaining two areas have similarly high ICCs. 40.7% and 40.3% of the variance can be attributed to inter-school differences in reading literacy and science literacy, respectively.

4.1 Basic overview of school performance for Czech students

Table 2 presents the parameters of the baseline two-level models explaining students' performance in the individual areas tested using gender and socioeconomic status at the student level and socioeconomic status at the school level, differentiating between state-funded and private-funded schools of different types. Reflecting the coding used for categorical variables included in the model (see section 3), the constant for each model corresponds to the mean performance of students in the given area at a state-funded basic school in CR.

Table 2 Baseline models using basic variables related to the performance of Czech 15-year-olds in tests of mathematical (M), reading (R) and science (S) literacy.

	Model 1 (M)	**Model 1 (R)**	**Model 1 (S)**
Fixed effects	***coef. (SE)***	***coef. (SE)***	***coef. (SE)***
Intercept	487.82 (3.64) ***	486.55 (4.73) ***	488.81 (4.15) ***
ESCS	22.47 (1.74) ***	21.89 (2.45) ***	20.87 (1.8) ***
GENDER[1]	−7.68 (1.42) ***	8.39 (1.73) ***	−8.38 (1.41) ***
ESCS (L2)	27.88 (6.35) ***	36.5 (7.79) ***	32.49 (6.94) ***
PRIVATE	−18.49 (8.08) *	−21.44 (8.4) *	−23.41 (8.07) **
G8−6	82.52 (8.2) ***	74.65 (9.51) ***	88.2 (8.29) ***
G4	73.52 (7.6) ***	77.73 (7.91) ***	76.92 (7.53) ***
STS	29.56 (5.19) ***	32.28 (6.39) ***	28.27 (5.4) ***
SVS	−35.19 (7.43) ***	−44.76 (7.65) ***	−37.28 (7.14) ***
Random effects	***Variance component***	***Variance component***	***Variance component***
Residual variance	4085.3 (94.2) ***	5118.3 (90.3) ***	4679.1 (52.5) ***
Intercept variance	420.1 (80.7) ***	601.1 (105.6) ***	521.1 (92.2) ***
Explained proportion of variance			
At the student level	0.088	0.063	0.072
At the school level	0.866	0.843	0.853

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$
[1] effect coding

[20] Since special and practical schools were excluded from the analyses, inter-school differences are somewhat lower compared with the above-mentioned national report.

Using the example of performance in mathematics, the coefficient of socioeconomic status at the student level may be interpreted by saying that increasing the index of economic, social and cultural status of a student by one point results in increasing their score in mathematical literacy by 22.47 points. The socioeconomic status at the school level is to be interpreted analogically. The effect of socioeconomic status on student performance is significant. While socioeconomic status at the student level has approximately the same effect on student performance in all areas tested, the effect of socioeconomic status at the school level is stronger for reading and science literacy than for mathematical literacy.

In the context of effect coding, which has been used, the gender coefficient means that girls score by 7.68 points worse than the average student in mathematical literacy and boys score by the same number of points better than the average. Girls are significantly better in terms of reading literacy while boys are better in terms of science literacy again. In all areas analysed, students studying at private-funded schools score significantly worse compared with students at state-funded schools. The coefficient for individual types of school reflects how much better or worse average students in the given type of school perform compared with primary schools.

We regard models in Table 2 as baseline models in the sense of including only the basic parameters commonly used to explain differences in student performance (i.e. socioeconomic status, gender and type of school). The explained proportion of variance at the student level and at the school level in these models will be regarded as reference value. The results obtained from the subsequent models, which will also include ICT-related parameters, will be compared with these baseline models. This will allow us to see to what extent ICT are a factor related to student performance beyond the basic factors included in the baseline models.

4.2 Performance of Czech students and ICT

Table 3 presents the basic ICT-related factors which turned to be significant in the individual areas of testing. We can see that in all three areas, the use of ICT by students in school (USESCH) is negatively correlated with their performance. The same holds for the index describing ICT as a part of students' everyday social life (SOIAICT). In contrast, a significantly positive relationship has been found between student performance and perceived autonomy in ICT use (AUTICT). Students who feel to be autonomous/independent in using technologies perform significantly better than other students.

In addition, both reading literacy and science literacy were significantly connected to ICT availability at home (ICTHOME) and use of ICT outside of school for school work (HOMESCH). In both cases however this relationship proves to be (perhaps surprisingly) negative. This means that students who can access ICT tools at home more easily and use them more to prepare for school perform worse in reading and science literacy.

The use of ICT by students outside of school for entertainment (ENTUSE) has not proved significant in any of the areas analysed. Similarly, no significant differences (with respect to school performance) have been recorded in terms of ICT availability in school (ICTSCH) or perceived competence in ICT use (COMPICT). It is also due to say that none of the variables concerning ICT equipment in school (i.e. numbers of PCs per student or per teacher and numbers of interactive whiteboards and data projectors in school) has proved to be statistically significant with respect to student performance in the individual areas analysed.

A comparison of the explained proportion of variance in the models including ICT-related factors with the above-described baseline models shows that the explained proportion of variance increased from 8.8% to 17% at the student level and from 86.6% to 89.9% at the school level. The difference therefore is 8.2% at the student level and 3.3% at the school level. The situation concerning reading (difference of 10% at the student level and 3.6% at the school level) and science (difference of 11.5% at the student level and 4.1% at the school level) is similar. This allows us to conclude that the connection between ICT and student performance is stronger at the individual level than at the school level. Also, the relationship between ICT and student performance is the highest in science and the lowest in mathematics.

Table 3 Basic ICT-related variables and performance of Czech 15-year-olds in mathematical (M), reading (R) and science (S) literacy (only significant ICT-related parameters are stated in the table to make it easy to read, although the models included all parameters used in the baseline models).

	Model 2 (M)	**Model 2 (R)**	**Model 2 (S)**
Fixed effects	***coef. (SE)***	***coef. (SE)***	***coef. (SE)***
Intercept	495.21 (3.07) ***	517.05 (8.58) ***	524.69 (7.73) ***
ICTHOME		−2.3 (0.82) **	−3.01 (0.75) ***
HOMESCH		−5.87 (1.75) **	−6.5 (1.64) ***
USESCH	−14.11 (1.37) ***	−15.35 (1.77) ***	−14.43 (1.65) ***
SOIAICT	−6.34 (1.68) ***	−6.06 (2.01) **	−5.07 (1.69) **
AUTICT	16.47 (1.46) ***	16.63 (2.15) ***	17.63 (1.77) ***
Random effects	***Variance component***	***Variance component***	***Variance component***
Residual variance	3752.3 (88.4) ***	4571.4 (83.3) ***	4166.4 (58.9) ***
Intercept variance	286.5 (61.9) ***	391.8 (96.2) ***	329.5 (76.2) ***
Explained proportion of variance			
At the student level	0.17	0.163	0.187
At the school level	0.899	0.879	0.894

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.3 Performance of Czech students and internet use

Table 4 presents models focusing on the use of the internet. All of them concern mathematical literacy. Compared with the previous model, the first model considers the use of the internet at school, the second the use of the internet at home on weekdays, and the third one the use of the internet during weekends. In all cases, the variable reflecting internet use was coded by dummy coding, with zero use or minimum use of the internet as the reference category.

We can see that internet use at school has a statistically significant negative effect for students whose school use of the internet exceeds one hour. Using the internet for over an hour per day at school is associated with worse performance in maths. The situation with internet use at home is however different. Students not using the internet or using it for only up to 30 minutes per day perform significantly worse than students using it for 31 minutes to 6 hours. Students using the internet at home for over 6 hours a day perform the same as students who do not use it at all or only up to 30 minutes a day. It therefore seems that excessive internet use by students (at home) has just as negative impact on their performance as zero or minimal internet use.

Due to the limited space of this article, the following table (table 4) presents only the models made for mathematical literacy. Nevertheless, the results for reading literacy and partly also science literacy were similar. The only exception was the model concerning science literacy and internet use during weekends. The differences in internet use did not prove significant in this case.

Table 4 Frequency of internet use and performance of Czech 15-year-olds in mathematical literacy (the basic parameters applied in the baseline models are not presented, even though they have been included).

	Model 3a (M)	**Model 3b (M)**	**Model 3c (M)**
Fixed effects	***coef. (SE)***	***coef. (SE)***	***coef. (SE)***
Intercept	500.41 (3.75) ***	479.93 (5.32) ***	483.07 (5.84) ***
USESCH	−11.76 (1.34) ***	−13.39 (1.31) ***	−13.85 (1.33) ***
SOIAICT	−6.68 (1.66) ***	−6.1 (1.66) ***	−6.65 (1.67) ***
AUTICT	16.95 (1.49) ***	17.03 (1.53) ***	17.08 (1.53) ***
Internet at school:			
1 to 30 minutes	−1.7 (3.34)		
31 to 60 minutes	−2.83 (4.43)		
1 to 4 hours	−10.05 (3.95) *		
over 4 hours	−29.24 (5.82) ***		
Internet at home:			
31 mins to 2 hours		23.28 (5.48) ***	
2 to 4 hours		22.12 (5.3) ***	
4 to 6 hours		13.44 (5.59) *	
over 6 hours		0.29 (6.2)	

Internet on weekends:			
31 mins to 2 hours			15.11 (6.25) *
2 to 4 hours			19.6 (6.63) **
4 to 6 hours			16.91 (6.11) **
over 6 hours			3.58 (6.02)
Random effects	***Variance component***	***Variance component***	***Variance component***
Residual variance	3683.3 (87.4) ***	3663.2 (85.1) ***	3689.8 (84.4) ***
Intercept variance	270.9 (59.1) ***	262.3 (58.9) ***	275.8 (62.6) ***
Explained proportion of variance			
At the student level	0.189	0.192	0.185
At the school level	0.904	0.906	0.902

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.4 ICT in interactions

Finally, we have analysed whether the relationship between the analysed ICT-related variables and student performance in mathematical, reading and science literacy is moderated by other variables. Due to the limited space, we only report selected results of the interaction analysis. Table 5 presents 3 different models addressing interactions with gender and type of school.

Table 5 Models with interactions (not all baseline models' parameters are presented, to keep the table easy to read).

	Model 4 (M)	**Model 5 (R)**	**Model 6 (S)**
Fixed effects	***coef. (SE)***	***coef. (SE)***	***coef. (SE)***
Intercept	495.27 (3.08) ***	515.4 (8.73) ***	525.59 (7.69) ***
ICTHOME		−2.25 (0.82) **	−3.05 (0.75) ***
HOMESCH		−6 (1.76) **	−6.43 (1.67) ***
SOIAICT	−6.54 (1.67) ***	−7.05 (2.04) **	−5.08 (1.69) **
AUTICT	15.95 (1.54) ***	14.16 (2.41) ***	17.57 (1.76) ***
GENDER	−7.18 (1.65) ***	7.78 (1.74) ***	−7.56 (1.47) ***
INTICT	0.48 (1.7)		
GENDER × INTICT	−2.36 (1.17) *		
COMPICT		3.05 (1.87)	
GENDER × COMPICT		−4.29 (1.41) **	
USESCH	−14.43 (1.35) ***	−15.36 (1.78) ***	−16.91 (2.11) ***
SVS	−27.63 (7.23) ***	−31.66 (7.41) ***	−26.5 (7.08) ***
SVS × USESCH			6.76 (2.94) *
STS	28.45 (4.76) ***	30.04 (5.93) ***	24.94 (5.02) ***

STS × USESCH			6.6 (2.51) **
G4	70.01 (7.25) ***	72.67 (7.5) ***	70.71 (6.9) ***
G4 × USESCH			7.72 (3) *
GV	76.5 (7.57) ***	64.73 (8.65) ***	78.46 (7.7) ***
GV × USESCH			−0.42 (3.98)
Random effects	***Variance component***	***Variance component***	***Variance component***
Residual variance	3735.9 (85.9) ***	4556.4 (85.2) ***	4153.9 (58.2) ***
Intercept variance	290.5 (64.9) ***	385.7 (96.1) ***	329.7 (77.4) ***
Explained proportion of variance			
At the student level	0.172	0.168	0.191
At the school level	0.897	0.881	0.894

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

It turned out above all that although initially interest in ICT (INTICT) and perceived ICT competence (COMPICT) did not seem to be significant factors, they turned out to be significant (in some cases) after including the interaction with gender. The relationship between ICT interest and student performance (Model 4) as well as the relationship between perceived ICT competence and student performance (Model 5) are significantly moderated by gender. In both cases high interest in ICT and high perceived ICT competence are associated with better results for boys and the trend tends to be opposite for girls. The decrease for girls is however milder than the increase for boys. Another interesting finding is that while the interaction between ICT interest and gender was significant only in connection with mathematical literacy (Model 4), the interaction between gender and perceived ICT competence was significant in all three areas of analysis (although Table 5 shows only the model for reading, i.e. Model 5).

The last of the models (Model 6) addresses the interaction between ICT use in school (USESCH) and type of school. Here, too, the analysed interaction proved significant, meaning that the relationship between ICT use at school and students' performance varies significantly depending on the type of school. It is true for all types of school that higher USESCH scores are associated with worse school performance, but the relationship is considerably stronger for primary schools and 8-year and 6-year gymnasiums than other types of school. This means that the worsening of school performance with increasing ICT use at school gets more serious with these two types of school.

5 Discussion

The results of the analyses allow us to conclude that neither ICT availability at school nor ICT equipment available at school seem to have a direct effect on student

performance. This finding may be regarded as a rather expectable one (also see the European Schoolnet research, 2013). As we have said in Section 1, most of the activities concerning ICT in education so far have addressed equipping schools with technologies and making sure that a basic level of availability of digital technologies will be the case in all schools. It therefore seems that a certain basic level of technology availability has been provided in Czech schools and the now existing differences in technologies available are not so pronounced any more to have a direct effect on student performance. This does not naturally mean that technological equipment in schools and ICT availability has ceased to be an important topic. The issue is, firstly, still topical due to the fast ageing of modern technologies (ČŠI, 2009, 2017b) and, secondly, technological availability is a basic prerequisite for technologies to be used in schools. Therefore, although their availability does not influence students' school performance directly, it still is an indirect influence.

One rather surprising finding concerning ICT availability is that ICT availability at home proved to be significant. It is firstly surprising that it turned out to be significant only in connection to performance in reading and science literacy but not in connection with mathematical literacy, and, secondly, because its effect on students' performance was negative. We are not able to explain this unequivocally and this topic will need to be further researched. The supplementary analyses, which unfortunately could not be fully described in this paper, however indicate that two factors might be at play. 1) The PISA questionnaire conceives the question concerning ICT availability not only in terms of its physical presence in students' homes but, to some extent, in terms of its use. 2) ICT availability at home seems to be moderated by the student's family's socioeconomic status, where for students with higher ESCS index greater ICT availability at home is related to better performance and for students with lower ESCS index with worse performance. It is therefore possible that the ICT availability index partly reflects ICT use by students while ways of ICT use by students vary depending on the socioeconomic status of the student's family. This would be in agreement with partial results of Czech School Inspectorate's secondary analysis (ČŠI, 2016), which identified certain differences in ICT use by students from families with varying socioeconomic status.

As for ICT use by students, attention was paid to their use of ICT at school and outside of school for school work and for leisure activities. As for ICT use at home, as it was to be expected, there was no significant link between ICT use in leisure time and school performance. In contrast, a rather surprising finding was the negative link between students' school performance and their ICT use at home for school work (such as doing homework using the PC etc.). Some supplementary analyses will need to be performed to explain this link. The results of the analyses made so far however do not indicate that this relationship has to do with the student's gender or socioeconomic status of their family. It is also due to say that the significantly negative effect of ICT use at home for school work was reflected only in terms of reading and science literacy.

A less surprising finding was the negative link between ICT use at school and students' school performance. Hints of this relationship occurred already in PISA data international analyses from previous years (OECD, 2015) as well as in the secondary analysis of PISA 2012 data by the Czech School Inspectorate (2016). To understand and explain the negative effect of ICT use in school on students' school performance, one primarily needs to be aware of how this index is conceived in PISA surveys. Rather than a measure of how (much) ICT are used in the given school, it is a measure of how (much) the given student uses ICT at school. It therefore reflects individual use of ICT by the student, which need not necessarily correspond to how ICT are used in the school as a whole. This is consistent with the partial results of the analyses, where the ICT use index aggregated at the school level did not seem significant. What also could play a role is the way respondents are selected as PISA surveys do not consider the level of the class (Straková, 2016). However, ICT use at school may differ significantly depending on the specific class and teacher. It has also turned out that the relationship between ICT use at school and students' school performance is significantly moderated by type of school, this negative relationship being stronger in primary schools and 8-year and 6-year gymnasiums. Thus, a broader context seems to play an important role in ICT use by students as well.

Moreover, the index of ICT use at school includes relatively general components not concerning ICT directly for school-related activities (the questionnaire, for instance, features items such as *Chatting online at school* or *Using email at school*). It is therefore easy to imagine that even students who do not pay attention to the learning content in class may have a high index of ICT use at school, chatting online with friends instead. Other items included in the index concern e.g. practising learning content or doing homework using a school PC, which are also likely to be more frequent with less successful students. Underperforming students may simply need to learn more at school (including using PC) and/or may try to catch up with homework at school after they have not given it enough time at home. This seems to be in line with the partial results of the secondary analysis by the Czech School Inspectorate (2016), according to which students from successful schools use school computers to practise learning content less. It may however also be that insufficient training of teachers for efficient ICT use in class also plays a role. This would be consistent with the results of TALIS 2013 research, which suggest that ICT skills are one of the most demanded topics in professional training of Czech teachers while the greatest proportion of Czech teachers call for more professional training in ICT skills for teaching (Kašparová, Boudová, Ševců, & Soukup, 2014). Inadequate skills among teachers were also indicated in research by Zounek and Šeďová (2009), according to which teachers tend to use ICT to reward students for working well in class and use them for their own teaching to a much smaller extent.

Besides ICT use in general, we have analysed the use of the internet by students during an ordinary day, at school, at home on a weekday, and at home on weekends. The results are, again, largely consistent with what we know from PISA 2012 survey (ČŠI, 2016; OECD, 2015). Attention so far has largely been paid to the negative ef-

fect excessive internet use has on student performance. The results of our analyses however indicate that excessive use of the internet (over 6 hours per day) at home is associated with the same – meaning equally bad – performance at school as minimum internet use (up to 30 minutes). The simple rule "The more internet at home, the worse performance at school" is thus not valid. Worse performance seems to be associated with "extreme" uses of the internet, at both ends of the range (i.e. too much and too little). The situation regarding the use of the internet at school is different: it indeed seems that the more students use the internet at school, the worse they perform. The explanation may however again be that using internet at school is a general label including not only internet use for learning but potentially conflicting with it (i.e. students may use the internet instead of paying attention to what they should be learning).

Finally, there were two variables concerning interest in ICT (i.e. INTICT and SOIAICT) and variables concerning ICT competence (COMPICT) and ICT autonomy (AUTICT) which were included in the analyses. These were however indexes newly introduced in PISA 2015, which cannot be simply related to the results from previous years. These variables have not been addressed by reports publishing results of PISA 2015 so far (see OECD, 2017a, 2017b, 2017c), therefore a basic comparison with countries included in PISA surveys is not possible. Nevertheless, focusing first on ICT competence and ICT autonomy, somewhat surprising is the strong positive relationship between students' autonomy in ICT use and their performance in tests of mathematical, reading and science literacy. This link may be due to the fact that better-performing students can generally work in more autonomous ways, which gets reflected in their autonomy in using digital technologies. Autonomy in ICT use, however, also requires some competence in using technologies; it was therefore rather surprising that the index of perceived ICT competence initially did not seem significant. A more detailed analysis however revealed that the connection between ICT competence and school performance varies significantly depending on gender. While the relationship is positive for boys (i.e. higher competence is related to better results) the relationship is mildly negative for girls.

The same is true of ICT interest, but only in connection to performance in mathematical literacy. Explanation for these findings may be found on the basis of ICILS survey results (Basl, Bird, Boudová, & Tomášek, 2015; Basl, Boudová, & Řezáčová, 2014; Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014), according to which girls did perform better at tests of computer and information literacy while their competencies in terms of advanced skills were assessed as worse. It was also boys rather than girls who expressed interest in ICT. It therefore seems that there are major differences between girls and boys in CR concerning interest in ICT, perceived ICT competencies and their effect on school performance. Last but not least, a rather surprising result is the negative correlation between the extent modern technologies play a role in everyday lives of students (SOIAICT) and school performance of students in all three researched areas. Unfortunately, in this case a comparison with other countries included in PISA surveys has not been available either nor are

more detailed analyses which could help explain the identified relationship more thoroughly.

6 Conclusions

Our study capitalizes on one of the opportunities provided by using data from ILSA to get a deeper insight into ICT in the context of the Czech educational system. We believe that the results of our analyses may yield important information for national educational policies on ICT, where the missing research and evaluation of the impact of ICT use on school education has been among the key problems.

Besides the results as such, we believe that our study illustrates also general possibilities of secondary ILSA analyses while pointing out also some of their limitations. Due to the rather broad focus of the study, its outputs include a series of new questions calling for more attention of research. Follow-up research could, for instance, address supplementary analyses of PISA data focusing on more specific questions concerning digital technologies. Another option is to use ICILS 2013 data, which has not been fully used in the Czech context yet. Our view is that these quantitative analyses should also be supplemented with qualitative research, which may mediate a better understanding of the contexts of ICT use by individuals at school and at home. Our further research therefore aims to go in this direction as well.

Last but not least it is due to mention the need for a specifically focused national survey to address not only links between ICT use by students and their school performance (measured in a traditional way). We believe that with respect to the so called competencies for the 21st century, attention should also be paid to whether and to what extent modern technologies can play a role in situations when students must combine their technological knowledge and skills with critical thinking, ability to work in teams and communicating with other actors (at school and beyond school) or come up with out-of-the-box and creative solutions to problems.

Acknowledgements

The study is one of the outputs of research project “Digital technologies in students’ everyday lives and learning”, supported by the Czech Science Agency (Grant no. 17-06152S). The authors express their thankfulness for the support.

References

Basl, J. (2010). Diferenciace v počítačové gramotnosti a nerovnosti v přístupu k informačním technologiím [Differentiation in computer literacy and inequalities in access to information technologies]. In P. Matějů, J. Straková, & A. Veselý (Eds.), *Nerovnosti ve vzdělávání: Od*

měření k řešení [Inequalities in education: From measurement to solutions] (pp. 208–226). Praha: Sociologické nakladatelství (SLON).

Basl, J. (2014). Statistika ve školství: Eurostat, OECD, PISA, IEA. In J. Hendl, et al., *Statistika v aplikacích* [Statistics in applications] (pp. 287–304). Praha: Portál.

Basl, J., Bird, L., Boudová, S., & Tomášek, V. (2015). *Mezinárodní šetření ICILS 2013: Shody a rozdíly v počítačové a informační gramotnosti mezi vybranými evropskými zeměmi* [International ICILS 2013 survey: Coincidences and differences in computer and information literacy between selected European countries]. Praha: Česká školní inspekce. Retrieved from http://www.csicr.cz/html/ICILS2013-ShodyRozdily/flipviewerxpress.html.

Basl, J., Boudová, S., & Řezáčová, L. (2014). *Národní zpráva šetření ICILS 2013: Počítačová a informační gramotnost českých žáků* [National report of ICILS 2013 survey: Computer and information literacy of Czech students]. Praha: Česká školní inspekce. Retrieved from http://www.csicr.cz/html/ICILS2013-NarodniZprava/flipviewerxpress.html

Biagi, F., & Loi, M. (2013). Measuring ICT use and learning outcomes: Evidence from recent econometric studies. *European Journal of Education*, *48*(1), 28–42. https://doi.org/10.1111/ejed.12016

BIFIE (2017). *BIFIEsurvey: Tools for survey statistics in educational assessment*. R package version 2.3-18. https://CRAN.R-project.org/package=BIFIEsurvey

Blažek, R., & Boudová, S. (2017). *Národní zpráva PISA 2015: Týmové řešení problému, dotazníkové šetření* [PISA 2015 national report: Teamwork on the problem, questionnaire survey]. Praha: Česká školní inspekce. Retrieved from http://www.csicr.cz/html/PISA_2015_NZ_reseni_problemu/flipviewerxpress.html.

Blažek, R., & Příhodová, S. (2016). *Mezinárodní šetření PISA 2015: Národní zpráva*. [PISA 2015 international survey: National report] Praha: Česká školní inspekce.

Caha, Z. (1986). Elektronizace ve výchově a vzdělávání [Electronization in education and upbringing]. *Pedagogika*, *36*(2), 133–136.

Comi, S. L., Argentin, G., Gui, M., Origo, F., & Pagani, L. (2017). Is it the way they use it? Teachers, ICT and student achievement. *Economics of Education Review*, *56*, 24–39. https://doi.org/10.1016/j.econedurev.2016.11.007

Cox, M. J., & Marshall, G. (2007). Effects of ICT: Do we know what we should know? *Education and Information Technologies*, *12*(2), 59–70. https://doi.org/10.1007/s10639-007-9032-x.

ČŠI. (2009). Úroveň *ICT v základních* školách *v* ČR: *Tematická zpráva* [ICT level in primary schools in CR: Thematic report]. Retrieved from http://www.csicr.cz/html/ICTvZS/flipviewerxpress.html.

ČŠI. (2013). *Výroční zpráva české školní inspekce za školní rok 2011/2012* [Czech School Inspectorate annual report for 2011/2012]. Praha: Česká školní inspekce. Retrieved from http://www.csicr.cz/html/vzcsi_2011_12/flipviewerxpress.html.

ČŠI. (2016). *Žáci a ICT: Sekundární analýza výsledků mezinárodních šetření ICILS 2013 a PISA 2012* [Secondary analysis of ICILS 2013 and PISA 2012 international surveys]. Praha: Česká školní inspekce.

ČŠI. (2017a). *Kvalita a efektivita vzdělávání a vzdělávací soustavy ve školním roce 2016/2017: Výroční zpráva české školní inspekce* [Quality and effectivity of the educational system in 2016/2017: Czech School Inspectorate annual report]. Praha: Česká školní inspekce. Retrieved from http://www.csicr.cz/html/Vyrocni_zprava_CSI_2016_2017/flipviewerxpress.html.

ČŠI. (2017b). *Využívání digitálních technologií v mateřských, základních, středních a vyšších odborných školách: Tematická zpráva* [The use of technologies in kindergartens, basic, secondary and higher vocational schools: Thematic report]. Retrieved from http://www.csicr.cz/html/tz_digitechnologie/flipviewerxpress.html.

Erdogdu, F., & Erdogdu, E. (2015). The impact of access to ICT, student background and school/home environment on academic success of students in turkey: An international comparative analysis. *Computers & Education*, *82*, 26–49. https://doi.org/10.1016/j.compedu.2014.10.023

European Schoolnet. (2015). *Czech Republic. Country report on ICT in education*. Retrieved from http://www.eun.org/cs/resources/country-reports.

European Schoolnet (2013). *Survey of Schools: ICT in Education. Benchmarking access, use and attitudes to technology in Europe's schools. Final Report*. Retrieved from https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/KK-31-13-401-EN-N.pdf.

Falck, O., Mang, C., & Woessmann, L. (2017). Virtually no effect? Different uses of classroom computers and their effect on student achievement. *Oxford Bulletin of Economics and Statistics*, *80*(1), 1–38. https://doi.org/10.1111/obes.12192

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age: The IEA international computer and information literacy study international report*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14222-7

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. New York, NY: Taylor & Francis.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*. New York, NY: Routledge.

Kašparová, V., Boudová, S., Ševců, M., & Soukup, P. (2014). *Národní zpráva šetření Talis 2013* [Talis 2013 survey, national report]. Praha: Česká školní inspekce. Retrieved from http://www.csicr.cz/html/TALIS2013-NZ/flipviewerxpress.html.

Kubiatko, M., & Vlčková, K. (2010). The relationship between ICT use and science knowledge for Czech students: A secondary analysis of PISA 2006. *International Journal of Science and Mathematics Education*, *8*(3), 523–543. https://doi.org/10.1007/s10763-010-9195-6.

Lei, J. (2010). Quantity versus quality: A new approach to examine the relationship between technology use and student outcomes. *British Journal of Educational Technology*, *41*(3), 455–472. https://doi.org/10.1111/j.1467-8535.2009.00961.x

Leuven, E., Lindahl, M., Oosterbeek, H., & Webbink, D. (2007). The effect of extra funding for disadvantaged pupils on achievement. *Review of Economics and Statistics*, *89*(4), 721–736. https://doi.org/10.1162/rest.89.4.721

Luu, K., & Freeman, J. G. (2011). An analysis of the relationship between information and communication technology (ICT) and scientific literacy in Canada and Australia. *Computers & Education*, *56*(4), 1072–1082. https://doi.org/10.1016/j.compedu.2010.11.008

Matějů, P., & Simonová, N. (2013). Koho znevýhodňuje škola: Chlapce, nebo dívky? Rozdíly v dovednostech, školních výsledcích a vzdělanostních aspiracích dívek a chlapců devátých tříd základních škol [Who is disadvantaged by school? Boys, or girls? Differences in skills, school performance and educational aspirations of girls and boys in Grade 9 of primary schools]. *Orbis Scholae*, *7*(3), 107–138.

Matějů, P., & Straková, J. (2006). *(Ne)rovné šance na vzdělání: Vzdělanostní nerovnosti v České republice* [(Un)equal educational chances: Educational inequality in the Czech Republic]. Praha: Academia.

Matějů, P., Straková, J., & Veselý, A. (2010). *Nerovnosti ve vzdělávání: Od měření k řešení* [Inequalities in education: From measurements to solutions]. Praha: Sociologické nakladatelství (SLON).

Michek, S., Vondroušová, J., & Vítová, J. (Eds.). (2017). *Vliv technologií v oblasti vzdělávání a v pedagogickém výzkumu: Sborník abstraktů z XXV. Konference České asociace pedagogického výzkumu konané ve dnech 13.–14. září 2017 v Hradci Králové* [The impact of technologies in education and pedagogical research]. Hradec Králové: Gaudeamus.

MPO. (n.d.). *Digitální Česko v. 2.0, cesta k digitální ekonomice* [Digital Czechia v. 2.0, pathway to digital economy]. Retrieved from https://www.mpo.cz/assets/cz/e-komunikace-a-posta/Internet/2013/4/Digi_esko_v.2.0.pdf.

MŠMT. (2014). *Strategie digitálního vzdělávání do roku 2020* [Strategy for digital education until 2020]. Retrieved from http://www.msmt.cz/file/34429.

MŠMT. (2017). *Průběžné hodnocení implementace Strategie digitálního vzdělávání do roku 2020 (rok 2016)* [Ongoing evaluation of the implementation of the Strategy for digital education until 2020; 2016]. Retrieved from http://www.msmt.cz/uploads/Implementace_SDV_zprava_za_rok_2016.pdf.

MŠMT. (2018). *Průběžné hodnocení implementace Strategie digitálního vzdělávání do roku 2020 (rok 2017)* [Ongoing evaluation of the implementation of the Strategy for digital education until 2020; 2017]. Retrieved from http://www.msmt.cz/uploads/zpra_va_SDV_vla_da_2017.docx.

OECD. (2005). *Are students ready for a technology-rich world? What PISA studies tell us*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/35995145.pdf.

OECD. (2011). *PISA 2009 Results: Students on line*. Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/9789264112995-en.

OECD. (2014). *PISA 2012 Technical report*. Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

OECD. (2015). *Students, computers and learning: Making the connection*. Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/9789264239555-en.

OECD. (2017a). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264266490-en.

OECD. (2017b). *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264267510-en.

OECD. (2017c). *PISA 2015 Results (Volume III): Students' Well-Being*. Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/9789264273856-en.

OECD. (2017d). *PISA 2015 Technical report*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf.

Palečková, J., Tomášek, V., & Basl, J. (2010). *Hlavní zjištění výzkumu PISA 2009. Umíme ještě číst?* [Principal findings of PISA 2009 survey. Can we still read?]. Praha: Ústav pro informace ve vzdělávání.

Ponzo, M. (2011). Does the way in which students use computers affect their school performance? *Journal of Economic & Social Research*, *13*(2), 1–27.

Potužníková, E., Lokajíčková, V., & Janík, T. (2014). Mezinárodní srovnávací výzkumy školního vzdělávání v České republice: Zjištění a výzvy [International comparative studies on school education in the Czech Republic: Findings and challenges]. *Pedagogická Orientace*, *24*(2), 185–221. https://doi.org/10.5817/PedOr2014-2-185

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Skryabin, M., Zhang, J., Liu, L., & Zhang, D. (2015). How the ICT development level and usage influence student achievement in reading, mathematics, and science. *Computers & Education*, *85*, 49–58. https://doi.org/10.1016/j.compedu.2015.02.004

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publications.

Soukup, P. (2006). Proč užívat hierarchické lineární modely? [Why use hierarchical linear models?]. *Sociologický časopis/Czech Sociological Review*, *42*(5), 987–1012.

Soukup, P. (2012). Mezinárodní výzkumy v oblasti vzdělávání [International surveys in education]. In J. Krejčí & Y. Leontiyeva (Eds.), *Cesty k datům: Zdroje a management sociálněvědních dat v České republice* (pp. 302–324). Praha: Sociologické nakladatelství (SLON).

Soukup, P. (2016). Možnosti praktické práce s daty z mezinárodních vzdělávacích studií: Problémy a jejich praktická řešení [Possibilities of practical work with data from international large scale educational assessments: problems and practical solutions]. *Orbis Scholae*, *10*(1), 97–120. https://doi.org/10.14712/23363177.2016.15

Spiezia, V. (2010). Does computer use increase educational achievements? Student-level evidence from PISA. *OECD Journal: Economic Studies*, *2010*(7), 1–22. https://doi.org/10.1787/eco_studies-2010-5km33scwlvkf

Straková, J. (2009). Vzdělávací politika a mezinárodní výzkumy výsledků vzdělávání v ČR [Educational policies and international surveys of educational outcomes in CR]. *Orbis Scholae*, *3*(3), 103–118. https://doi.org/10.14712/23363177.2018.200

Straková, J. (2010). Dopad diferenciace vzdělávacích příležitostí v povinném vzdělávání na vývoj nerovností ve výsledcích žáků v ČR po roce 2000 [The impact of differentiation of

educational opportunities in compulsory education on the development of inequality in pupils' results in the CR since 2000]. *Pedagogika*, *60*(2), 21–37.

Straková, J. (2016). *Mezinárodní výzkumy výsledků vzdělávání: Metodologie, přínosy, rizika a příležitosti* [International surveys of educational outcomes: methodologies, merits, risks and opportunities]. Praha: Univerzita Karlova v Praze, Pedagogická fakulta.

Sucháček, P. (2014). Spor o víceletá gymnázia: Historický kontext a empirická data [The dispute about classical grammar schools: Historical context and empirical data]. *Studia Paedagogica*, *19*(3), 139–154. https://doi.org/10.5817/SP2014-3-8

Woessmann, L., & Fuchs, T. (2004). Computers and student learning: Bivariate and multivariate evidence on the availability and use of computers at home and at school. *CESifo Working Paper Series No. 1321*. Retrieved from https://ssrn.com/abstract=619101

Wittwer, J., & Senkbeil, M. (2008). Is students' computer use at home related to their mathematical performance at school? *Computers & Education*, *50*(4), 1558–1571. https://doi.org/10.1016/j.compedu.2007.03.001

Zounek, J., & Šeďová, K. (2009). *Učitelé a technologie: Mezi tradičním a moderním pojetím* [Teachers and technologies: Between the traditional and the modern approach]. Brno: Paido.

Zounek, J., & Tůma, F. (2014). Problematika ICT ve vzdělávání v českých pedagogických časopisech (1990–2012) [Issues related to ICT in education from the perspective of Czech educational journals (1990–2012)]. *Studia Paedagogica*, *19*(3), 65–87. Retrieved from http://www.phil.muni.cz/journals/index.php/studia-paedagogica/article/view/902.

Mgr. Bc. Libor Juhaňák
Department of Educational Sciences
Faculty of Arts, Masaryk University
Arne Nováka 1, 602 00 Brno
juhanak@phil.muni.cz

Assoc. Prof. Mgr. Jiří Zounek, Ph.D.
Department of Educational Sciences
Faculty of Arts, Masaryk University
Arne Nováka 1, 602 00 Brno
zounek@phil.muni.cz

Mgr. Klára Záleská
Department of Educational Sciences
Faculty of Arts, Masaryk University
Arne Nováka 1, 602 00 Brno
zaleska@phil.muni.cz

Mgr. Ondřej Bárta
Department of Educational Sciences
Faculty of Arts, Masaryk University
Arne Nováka 1, 602 00 Brno
barta@phil.muni.cz

Mgr. Kristýna Vlčková
Department of Educational Sciences
Faculty of Arts, Masaryk University
Arne Nováka 1, 602 00 Brno
vlckova.k@mail.muni.cz

Comparing results of TIMSS and the Hungarian National Assessment of Basic Competencies

Ildikó Balázsi, Ildikó Szepesi

Hungarian Educational Authority, Department for Analyses of Public Education / Department of Assessment and Evaluation

Abstract: In our paper, we compared some characteristics of TIMSS 2015 and the National Assessment of Basic Competencies (NABC) 2015. The NABC assesses all students' reading and mathematics performance in Grades 6, 8 and 10. Both studies assessed Hungarian Grade 8 students' mathematics abilities in the spring of 2015. We linked data of the two studies on the student level using Student Measurement IDs.

We compared TIMSS and NABC mathematics scales based on the Assessment Framework of the two studies along with the results of students in the two assessments. The comparison of the Frameworks revealed that although the two tests use similar content and cognitive categorizations, there are crucial differences between the two constructs. While the basis of TIMSS's mathematics construct is the common part of mathematics curricula of participating countries, NABC intends to measure mathematical literacy, the ability of students to use their mathematical knowledge and competencies in real life situations. The correlation between the TIMSS and NABC mathematics test results (0.79) also confirms that the two tests measure related, but not identical abilities.

To evaluate the representativeness of the TIMSS sample we used school- and class-level weight factors of TIMSS and the student-level weights of NABC combined. The mean performances of the TIMSS sample are only slightly lower than the full NABC cohort's, the effect size of the difference is 0.042 and 0.046 in mathematics and reading respectively. The differences in the standard deviations are somewhat but not considerably larger. The SES-index shows a very good match with no statistically significant differences in the mean and standard deviation of the sample and the full cohort. Our analysis confirms that estimations of population parameters based on TIMSS samples are of a good quality.

Keywords: TIMSS 2015; Hungarian NABC; mathematics test; representativeness

The Hungarian Evaluation and Assessment Framework uses both international and national student assessments to evaluate the performance and other characteristics of the school system (Sinka, 2010). TIMSS, as one of the international large scale student assessments, describes the school systems' characteristics and quality by measuring their students' performance in an international context every four year. The International Reports of TIMSS compare countries based on their student achievement in mathematics and science along with student, teacher and school characteristics in Grades 4 and 8 (Mullis, Martin, Foy, & Hooper, 2016).

In contrast, the Hungarian National Assessment of Basic Competencies (NABC) evaluates individual schools' results, reporting their students' performance in reading and mathematics in Grades 6, 8 and 10 annually (Balázsi, Lak, Ostorics, Szabó, & Vadász, 2016). The main aim of NABC is to empower schools and the wider public

https://doi.org/10.14712/23363177.2018.293
www.orbisscholae.cz

with objective, reliable and comparable data on students' performance in areas crucial for the well-being and prosperity of students in their later life. The reports show students' results by various background characteristics, like settlement type and size, school type and size, socio-economic background and baseline performance from two years earlier. Therefore, schools, school maintainers and parents can evaluate the schools' results taking into account these background characteristics.

The two studies serve different purposes and complement each other. From policy perspective, TIMSS is used to put the results and features of the Hungarian school system in international context, to evaluate strengths and weaknesses in comparison with similar countries or with countries which can be seen as a role model for some reason (see for example Szalay, Szepesi, & Vadász, 2016, pp. 270–271). NABC is mainly used on school level, although detailed analyses of within country structure of the education system are also available (Balázsi et al., 2016), as well as secondary analyses of the data used alone or complementing other primary data collections (see for example Horn, 2013; Kertesi & Kézdi, 2016).

Transparency, validity and reliability of the data presented in these assessments is crucial from educational policy perspective. To ensure these, both TIMSS and NABC published their methods and processes followed during test development, sampling, data collection and reporting (Martin, Mullis, & Hooper, 2016; Aux-Bánfi et al., 2015). In our research, we attempt to analyze retrospectively the validity of the results, based on crosschecking the data in the two databases.

In our paper, two research questions are addressed. Since both TIMSS and NABC assessed the mathematics performance of students in Grade 8, our first question is whether the two constructs are the same and if not, what the main differences are. Mathematics is a complex construct, and different assessments might define it somewhat differently according to their aims. Besides, during test item development and test item selection, different content or cognitive areas of mathematics might get different emphasis in each study, and item types can differ as well. According to a research conducted by the U.S. Department of Education Institute of Education Sciences: "although the NAEP, TIMSS, and PISA 2003 mathematics frameworks address many similar topics and require students to use a range of cognitive skills and processes, it cannot be assumed that they measure the same content in the same way" (Neidorf, Binkley, Gattis, & Nohara, 2006, p. iv.). To interpret the mathematics results from TIMSS and NABC correctly, we have to take into account any differences of the two studies' mathematics scale constructs. Besides, with linking the TIMSS and NABC databases on student level, we can analyze the correlation between the results in the two studies, giving empirical support for our findings based on the content comparison of the tests.

Our second research question addresses the representativeness of the TIMSS 2015 sample. TIMSS has rigorous procedures for sampling and participation, makes a great effort to ensure that its findings are valid for the whole educational system of the participating countries and regions (LaRoche, Joncas, & Foy, 2016). However, linking TIMSS and NABC data on student level, we can evaluate the representativeness of

the TIMSS sample retrospectively and independently from the databases and procedures used for sampling in TIMSS. Although scientifically might seem unnecessary to evaluate the representativeness of the TIMSS sample due to the scientific rigor of sampling in TIMSS, skeptical views on international large scale studies among educators, policy makers or the general public from time to time question the quality of the sample or the relevance of results coming from a sample to the whole population. For example, László Mendrey, the president of the Hungarian Democratic Trade Union of Teachers at that time stated to an online newspaper that "… only 150 Hungarian schools' 4th Grade students participated in PILRS. […] The problem is that PIRLS represents only the results of a fragment of schools, as there are more than four thousand public education institutions in Hungary."[1] Therefore, having data of the whole population on which TIMSS sampling is based gives an excellent opportunity to prove the relevance of the results of international large scale assessments for the whole educational system. Proving representativeness of the sample independently from the study itself makes a strong argument easily comprehensible for a nonprofessional audience as well.

1 Data and Methods

In our research, we chose to link and analyze TIMSS 2015 and NABC 2015 Grade 8 databases, since these assessed the same student population, and have mathematics as one of their cognitive domain in common. Student measurement identification (SMID), which, introduced in 2008, is used in every international and national large-scale assessment in Hungary, and allows us to link the data from the two studies on student level. The Hungarian TIMSS data was collected between 30th March and 28th April, NABC 2015 was administered on 27th May in the same school year. As NABC is a census, the TIMSS sample is approximately[2] a sample form the NABC population. Indeed, all but 8 students of the TIMSS sample are present in the NABC data file (Table 1). These 8 students either dropped out from the school system or moved abroad between the two data collections, or some database error in one of the studies prevented linking the SMIDs.

We compared the mathematics test contents based on the frameworks of the two studies (Mullis & Martin, 2013; Balázsi et al., 2014). Simultaneously, we evaluated the similarities of the two results using correlations. We have used TIMSS plausible

[1] Article published on December 6th, 2017 on https://24.hu/belfold/2017/12/06/a-magyar-diakok-minden-eddiginel-jobb-eredmenye-nem-pont-az-aminek-latszik, the quoted sentences were translated by the author.

[2] Not exactly, as exclusion policies differ in the two studies. TIMSS, trying to be as inclusive as possible, asks schools to exclude students with special education needs (SEN) only if their disabilities would seriously affect their test writing and results. Schools with only such SEN students were excluded before sampling. In contrast, in NABC all SEN students were excluded to ensure school comparability. However, SEN students learning in inclusive schools are included in the database of NABC, although they are marked as non-eligible. The overall exclusion rate is 5.4% in TIMSS (Mullis et al., 2016, Appendix C.2) and 6.4% in NABC (own calculation).

Table 1 Number of students in TIMSS 2015 and NABC 2015 Grade 8 databases.

Number of students (unweighted)	Total	In TIMSS sample (eligible students)	With performance data (in TIMSS database)	With data on parents highest education
Total		5,058	4,893	4,857
In NABC database	88,967	5,050	4,891	4,855
NABC eligible	84,113	4,887	4,738	4,705
With NABC performance data	78,985	4,615	4,492	4,463
With NABC SES-index	62,317	3,836	3,736	3,715

values and NABC IRT ability scores to calculate correlations along with the weighting variable and jackknife error calculation methods of TIMSS (Martin et al., 2016, Aux-Bánfi et al., 2015).

TIMSS and NABC use slightly different methods and software for calculating the performance of students (Yamamoto & Kulick, 2016; Aux-Bánfi et al., 2015, pp. 91–105). Also, TIMSS item parameters are calculated based on data from every participating country using equal weights for every country (Foy & Yin, 2016). These differences might affect Hungarian students' TIMSS 2015 performance scores and hence the correlation between the two tests' results. In order to exclude these effects from our comparison, we have also compared NABC scores with a performance score of the TIMSS mathematics test calculated from item response level data of the Hungarian students with methods and software used in NABC.

To evaluate the representativeness of the TIMSS sample we compared NABC mathematics and reading results of students in the TIMSS sample to the results of the overall NABC cohort. We have also used students' SES-indices to compare the TIMSS sample to the whole population. The SES-index is based on some questions of the non-compulsory student background questionnaire of NABC, so due to a large number of missing values we should interpret results based on SES carefully. However, the response rate is high, 79% of non-missing students have SES data as well. The index consist variables related to highest education of parents, number of books at home and educational and economical resources possessed by the students' family. It was anchored so the average SES is 0 and the standard deviation is 1 for the overall student population of the three grades involved in the assessment (Aux-Bánfi et al., 2015).

For the analyses of the representativeness of the TIMSS sample, using NABC student weights alone is not appropriate. TIMSS weights consist six different factors, school-, class- and student-level sampling weight factors[3] are supplemented with school-, class- and student-level adjustment weight factors to adjust for non-re-

[3] Student level sampling weight factors are 1 for all students, as all students of a selected class are added to the sample.

sponse on school-, class- and student-level (Martin et al., 2016). In NABC, as it is a census and for schools, classes and students it is compulsory to participate except of students missing from school on the day of assessment, the only weight factor not 1 by definition is the student adjustment factor for non-response (Aux-Bánfi et al., 2015). If we had analyzed the results of TIMSS students (i.e. students selected for participation in TIMSS) unweighted or using NABC weights alone, we would have neglected the correction effects of weighting used in TIMSS to clear estimates of population parameters from biases rising from unequal sampling probabilities and different response rates.

The TIMSS-sample file used in our analysis contains every student in responding schools and classes, however, non-responding schools and classes are not included. Hence, to analyze correctly the characteristics of students in the TIMSS sample we used school- and class-level weight and adjustment factors of TIMSS and the student-level weight factor for non-response of NABC combined, along with the block bootstrap method of NABC for error calculations. We have also calculated the effect sizes of differences, the difference divided by the standard deviation using the estimation method described in Hedges (2007) for nested data with unequal school sizes.

2 Results

2.1 Similarities and differences in the TIMSS and NABC Grade 8 mathematics constructs and results

Both TIMSS and NABC declare their scope and content in their Assessment Frameworks (Mullis & Martin, 2013; Balázsi et al., 2014). The TIMSS framework derives its mathematics scale and test items primarily from the mathematics curricula of the participating countries. In contrast, NABC intends to measure mathematical literacy: "the ability of an individual to understand and analyze the role of mathematics in the real world; the skillful use of mathematical tools; the willingness and ability to use the acquired mathematical knowledge in real life situations; the use of mathematical tools in communication and cooperation during social interactions – on a level adequate for the age of the individual". During test development, the mathematics Core Curriculum is taken into account to ensure students do not face problems involving mathematical tools and knowledge they did not learn up to that grade. However, according to the definition of mathematical literacy, in NABC mathematics test items are usually not purely mathematical and do not resemble simple textbook examples, but the mathematical problem students need to solve is embedded in some situation similar to the situations in which students should use mathematical tools and knowledge in their everyday life.

The differences in the definition of the two constructs have definite effect on the test booklets. The NABC 2015 mathematics test have a much higher reading load,

students need to read approximately two times as many words in NABC then in TIMSS during a 45 minute test period.[4]

Both TIMSS and NABC categorizes test items according to two aspects: their content and their cognitive demand, named content and cognitive dimensions in TIMSS, content areas and thinking processes in NABC. TIMSS assigns target percentages of testing time to different content and cognitive domains. NABC assigns target percentage intervals based on the number of items to every content area – thinking process category pair.

Both TIMSS and NABC have four content categories in Grade 8. TIMSS items can belong into the *Number*, the *Algebra*, the *Geometry* or the *Data and Chance* cognitive category, while in NABC the four categories are named *Quantities, numbers, operations*, *Assignments, relationships*, *Shapes, orientation*, and *Statistical characteristics, probability*. TIMSS testing time is divided so 30% of testing time used to solve *Number* items, 30% used for *Algebra* items, 20% used for *Geometry* and also 20% used for *Data and Chance*. In NABC, 35–40% of items belong to *Quantities, numbers, operations*, 25–30% belong to *Assignments, relationships*, 20–25% belong to *Shapes, orientation*, and 12–15% belong to *Statistical characteristics, probability*. The two divisions of mathematical contexts highly overlap (Table 2). For example, most topics from NABC's *Quantities, numbers, operations* category appear in TIMSS's *Number* category, however, NABC puts calculations of specific quantities in relation with geometric shapes, like calculating the volume or area of a geometric shape or using the Pythagorean theorem into the *Quantities, numbers, operations* category, while in TIMSS these items belong to the *Geometry* content category.

On the cognitive dimension, categorizations of the cognitive procedures necessary to solve the test items are even more similar in the two studies. TIMSS uses the *Knowing*, *Applying* and *Reasoning* categories in a way that 35%, 40% and 25% of testing time is devoted to each. In NABC, 25–30% of items belong to the *Knowledge of facts and simple operations* cognitive category, 45–55% belong to the *Application, integration* category and 20–25% belong to the *Complex solutions and evaluation* category. The three categories used in NABC are almost equivalent to the categories used in TIMSS, both based on Bloom's taxonomy of the cognitive domain. Both TIMSS and NABC uses multiple-choice and constructive response items, in TIMSS at least half of the items, in NABC 55–65% of items are multiple choice according to the Framework.

The correlation coefficient between the TIMSS and NABC mathematics test results is 0.79 (Table 3). While this is a high value, not as high as would be anticipated in two

[4] To analyze the reading load of the tests, we counted the number of words (with Word's word-counting function) in the 45 minutes long mathematics parts of the fourteen TIMSS 2015 booklets. We have also counted the words in eight 45 minutes long mathematics blocks of the NABC booklets between 2012 and 2015. Pairwise comparison of word counts in the two studies showed that in average the ratio of the number of words in the 45 minutes long blocks in NABC versus TIMSS was 189% (the standard deviation of the ratios was 19 percentage points).

Table 2 Content topics of NABC matched to the content categories of TIMSS (the percent figures in the first column indicate the percent of test items devoted to individual content area).

Content areas in NABC	
Quantities, numbers, operations (35–45%)	Covered in the ***Number*** category of TIMSS
	Numbers: number line; intervals; place value; fractions and decimals (equivalence, comparison, reduction, visualization) *Calculations, operations*: multiple operations (e.g. writing, perform, powers, square root, rounding), data needed for a calculation; calculating percentages of a value, conversion between per cents and fractions, their visualization; calculating with ratios; proportion compared to 1 *Measurements*: scales (reading and representing data) e.g. thermometers, clocks; comparing quantities; conversion of units; computing with time *Divisibility*: common divisors, greatest common divisor, smallest common multiple, remainders, divisibility rules
	Covered in the ***Algebra*** category of TIMSS
	Calculations, operations: substituting a value into an algebraic expression without rearrangement
	Covered in the ***Geometry*** category of TIMSS
	Calculations, operations: operations with geometric shapes (e.g. perimeter, area, volume, Pythagorean Theorem)
	Not explicitly covered in the TIMSS framework
	Numbers: scientific notation; *Measurements*: time zones
Assignments, relationships (25–30%)	Covered in the ***Number*** category of TIMSS
	Proportionality (direct and inverse proportionality, examples of proportions where each value is different from 1): ratio of numbers and quantities; scaling compared to other numbers than 1; calculating the total from percentages and the percentage value of a quantity
	Covered in the ***Algebra*** category of TIMSS
	Matching quantities (tables, functions, diagrams, graphs, etc. - not statistical data): reading relationships (value, slope, continuation, evaluation, etc.); representation of relationships (e.g. on graphs, diagrams), examination of representations; writing and application of relationship rules, parameterization, general formula, etc., relationship between variables *Parametric algebra*: operations with algebraic expressions and formulas with rearrangement; equations and inequalities *Sequences*: finding the next or a given element using the rule, finding the sequence number of an element, finding the sum of elements (without formula)
Shapes, orientation (20–25%)	Covered in the ***Geometry*** category of TIMSS
	Two-dimensional shapes: knowledge of geometric characteristics (e.g. diagonal of a square, angles of a triangle, angles and diagonals of regular and irregular polygons, parts of the circle); transformations in two dimensions: congruence (reflection through a line or a point, translation,

Shapes, orientation (20–25%)	rotation), symmetry, similarity (only based on intuition), completing a pattern; perimeters and areas of two-dimensional shapes (estimation, covering, rearrange parts, relation between parameters) *Three-dimensional shapes, dimensions*: representations of three dimensional objects (views, nets, components, etc.), bounding volumes (e.g. choosing the right box for a present); three-dimensional transformations (rotation, translation, similarity, reflection across a plane– recognizing the result of a transformation based on intuition); relationship between parameters of a three dimensional shape and its volume and surface *Orientation*: directions and cardinal directions, angle of view (based on intuition), locations in coordinate systems (e.g. chessboard, the globe, the Cartesian plane, contour maps)
Statistical characteristics, probability (12–15%)	Covered in the ***Data and Chance*** category of TIMSS
	Collecting statistical data from tables/diagrams: reading data, comparing data (smallest, largest, differences), evaluating and analyzing data *Statistical representation and data matching*: representing and matching data given in different forms (e.g. in written text, in tables, in diagrams) *Statistical calculations*: e.g. mean (average, weighted average), median, range, mode *Statistical methods*: e.g. choosing, interpreting, using, evaluating the appropriate statistics, identifying the data necessary for a statistics, identifying the statistical properties inferable from a statistical representation *Probabilities*: certain, impossible, possible events, chance, more likely, less likely, frequency, relative frequency etc. *Combinatory*: counting
	Not explicitly covered in the TIMSS framework
	Event graphs: counting the edges, paths; *Sets*: basic operations and their properties; *Formal logic*: logical values, operations

measurements of the same construct assessed two times in a two-month timeframe.[5] The correlation between the TIMSS mathematics and the NABC reading results is only slightly lower, 0.75. Recalculating TIMSS mathematics scores based on NABC's methods using only Hungarian students' data increased the correlation insignificantly (to 0.80), therefore our conclusions are not affected by the methods used for scaling.

Table 3 Correlation between the TIMSS 2015 and the NABC 2015 results.

	NABC mathematics		**NABC reading**	
	correlation	(*SE*)	correlation	*(SE)*
TIMSS mathematics score	0.79	(0.013)	0.75	(0.013)
TIMSS science score	0.74	(0.015)	0.73	(0.014)

[5] Cronbach's coefficient alpha, which is a lowerbound estimate of the internal consistency of a test and also can be seen as an estimate of the correlation between two tests measuring the same construct, is 0.91 for both mathematics tests for the Hungarian students (Martin, Mullis & Hooper, 2016 – Exhibit 11.8, pp11.16, Lak et al., 2016 – Table 2, p. 6).

While the two mathematic tests measure slightly different abilities, NABC still can be used to evaluate the TIMSS sample. As the TIMSS 2015 sample is practically a sample of the NABC cohort, the TIMSS sample should be a representative sample of the students in the NABC. The NABC results, as well as other characteristics of students in the TIMSS sample should be similar to the overall national results when using the TIMSS school- and class-level weights combined with NABC student-level weights.

First, we selected to compare the mathematics results of the TIMSS sample with the NABC cohort's (Table 4). The TIMSS sample's average performance was 8.2 points lower than the overall performance of students in Grade 8, which is a statistically not significant difference. Standard deviation estimated from the TIMSS sample was 14.6 points higher than the population parameter, which is a significant difference on a 0.05 significance level, but not on the 0.001 significance level.

Table 4 NABC 2015 Grade 8 mathematics performance of students.

	Number of students (weighted)	Mean performance (SE)	Standard deviation (SE)
Students in the TIMSS sample	81,836	1609.4 (11.47)	209.0 (6.00)
All students	84,108	1617.6 (2.50)	194.4 (1.14)

To evaluate the relevance of the difference in the mean performances estimated from the TIMSS sample and the NABC participants, we also estimated the effect size as a proportion of the standard deviation of the whole population, which is 0.042. A difference of the same effect size in Hungarian students' average performance on the TIMSS scale would be 3.9 points,[6] not statistically significant. Although the two tests' contents are not identical, and a difference on the NABC mathematics scale could not be transferred directly to the TIMSS scale, these findings confirm that the national average based on the whole NABC cohort probably would lay in the confidence interval of the mean performance of the Hungarian national average published by TIMSS. The two estimations of the standard deviations differ slightly more, indicating that it is possible that on the TIMSS scale, the standard deviation is somewhat overestimated.

Comparing reading results of the TIMSS sample to the whole NABC cohort's leads to the same results: TIMSS students' reading performances are somewhat, but not statistically significantly lower than the national average, the effect size of the difference is 0.046. The standard deviation of reading performances of TIMSS students is 13.2 points higher than the standard deviation of the whole population, which is also a significant difference on the 0.05 level and not on the 0.001 level.

We also compared TIMSS students based on their socio-economic status, where differences were even smaller (Table 5). The effect size of difference in the average SES was 0.009.

[6] In TIMSS, the average performance of the Grade 8 population was 514 score points (*SE* 3.8), the standard deviation is 93 (*SE* 2.2) (Mullis et al., 2016).

Table 5 Socio-economic status of students based on their answers to the NABC 2015 Grade 8 Student background questionnaire.

	Number of students (weighted)	Mean (SE)	Standard deviation (SE)
Students in the TIMSS sample	67,435	−0.027 (0.0643)	1.037 (0.0313)
All students	66,277	−0.018 (0.0161)	1.019 (0.0076)

Q-Q plots comparing the percentiles of the distribution of mathematics and reading results and SES-indices of the TIMSS sample to the NABC full cohort's same values also show that the sample represents the full cohort very well (Figure 1). The mathematics and reading results show minor differences on the lower and upper end of the distribution: somewhat more students have low results and slightly less students have high results in the TIMSS sample than in the full cohort. The distribution of the SES-index shows an almost perfect match between the two groups of students.

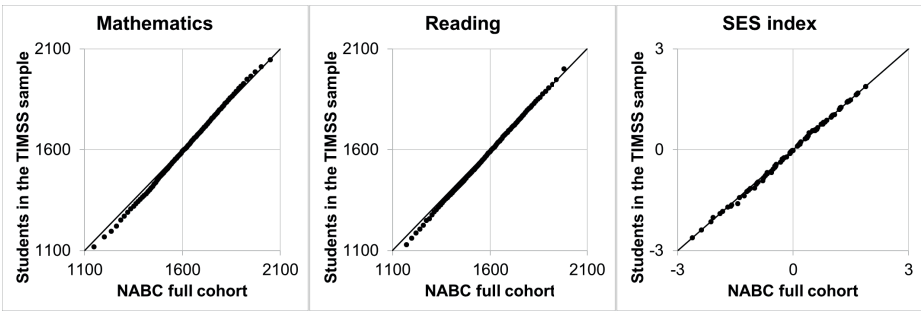


Figure 1 Q-Q plots comparing the distribution of the TIMSS sample to the full cohort's.

3 Discussion

In our research, we compared TIMSS and NABC mathematics scales based on the Framework of the two studies along with the results of students in the two assessments. Although TIMSS and NABC both measure mathematical abilities of students, there are some differences in the two constructs. NABC test items are usually not purely mathematical but mathematical problems embedded in real life situations. While TIMSS also uses problem situations in some of their items, the TIMSS mathematics test mainly contains items more similar to regular examples in a mathematical textbook. Furthermore, while the content and cognitive categorizations and the share of items from the different categories are similar in the two constructs, some minor differences do exist in the frameworks. For example in NABC there is slightly more emphasis given to the *Application, integration* cognitive category, while TIMSS has a little higher percent of testing time for the *Knowing* category. Examining the

correlation coefficient between the two mathematics results (0.79) also confirms that in accordance with the findings from the comparison of the frameworks, the two tests measure related, but not identical abilities.

We evaluated the representativeness of the TIMSS sample using NABC mathematics and reading results and the SES-index of students. Our analysis confirms that the sample of TIMSS represents very well the full NABC cohort, and estimations of population parameters based on TIMSS samples are of a good quality.

4 Further Research

Beside mathematics results, in the continuation of the research presented in this paper, we intend to compare other characteristics of the students measured in the two studies. Student's socio-economic status (SES) is highly correlated with their abilities, and the NABC uses SES as one of the main characteristics in school reports. Hence, its validity is crucial for the study. TIMSS measures the same or similar socio-economic variables, using them to provide international comparisons of the effect of SES on mathematics and science abilities. Therefore, crosschecking the stability of these variables can support the validity and relevance of reports based on SES for both studies.

We also intend to analyze how missing data of non-participating students can distort the results of the studies. On the one hand, we are going to analyze how the NABC achievement of students in the TIMSS sample with missing data compares to the NABC achievement of TIMSS participants and how their participation would have affected the national TIMSS result of Hungary. And, vice versa, we are going to analyze, how students with missing data in the NABC performed in TIMSS. TIMSS uses follow-up sessions for absent students to maximize participation rate, while NABC is written on the same day in every school without any possibility to reach students absent on the day of assessment. Accordingly, the later has a somewhat lower participation rate on student level, 94% compared to 97% in TIMSS. Our research question is whether there are systematic patterns in absent students' characteristics and abilities, and how missing data affects school level and overall results. Similarly, we are going to examine the consequences of non-responding to the student questionnaire in NABC.

References

Aux-Bánfi, I., Balázsi, I., Balkányi, P., Balogh, V. K., Gyapay, J., Lak, Á. R., Ostorics, L. I., Palincsár, I., Rábai-Szabó, A., Rózsa, Cs., Szabó, Á., Szabó, L. D., Szepesi, I., Szipőcs-Krolopp, J., & Vadász, Cs. (2015). *Országos kompetenciamérés. Technikai leírás*. Retrieved from https://www.oktatas.hu/kozneveles/meresek/kompetenciameres/tanulmanyok_publikaciok.

Balázsi, I., Balkányi, P., Bánfi, I., Szalay, B., & Szepesi, I. (2012). *A PIRLS és TIMSS 2011 tartalmi és technikai jellemzői*. Retrieved from https://www.oktatas.hu/kozneveles/meresek/timss/timss_2011_meres.

Balázsi, I., Balkányi, P., Ostorics, L., Palincsár, I., Rábai-Szabó, A., Szepesi, I., Szipőcs-Krolopp, J., & Vadász, Cs. (2014). *Az Országos kompetenciamérés tartalmi keretei – Szövegértés, matematika, háttérkérdőívek*. Retrieved from https://www.oktatas.hu/kozne veles /meresek/kompetenciameres/tanulmanyok_publikaciok.

Balázsi, I., Lak, Á. R., Ostorics, L., Szabó, L. D., & Vadász, Cs. (2016). *Országos kompetenciamérés 2015 – Országos jelentés*. Retrieved from https://www.oktatas.hu/kozneveles /meresek/kompetenciameres/eredmenyek.

Foy, P. & Yin, L. (2016). Scaling the TIMSS 2015 Achievement Data. In M. O. Martin, I. V. S. Mullis, & M. Hooper, (Eds.), *Methods and procedures in TIMSS 2015*. Retrieved from http:// timssandpirls.bc.edu/publications/timss/2015-methods.html.

Hedges, L. V. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370.

Horn, D. (2013). Diverging performances: The detrimental effects of early educational selection on equality of opportunity in Hungary. *Research in Social Stratification and Mobility*, *32*(June), 25–43.

Kertesi, G. & Kézdi, G. (2016). On the Test Score Gap between Roma and non-Roma Students in Hungary and its Potential Causes. *Economics of Transition*, *24*(1), 135–162.

Lak, Á. R., Palincsár, I., Szabó, L. D., Szepesi, I., & Szipőcs-Krolopp, J. (2016). *Országos kompetenciamérés 2015 – Feladatok és jellemzőik. Matematika, 8.* évfolyam. Retrieved from https://www.oktatas.hu/kozneveles/meresek/kompetenciameres/feladatsorok.

LaRoche, S., Joncas, M., & Foy, P. (2016). Sampling design in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper, (Eds.), *Methods and procedures in TIMSS 2015*. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Martin, M. O., Mullis, I. V. S., and Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International results in mathematics*. Retrieved from http://timssandpirls.bc.edu/timss2015/international -results.

Mullis, I.V.S. & Martin, M.O. (Eds.) (2013). *TIMSS 2015 assessment frameworks*. Retrieved from http://timssandpirls.bc.edu/timss2015/frameworks.html.

Neidorf, T.S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 assessments* (NCES 2006-029). Retrieved from http://nces.ed.gov/pubsearch.

Sinka, E. (2010). *OECD review on evaluation and assessment frameworks for improving school outcomes. Country background report. Hungary.* Retrieved from http://www.oecd.org /edu/school/50484774.pdf.

Szalay, B., Szepesi, I., & Vadász, Cs. (2016). *TIMSS 2015* Összefoglaló *jelentés*. Retrieved from https://www.oktatas.hu/kozneveles/meresek/timss/timss_2015_meres.

Yamamoto, K. & Kulick, E. (2016). TIMSS 2015 achievement scaling methodology. In M. O. Martin, I. V. S. Mullis, & M. Hooper, (Eds.), *Methods and procedures in TIMSS 2015*. Retrieved from http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Ildikó Balázsi
Department for Analyses of Public Education
Hungarian Educational Authority
Szalay utca 10-14, 1055 Budapest
balazsi.ildiko@oh.gov.hu

Ildikó Szepesi
Department for Assessment and Evaluation
Hungarian Educational Authority
Szalay utca 10-14, 1055 Budapest
szepesi.ildiko@oh.gov.hu

Linking Mathematics TIMSS Achievement to National Examination Scores and School Marks: Unexpected Gender Differences in Slovenia

Barbara Japelj Pavešić
Educational Research Institute, Ljubljana, Slovenia

Gašper Cankar
National Examinations Centre, Ljubljana, Slovenia

Abstract: In the article, we present the results of the Slovene national study of three different assessments of mathematics for students in Grade 8 and Grade 13: the independent TIMSS or TIMSS Advanced outcomes, the national external examination scores and internal teacher's marks. Grade 8 students who participated in TIMSS also took the national assessment (NA) one year after TIMSS assessment; TIMSS Advanced math students took the Matura examination from mathematics two months after the TIMSS Advanced assessments. The data on school marks from mathematics were collected with the nationally added questions to the international TIMSS and TIMSS Advanced questionnaires for students, together with the series of questions about the effort put into solving the test. One year later, the outcomes from TIMSS assessment, national examinations (Grade 9 and Grade 13) and school grades for each student were linked and the differences between boys and girls, attitudes toward mathematics and plans for future education were analysed. It was found that gender differences at scores from national exams as well as in school marks differ from gender differences in TIMSS and TIMSS Advanced achievement (i.e. Grade 13 students' Matura results are slightly in favour of girls while TIMSS Advanced show better achievement for boys). Comparison of three outcomes reveal some characteristics of both national examinations and teachers' marking not evident otherwise. Matura gives to the most able students proportionally less opportunity to demonstrate the highest cognitive level of knowledge. Boys who demonstrated the same knowledge in TIMSS as girls get lower national marks as girls, in exams and by teachers. Girls put less effort than boys in solving the TIMSS test which could help to explain the changing gender gap from TIMSS to the national examinations. In Grade 8, the marks and TIMSS scores also show inconsistencies on student level. They are differently associated with attitudes toward mathematics which can provide some ideas for improvement of low motivation for learning mathematics in Slovenia.

Keywords: TIMSS; mathematics; school grades; national examination; gender difference

In Slovenia, the central database of the national examination results and participation in the international assessments of mathematics enabled the national study of links between international and national measurements of knowledge of mathematics aimed for better understanding of different gaps in achievement between specific groups of students in both measurements. TIMSS assessed the representative sample of all students at the penultimate grade of elementary school, at Grade 8, and TIMSS Advanced, at the end of general upper secondary school, at Grade 13 from

https://doi.org/10.14712/23363177.2018.294
www.orbisscholae.cz

mathematics and collected also the rich set of background data about learning and teaching mathematics, every four years from 1995 to 2015.

All Slovene students in Grade 9 take the national external assessment (NA) from mathematics, Slovene language and one of other school subjects that varies between years and schools. NA scores are used as secondary criteria for admission to the upper secondary schools with limited number of accepted candidates, which are usually highly demanding general gymnasia. All students at the end of Grade 13 in general upper secondary schools, gymnasia, take compulsory national external mathematics examination, Matura. Passing the examination from mathematics and additional 4 subjects is required for admission into any academic university study and scores are used as criteria for the most elite university studies, such as medicine, biological sciences and law with limited numbers of places for new students. Besides the assessments, all students are given marks for mathematics knowledge internally by their school teachers. Teachers' marks from elementary school are used as the main criteria for student's admission into upper secondary schools and marks from those schools are used as the secondary criteria for students' admission to some elite university studies. There is no central database of teachers' internal marks and no specific national study of marking.

The main reason for the study were unexplained gender differences in mathematics achievement that are not consistent across all measurements of knowledge. In Grade 13, the independently measured mathematics achievement from TIMSS Advanced 2008 and 2015 assessments is higher for boys while Matura scores are traditionally higher for girls. In Grade 8, TIMSS achievement didn't differ among boys and girls in last 20 years and same is true for results from the NA, but school marks are still higher for girls. Comparisons within elementary schools (Grades 1–9) also show unexplained trends of increased frequency of the highest marks. In view of the importance of assessment results for individual students, especially their influence on admission into demanding academic study courses or upper secondary schools, our research questions were following: Do the highest achieving students in all assessments form the same groups in Grade 13 and Grade 8? How are scores from assessments related to teacher's marks? Are girls in Slovene schools in mathematics marked higher than boys because an assessment includes oral part? Does the background data suggest any explanation why their marks are higher? What are possible differences in scope and content of TIMSS Advanced and National tests that produce observed differences?

1 Gender differences in mathematics

Gender differences in mathematics are well researched area. In general, many articles address the gender differences in achievement, characteristics of differences in externally measured achievement or teacher's marks, across time, countries and different student samples. Research also often focuses on personal or sociological

factors leading to higher achievement of boys or girls. Our research is focused on the differences between externally measured achievement and internal school marks. TIMSS Advanced study found the notable gender differences in pre-university mathematics achievement in most participating countries, all in favor of boys. Even more, there was no country where girls outperformed boys, even if majority of advanced mathematics students in some of the participating populations were girls, as it is the case of Slovenia and Portugal (Mullis, Martin, Foy, & Hooper, 2016a). International examination of excellence gap trends from TIMSS data up to 2012 found shrinking sex-based gaps in Grade 8 mathematics (Rutkowski, Rutkowski, & Plucker, 2012). These findings are in line with other research findings of increasing gender difference over years of schooling, from an elementary to a high school. The increasing trend of gender differences in favor of boys was found in solving complex math problems although so small that results of the study still supports gender similarities hypothesis (Hyde, 2005). From the meta-analysis of 242 studies of gender differences in mathematics published from 1990–2007 it is evident that young boys and girls are similar in mathematics, but small difference exists in high school in favor of boys when solving complex problems (Lindberg, Hyde, Petersen & Linn, 2010). In international comparisons, across different countries or cultures, gender differences are not consistent. The general gender gap could be different in different cultures, samples or age of students present many articles and were reviewed by Ceci, Williams, and Barnett (2009).

In general, gender gap in academic achievement requires complex studies on all aspects, starting from early schooling (Halpern, 2014). Baye and Monseur (2016) studied gender differences of students aged 14 to 15 years from TIMSS and PISA, from 1995 to 2015 in an international perspective and confirmed that gender differences vary by content area, students' educational levels, and students' proficiency levels and that males are more frequently among the highest performing students in mathematics and science. Opposite, the meta-analysis of more than three hundred studies of differences in school marks in mathematics, mostly from North America, confirms that small difference exists in favor of girls. Authors found that differences were smaller in the elementary school than in college, larger in North America than in few other countries, but not linked to the year of study or the scale of marks (Voyer & Voyer, 2014).

With the development of statistical methods for meta-analysis, researchers often observe the gender gap of mean outcomes and the gap of outcomes at the tails of distributions which could help to explain small gender differences in means but the higher number of most successful male students in mathematics and science internationally (Ceci, Williams, & Barnett, 2009). Bergold, Wendt, Kasper, and Steinmayr (2017) found that already in Grade 4, across many countries participated in TIMSS and PIRLS 2011, boys were more likely than girls to perform at the top level on general academic performance tests, not only in mathematics. Reilly, Neumannn and Andrews (2015) in their meta-analysis confirmed the small mean differences but greater male variability in mathematics achievement of Grade 12 USA students

in NAEP studies from 1990–2011. Many studies tried to find explanations for gender differences in achievement. For Van Houtte, different study culture of boys and girls can be linked to lower achievement of boys (2004). Self-discipline and higher ability self-concept of German girls are linked to their higher math performance according to Steinmayr and Spinath (2008). Study with focus on relation of gender differences to attitudes of girls and women in different cultures showed small gender differences in mathematics achievement of students in Grade 8 or aged 15 years, from TIMSS and PISA 2003 in the meta study of 69 countries, but found more positive math attitudes and affect among boys (Else-Quest, Shibley Hyde, & Linn, 2010). However, the study of Ceci et al. (2009) found that factor connected to the underrepresentation of women in science could be "that mathematics-capable women disproportionately choose non-mathematics fields and that such preferences are apparent among math-competent girls during adolescence". The study of Implicit Association Tests in 34 countries revealed that people implicitly associate science with males more than with females which has consequences for gender differences in Grade 8 science and mathematics achievement on national level (Nosek et al., 2009). Analyses of external achievement and internal school marks are rare in the literature compared to studies of gender gaps in one measurement. However, the study of school marks and national assessment outcomes in Sweden identified some factors related to the variation in assigning school marks in schools beside the individual student's achievement, such as education of parents (Klapp Lekholm & Cliffordson, 2008). The study of differences in school marks and scores from national exams among Grade 8 students in Croatia found that there is slight gender difference in achievement between both but no differences in marks given by female and male teachers although authors did not observe mathematics but other school subjects (Burusic, Babarovic, & Seric, 2012). Girls are receiving higher marks for their demonstrated knowledge because they have better verbal intelligence, higher agreeableness, stronger self-discipline, as well as certain aspects of the motivation by Spinath, Eckert, and Steinmayr (2014).

2 Slovene assessments of mathematics

In Slovenia, the target population for TIMSS Advanced 2015 was the population of general upper secondary school students who took the Matura Examination two months after TIMSS Advanced assessment. Slovenia differs from other countries in TIMSS with very large mathematics coverage index (34%), which is the percentage of students taking the most advanced mathematics pre-university course as part of the whole appropriate age cohort in Slovenia. Slovene school system requires from all future students of academic university studies to finish the same, most advanced, general 4-year upper secondary school programs for all school subjects and pass the final National Examination (called Matura) at the end. The final examination consists of exams from three compulsory and two chosen subjects. Mathematics is compul-

sory and can be taken at basic level (BL) or advanced level (AL) by student. All students follow the same mathematics curriculum and they finally decide individually about their level of mathematics exam only few weeks before the examination. With passing the advanced level of math exam, students can reach higher maximum math score (up to 8 on AL instead of up to 5 on BL). A sum of examination scores from all five subjects is used as a student's qualification for the admission to the university. Therefore, mathematics examination results are important motivation for students who compete to enter the most elite university studies, regardless of whether they are going to study mathematics further or not.

Elementary schools in Slovenia include nine grades for students aged 6 to 15 years. Mathematics is one of the most important subjects, it is taught at least 4 lessons per week in every grade and from Grade 6 by specialist teachers with finished university degree from educational mathematics. Students in elementary and upper secondary schools are given marks from 1 (unsatisfactory) to 5 (excellent) by their teachers, for written tests and oral questioning, according to their demonstrations of reaching the national standards of knowledge prescribed in curriculum. At the end of school year, for each student, a teacher summarises all marks given that year in one final mark for the official student's report. In elementary school, the large majority of students receive marks 4 and 5 even from mathematics. At the end of Grades 6 and 9 students take national examinations from mathematics. The student's result and teachers' marks for every school subject for Grades 7 to 9 are used for student's placement in the upper secondary school. The procedure is centralised and include also the private upper secondary schools. While majority of students is placed in the chosen schools, there is usually a strong competition for few of the most demanding general schools (gymnasia) leading to most prestigious university studies. Three school marks for mathematics from Grades 7 to 9 and one from NA are therefore important especially for the best students when competing to reach the desired future academic education.

From the report on gender differences in mathematics achievement found in data from PISA 2012 (OECD, 2015), Slovenia was among countries with notable gender gap between high achievers in favour of boys while no gap was found among low achieving 15-year-old students. Also, the difference in achievement among highest achieving students was likely to change into being in favour of girls after accounting for gender differences in mathematics self-beliefs.

In international comparative assessment of mathematics, TIMSS 2015, mean mathematics achievement of Slovene Grade 8 students is above the international average but not extremely high. The achievement from science of the same students scored higher on the international scale than mathematics, and both increased over all 5 measurements in the last 20 years (Mullis, Martin, Foy, & Hooper, 2016b). TIMSS Advanced results are average compared to other nations in general but very high for Slovene subgroup of students taking the advanced level of national examination. This subgroup is in fact more comparable to populations from other countries than the extremely large whole population of all future university students in Slovenia.

TIMSS and TIMSS Advanced assessments covered almost the same mathematics contents as the national mathematics examinations and national curriculum, as can be seen from the analysis of contents covered by teacher reports in TIMSS and TIMSS Advanced (Mullis et al., 2016a; 2016b). Therefore, the data from these international assessments enables the direct comparison of independent measurement of mathematics achievement with the mathematics knowledge demonstrated through examinations and school marks on student level.

3 Data and methods

Data sources for our study were databases of TIMSS (IEA, 2017b) and TIMSS Advanced 2015 (IEA, 2017a) student background data with mathematics achievement for Slovenia, extended with data about school marks, effort put into TIMSS or TIMSS Advanced assessments, and relevant national examination scores from mathematics. The questions about student final summarised school mark from mathematics and some other school subjects and series of questions about effort students put into solving the test were added as national options in TIMSS and TIMSS Advanced background questionnaires for students. Upper secondary school students were also asked about their chosen level of the national math exam (basic or advanced).

There were some differences between TIMSS and TIMSS Advanced. TIMSS sampled Grade 8 students from the whole national population while TIMSS Advanced target population were students enrolled into most advanced mathematics programmes in the country and could be in different grades in each country. In all participating countries, target populations for TIMSS Advanced covered less than 20% of age cohort in the country, except Slovenia, where all future university students (34% of age cohort) follow the same most advanced mathematics in their gymnasia. The students sampled were given the test with math items only. Items were mostly multiple choice or extended open ended questions. Students were given the list of formulae and were allowed to use calculators, although items were designed to be independent of calculator use. In case of using calculator, student was required to describe his procedure and write all partial results to get full credit. They were not allowed or needed to use geometry tools. Students taking the physics test were sampled separately and were in general not the same students as in the math sample. Assessments, coding of the answers and preparations of the dataset were done similarly in TIMSS and TIMSS Advanced.

Grade 8 students who participated in TIMSS in Spring 2015, all took NA one year later; all TIMSS Advanced math students took the Matura examination from mathematics two months after the TIMSS Advanced assessments in Spring 2015. A year and half after the TIMSS assessments, the Grade 8 student background data were merged with their scores from NA math examinations from the end of their Grade 9. Data for upper secondary school students were merged with scores from their Matura mathematics examination. Although all personal information was removed from

resulting datasets, they were still considered sensitive and all statistical analyses were done according to the policies for highly sensitive individual examination data inside the safe room at the National Examination Center who holds databases on national examinations in Slovenia. We used IDB-Analyzer and SPSS statistical package as recommended for all the large-scale assessments data which report learning outcomes in the form of plausible values, to calculate descriptive statistics, correlations, regressions and means inside benchmark levels of the achievement. Some of the results are also presented via ordinal dominance graphs (Cankar, 2016). Ordinal dominance (OD) graphs are a unique way to represent information which is exact, easily understandable and allows comparison of ordinal data. They are associated with Mann Whithey's *U*-test and the area under curve can be meaningfully interpreted and compared with other OD graphs (Bamber, 1975). The area under curve that is sometimes called ordinal mean effect (OME(X)) can be literally interpreted in the following way (Jewett, 1983): If we took randomly one subject from each group compared, OME(X) would be the probability that the subject from Group A (that's on the X axis) has higher score or equal to the subject from Group B (on the Y axis).

4 Results

4.1 Comparisons of gender differences in international assessments of TIMSS

National analysis of TIMSS Advanced data for Slovene upper secondary school students revealed high differences in achievement between the students who intended to take advanced or basic level of national mathematics exam. In international comparisons, the group of Slovene students taking advanced national mathematics exam reached TIMSS Advanced mathematics scores similar to the highest ranked advanced Russian student subpopulation. The mathematics coverage index for Slovene students was higher, 8.2% of age cohort, showing that almost a quarter of the whole future university student population (which represents 34.4% of age cohort) in Slovenia takes advanced level of the national exam. The mean score of students taking basic level of national exam was similar to the achievement of Swedish and Italian students as seen from the Table 1 (source: original TIMSS Advanced 2015 Exhibits M1.2 and M1.6 with added Slovene subpopulations with regard to the chosen level of national exam from mathematics (NE) data).

TIMSS Advanced mathematics achievement of boys was higher than achievement of girls in 6 countries; three other countries showed no gender difference. Slovene boys reached higher scores than girls overall and in two subpopulations. Because proportions of boys and girls are unequal in basic and advanced exam group in Slovenia, the overall difference for both groups together is even larger (27 points) than differences in each group (23 and 22 points respectively) – a situation known in statistics as Simpson's paradox (Cankar, 2010). The percentage of boys and girls

Table 1 Distribution of TIMSS Advanced 2015 mathematics achievement by gender, Grade 13.

Country	Coverage index	Total		Girls		Boys		Difference (boys–girls)	
		Mean	*SE*	Mean	*SE*	Mean	*SE*		*SE*
Slovenia, advanced level of NE	8.2%	549	(3.4)	538	(3.6)	562	(6.0)	−23*	(3.3)
Russian Federation 6 hours+	1.9%	540	(7.8)	530	(9.0)	549	(4.5)	−20*	(5.2)
Lebanon	3.9%	532	(3.1)	533	(4.8)	531	(3.9)	−2	(6.1)
USA	11.4%	485	(5.2)	470	(5.3)	500	(6.4)	−30*	(5.8)
Russian Federation	10.1%	485	(5.7)	480	(6.0)	489	(6.2)	−9*	(4.3)
Portugal	28.5%	482	(2.5)	481	(3.0)	483	(3.1)	−2	(3.6)
France	21.5%	463	(3.1)	449	(3.1)	475	(3.4)	−26*	(2.8)
Slovenia	34.4%	460	(3.4)	449	(3.5)	476	(4.9)	−27*	(4.7)
Norway	10.6%	459	(4.6)	453	(5.1)	463	(5.2)	−10*	(4.8)
Slovenia, basic level of NE	26.2%	433	(3.3)	425	(4.0)	447	(4.5)	−22*	(3.2)
Sweden	14.1%	431	(4.0)	424	(5.1)	436	(4.6)	−13*	(5.3)
Italy	24.5%	422	(5.3)	427	(6.1)	419	(6.6)	−8	(7.5)

Notes: Coverage index is fraction of student population as part of whole age cohort in a country; * Significant difference at the 0.05 level; Russian Federation 6 hours+ is subpopulation of students having 6 hours of math per week.

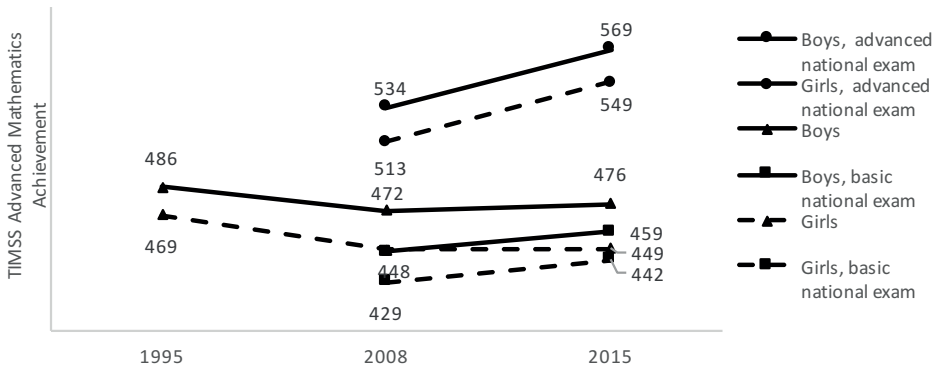


Figure 1 Trends in TIMSS Advanced 2015 mathematics achievement in Slovenia by gender and subpopulations, Grade 13.

from the whole age cohort reaching the top benchmark was similar (boys 8% and girls 7%), although there was share of 60% of girls among all TIMSS advance population and therefore, there are more males in the upper tail of distribution of achievement. Gender differences in TIMSS Advanced mathematics achievement in Slovenia stayed unchanged over the last 20 years (Figure 1).

For Grade 8 students, the analysis of trends of mathematics achievement gap between girls and boys shows no gender differences in the last 20 years. However, in 2015 boys outscored girls in one content domain (numbers) and girls slightly outscored boys in one other content domain (algebra), but no differences were seen in scores achieved in different cognitive domains of mathematics knowledge (Mullis et al., 2016b). Ratio between variance of achievement of boys over the variance of girls was 1.02, not significant according to the published literature.

4.2 Comparisons of gender differences in national marks from mathematics

Upper secondary school students can get a maximum of 5 points if they choose the basic level of the national math exam and maximum of 8 points at the advanced level. Scores from the national mathematics examination of upper secondary school students are higher for girls. As shown in the Table 2, in general, girls scored about 0.5 point (out of maximum 8 points) higher than boys. However, girls outperformed boys only among students who chose basic level of the national exam. There is no difference between boys and girls in their mean scores at the national exam at the end of elementary school at Grade 9 from mathematics, but the difference exists in scores from Slovene language, in favour of girls.

Table 2 School marks and results from national test in mathematics for Grade 13.

	School marks					National test in mathematics				
Grade 13	**Boys**	***SE***	**Girls**	***SE***	**Difference (girls–boys)**	**Boys**	***SE***	**Girls**	***SE***	**Difference (girls–boys)**
Mathematics										
Total	3.25	0.04	3.31	0.05	0.06	4.08	1.72	4.64	1.47	0.56*
Advanced exam	4.49	0.07	4.55	0.03	0.06*	6.02	1.60	6.11	1.39	0.09
Basic exam	2.82	0.04	2.97	0.05	0.15*	3.28	0.98	3.84	0.68	0.56*
Slovene language										
	3.66	0.05	3.90	0.05	0.24*					

* Difference is statistically significant at the 0.05 level.

Grade 13 girls are marked higher from Slovene language and from mathematics, when we compare two subgroups of students by the level of intended mathematics examination separately.

Although girls and boys reach similar marks at the national examination at the end of elementary school, their school marks given by teachers in classes differ. In mathematics and all science subjects, Grade 8 girls are marked significantly higher than boys (Table 3). The ratio of variances of marks of boys over girls was 1.1, slightly higher than in case of achievement, but still small.

Table 3 School marks and results from national test in mathematics for Grade 9[1].

	School marks					National test in mathematics				
Grade 8	**Boys**	**SE**	**Girls**	**SE**	**Difference (girls–boys)**	**Boys**	**SE**	**Girls**	**SE**	**Difference (girls–boys)**
Mathematics	3.46	0.03	3.63	0.03	0.17*	51.09	0.76	52.07	0.78	0.98
Biology	3.67	0.04	4.09	0.04	0.42*					
Geography	3.72	0.03	4.00	0.04	0.28*					
Chemistry	3.64	0.04	3.94	0.03	0.30*					
Physics	3.59	0.03	3.72	0.04	0.13*					
Slovene lang.	3.47	0.03	4.02	0.03	0.55*					

[1] Grade 9 was the TIMSS Grade 8 population one year later.
* Difference is statistically significant at the 0.05 level.

Different pattern of gender gap in assessments and in school marks for both populations was not expected. Students who reach each TIMSS Advanced benchmark level and took basic level of national exam scored lower than student who took advanced level, although the same math knowledge should be valued similarly (Table 4). School marks of girls who scored below high benchmark are higher than marks of similarly successful boys. The gap in marks disappears in groups of student who scored above the high TIMSS Advanced benchmark, the most able students. However, percentages of boys and girls over benchmarks vary, more boys than girls reached higher benchmark levels in both subpopulations.

Table 4 Scores from the national math exam and school marks by reached TIMSS Advanced benchmark levels of mathematics knowledge, Grade 13.

TIMSS Advanced benchmarks	**% of students**	**Mean national exam score (SE)**		**% of girls**	**Mean school mark of girls (SE)**		**% of boys**	**Mean school mark of boys (SE)**	
Basic level of the national exam									
Below intermediate	69	2.92	(0.05)	76	2.77	(0.05)*	64	2.55	(0.05)
Intermediate to High	27	3.78	(0.05)	22	3.53	(0.07)*	30	3.23	(0.07)
High to Advanced	4	4.25	(0.12)	2	4.16	(0.28)	6	3.64	(0.18)
Above advanced	0	–	–						
Advanced level of the national exam									
Below intermediate	10	4.03	(0.28)	14	4.02	(0.17)	7	3.67	(0.26)
Intermediate to High	39	5.42	(0.10)	43	4.49	(0.06)*	35	4.24	(0.11)
High to Advanced	40	6.42	(0.10)	36	4.77	(0.05)	43	4.69	(0.09)
Above advanced	11	7.21	(0.16)	7	4.80	(0.12)	15	4.92	(0.06)

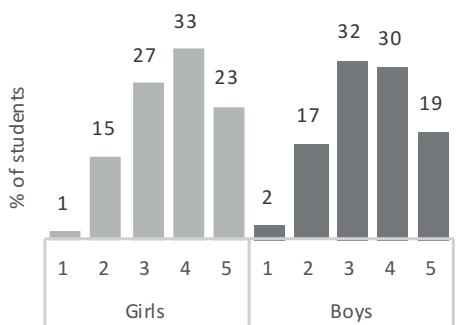
* The mean is significantly higher than the mean of the opposite sex (at the 0.05 level).

Having the mark satisfactory (2) reported 26% students, good (3) reported 29% students, very good (4) reported 23% students and excellent (5) reported 19% of students. However, students are distributed to the reached benchmark levels not consistently with their school mark. Almost half of student with excellent mark (48%) from mathematics reach at most the high knowledge in TIMSS Advanced, while 43% of girls but 64% of boys scored above the high benchmark. From this data, we can't confirm that high achievers from TIMSS are at the same the high achievers recognised by Matura examination.

Grade 8 students show no gender gap in mathematics achievement, but evidently higher marks for girls. Graphs of distributions of marks differs from almost equal distributions of TIMSS scores for both genders. Therefore, we examined if the similar pattern of lower marks for high achieving boys than for high achieving girls exists already in Grade 8. We assigned students to groups by percentiles determined through marks: the TIMSS achievement was calculated at percentiles defined by the distribution of students by school marks (Table 5) and scores used as limits of intervals for levels of knowledge. 21% of students were marked excellent and top 21% scored 572 points or more on TIMSS scale. In we estimate students who are marked differently than expected if marks and TIMSS scores would be aligned. Lower marked students are in majority boys and there are more girls among students with the highest school mark.

Table 5 Distribution of students by school marks and TIMSS achievement, Grade 8.

School marks	% of students	Percentiles of TIMSS math scores according to given school marks	% of girls	% of boys	Difference	T
Unsatisfactory (1)	2	below 372	35.47	64.53	29.06	2.73
Satisfactory (2)	16	from 372 to below 452	44.97	55.03	10.06	2.47
Good (3)	30	from 452 to below 515	45.03	54.97	9.95	3.20
Very good (4)	31	from 515 to below 572	51.04	48.96	−2.08	−0.81
Excell-ent (5)	21	at or above 572	53.84	46.16	−7.69	−2.43



Distributions of marks girls and boys

If we look at the intervals defined by TIMSS scores from Table 5 as the expected marks for students, the analysis of difference reveals that only 49% of students are marked as expected from TIMSS score (Figure 2). However, girls more often have higher given mark than it would be expected (29% vs 21%) and boys have more often lower mark than it would be expected (29% vs 22%) from their TIMSS score.

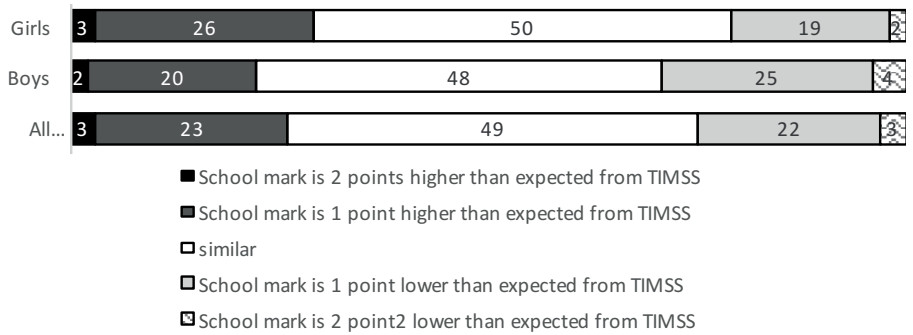


Figure 2 Distribution of percentages of girls and boys by differences between given school marks from mathematics and expected marks from TIMSS score, grade 8.

Precise data show that girls get higher marks than boys and that differences increase with marks. Out of 21% highest achievers in TIMSS, the excellent mark was given to only 56% of boys but almost 70% of girls. 9% of boys but 2% of girls received middle mark (3, good) although they were top achievers in TIMSS. We may conclude that school marks base on some additional skills and knowledge not assessed by TIMSS but giving priority to girls. This is the opposite as expected as in Slovene system, by curriculum requirements, school marks should strictly reflect the achieved standards of knowledge and not subjective teacher estimate of student effort in school.

4.3 The frameworks and structure of tests

The gender differences are higher in upper secondary schools, therefore the first step in searching for explanation for gender differences was the analysis of the content and cognitive domains covered by TIMSS Advanced and the national mathematics exam. The comparison of the *TIMSS Advanced Framework* (IEA, 2013) and the *Standards for the National Mathematics Examination* (Benko et al., 2015) shows that the national tests covered more contents than TIMSS Advanced in the area of logic, sets, probability and statistics as well as in algebra (Table 6) with trigonometry divided among functions (algebra) and shapes (geometry).

Table 6 TIMSS Advanced frameworks and national mathematics exam contents, Grade 13.

Content domains		**TIMSS Advanced**	**National exam**		
Algebra		Expressions and operations; Equations and inequalities; Functions	Expressions and operations; Equations and inequalities; Functions; Conic sections; Sequences and series; Differential of functions		
Calculus		Limits; Derivatives; Integrals	Limits; Derivatives; Integrals		
Geometry		Non-coordinate and coordinate geometry; Trigonometry	Geometry on plane and in space; shapes; Vectors on plane and in space		
Logic & sets			Basics of logic; Sets, Number sets		
Probability & Statistics			Combinatorics, Probability, Statistics		
		National Exam			
Cognitive domains	TIMSS Advanced	Written part 1 (Basic and Advanced level)	Written part 2 (Advanced level)	Oral exam (Basic level)	Oral exam (Advanced level)
Knowing	35	at least 30	at least 10	at least 30	at least 10
Applying	35	30–50	40–60	30–50	40–60
Reasoning	30	up to 30	up to 40	up to 30	up to 40

The attention to each cognitive domain in TIMSS Advanced tests is given in percentages of time for solving test items. The Matura mathematics exam covers the same cognitive domains with different attentions at the basic (BL) and advanced level (AL). The exam consists of two parts, written and oral. The written part is further divided into the first test which is the same for all students and second additional test for students who take AL. Regarding cognitive domains, the first written and oral parts for BL are similar to TIMSS Advanced test. But the second written part and oral part for AL give twice less attention to knowing and half more attention to applying and reasoning than TIMSS test.

The gender differences in percentages of student solving each TIMSS Advanced test item were mostly in favour of boys. In Slovenia, only few items out of 114 were solved better by girls. Average points of gender differences were larger for items from reasoning domain (9.21 points) and smaller for items from knowing and applying domains (5.2 and 7.1 points). Therefore, the gender difference in TIMSS Advanced is increasing with the cognitive expectation of items.

4.4 Relations between TIMSS achievement, national scores and school marks

The final score of the national exam for upper secondary school students is the sum of scores for each part. Score given at the oral part of examination contributes only with 20% to the final score. The correlation between TIMSS scores and grades from the Matura mathematics exam were found to be high for the written part and low for the oral parts of the exam. That is most certainly due to ceiling effect since most students perform excellent on the oral part and variability is rather low. The oral questioning of the exam is assessed by the teachers from the student's school and is regarded by students as not fully external although the questions are prepared nationally in advance and randomly blindly drawn from the collection by the student at the exam. The final grades of the Matura exam are also highly correlated with TIMSS Advanced scores, higher at AL than at BL (Table 7).

Table 7 Correlations between TIMSS and national scores by gender.

Correlated scores	Girls	Difference	Boys
Grade 13			
Scores for Matura, written part, BL & TIMSS Adv.	0.56	=	0.57
Scores for Matura, written part, AL &TIMSS Adv.	0.59	>	0.56
Scores for Matura, oral part, BL & TIMSS Advanced scores	0.27	>	0.24
Scores for Matura, oral part, AL & TIMSS Advanced scores	0.23	<	0.26
Matura final marks, BL (1–5) & TIMSS Advanced scores	0.56	=	0.56
Matura final marks, AL (1–8) & TIMSS Advanced scores	0.58	>	0.55
School marks (1–5) & TIMSS Advanced scores	0.61	<	0.65
School marks (1–5) & Matura final marks	0.66	<	0.76
Grade 9			
Scores for NA & TIMSS scores from Grade 8	0.77	=	0.77
School marks (1–5) & TIMSS scores from Grade 8	0.52	>	0.48

In upper secondary schools, correlations between school marks and grades from the Matura Exam are the highest. Unexpectedly, the TIMSS Advanced scores have the highest correlation with school marks although these marks are not external and come from written and oral assessments. In elementary schools, the correlation of TIMSS scores with external national exam scores is higher than with the internal school marks.

In order to describe the gender gaps, we use ordinal dominance graphs, graphical comparison of two distributions that quickly shows which group dominates and in which part of distribution (Figure 3). Results along diagonal line would indicate equal groups without any dominance. The ordinal dominance graphs by gender for the different test scores in Slovenia show equality of girls in boys on TIMSS in Grade 8, slight dominance of girls when observing results from NA in Grade 9 and larger dominance of girls when observing school marks for same population. Looking at Grade 13

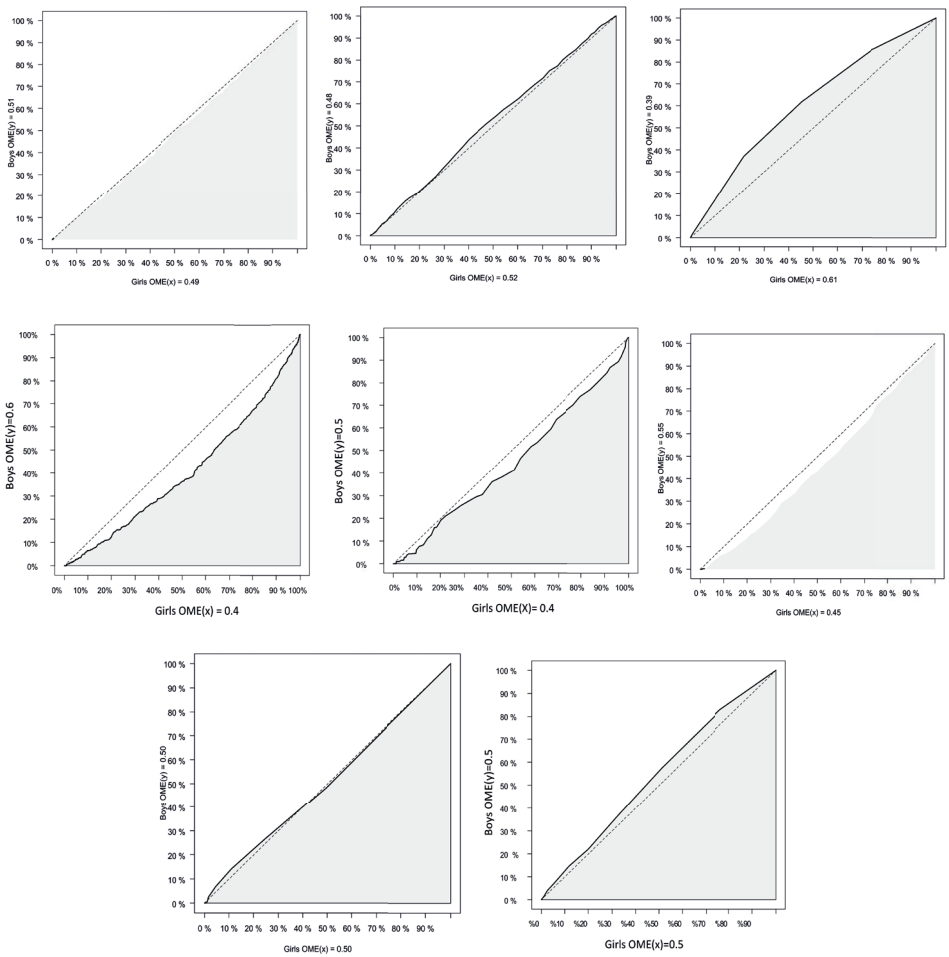


Figure 3 Ordinal dominance graphs for TIMSS and TIMSS Advanced test, national mathematics examination and school marks by gender, Grade 13.

population, results of boys dominate on TIMSS Advanced and written part of Matura at AL, while girls slightly dominate on oral part at BL.

Grade 8 girls are essentially given higher grades from teachers at mathematics lessons in schools than boys but reached similar score as boys at NA exam from mathematics and from TIMSS test. In upper secondary school, the highest gender differences in TIMSS Advanced scores occur around the middle part of the scale. Boys scored higher at the upper half of the scale for written part of the Matura exam at AL and BL, but differences are smaller than for TIMSS Advanced. Girls who reached the upper half of the scale for the oral part of the Matura Exam at BL received higher grades than boys. There was no gender difference in oral part at the AL of the Matura mathematics exam.

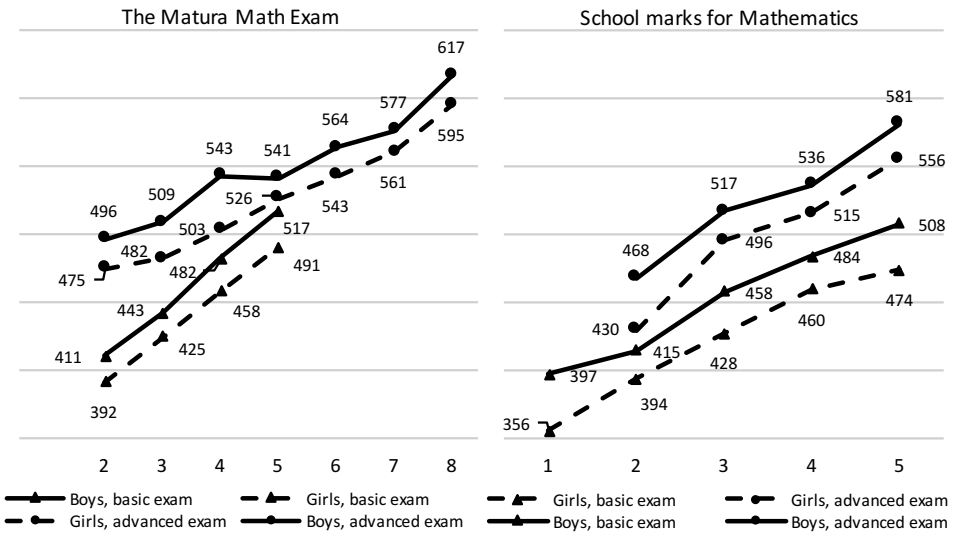


Figure 4 TIMSS Advanced achievement by the Matura exam and school mathematics marks, Grade 13. **Note: Differences** are not significant only between girls with school marks 4 and 5 taking the basic level, and between girls and boys taking the advanced level of the math exam and graded by 2 and 3 or 3 and 4.

TIMSS Advanced scores by national marks given at the Matura exam and by teachers in classes of upper secondary schools (Figure 4) describe the inconsistent links between both scores. Although having the same national score, boys reached higher TIMSS Advanced score than girls. Or, girls with similar TIMSS Advanced score as boys were given higher national mark than boys. Additionally, students with the same TIMSS Advanced score who choose the BL of the Matura math exam reached higher final national mark than students who choose the AL of the Matura math exam. For example, girls with TIMSS Advanced achievement of 475 points reached 2 points if taking the AL but 4 or 5 points if taking the BL. Boys with 517 points from TIMSS Advanced reached 5 points at BL of Matura exam but achieved 3 or 4 points if taking the AL. Results show that marking at the basic and advanced level of the Matura math exam is not completely consistent. The same pattern is seen from the comparison by school marks. Girls with TIMSS Advanced achievement of about 430 points are given mark 3-good if they intend to take basic level of the national exam and mark 2-satisfactory if they choose advanced level of the exam. Results are problematic. Regular mathematics course is following the same standards and curriculum and students are taught in mixed classes of students that choose any level of exam. Therefore, students' marks for the same math knowledge should be the same for both genders and student of both Matura levels.

5 Searching for explanations of gender gaps

From the comparison of grades and TIMSS Advanced achievement it seems that national assessments do not recognize some knowledge of boys. The Table 8 of TIMSS Advanced scores from each cognitive domain clearly shows the largest gender differences between equally marked boys and girls in reasoning. While there are no differences by gender between students of the advanced level of the Matura Exam for knowing and applying domains (scores 6, 7 and 8), in reasoning, boys with any score from the Matura math exam outperformed girls with the same score at TIMSS advanced. Also, boys with lower scores (5 or less) at Matura outperformed girls in all three cognitive domains on TIMSS Advanced.

Table 8 Gender differences of achievement in cognitive domains by the Matura exam scores, TIMSS Advanced, Grade 13.

Nat. exam score	Mean achievement – reasoning					Mean achievement – applying					Mean achievement – knowing				
	Boys	*(SE)*	Girls	*(SE)*	Diff.	Boys	*(SE)*	Girls	*(SE)*	Diff.	Boys	*(SE)*	Girls	*(SE)*	Diff.
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	402	(8.3)	373	(8.6)	29*	420	(6.9)	402	(6.6)	18*	420	(7.8)	402	(5.1)	18*
3	433	(6.2)	407	(6.1)	26*	452	(5.2)	435	(4.3)	16*	450	(7.4)	431	(4.6)	19*
4	475	(5.6)	439	(5.0)	37*	493	(5.8)	469	(3.9)	24*	491	(8.2)	468	(3.2)	23*
5	514	(8.0)	481	(5.9)	33*	533	(7.0)	509	(4.9)	23	531	(6.9)	510	(4.8)	21*
6	555	(13.1)	521	(6.3)	34*	569	(11.9)	551	(8.9)	19	568	(9.7)	551	(5.5)	17
7	570	(7.6)	544	(7.4)	26*	581	(8.8)	567	(5.5)	14	579	(9.9)	573	(6.5)	6
8	617	(9.2)	585	(8.6)	33*	623	(7.9)	602	(11.0)	21	621	(8.6)	604	(7.2)	17

* Difference is statistically significant at the 0.05 level.

These results suggest that the Matura from mathematics measures similar knowledge between boys and girls in knowing and applying domain at advanced level of exam. But it does not recognize the knowledge of boys in reasoning that was required by items in TIMSS Advanced assessment. Even worse, at BL it seems that Matura does not see and grade the mathematics knowledge of boys that was measured by TIMSS Advanced in all three cognitive domains. The findings warn that there are conceptual differences in the content domains of both tests.

Similar, in Grade 8 students with the highest excellent teacher's mark show no gender difference in TIMSS achievement across cognitive domains. However, among student with the middle mark "good", boys outscored girls in all three domains. Therefore, we conclude that assigning marks in Slovene schools is not well focused to recognise and award the intermediate knowledge of boys. Giving lower marks leads into less opportunities for placement into more demanding upper secondary schools for boys as it influences admission where they require high marks from el-

Table 9 Gender differences of achievement in cognitive domains by the teacher marks, TIMSS, Grade 8.

School mark	Mean achievement – reasoning					Mean achievement – applying					Mean achievement – knowing				
	Boys	(*SE*)	Girls	(*SE*)	Diff.	Boys	(*SE*)	Girls	(*SE*)	Diff.	Boys	(*SE*)	Girls	(*SE*)	Diff.
1	438	(10.6)	431	(14.3)	7	442	(10.0)	427	(11.8)	15	437	(9.1)	430	(12.8)	7
2	449	(5.8)	441	(6.1)	8	455	(3.8)	441	(4.8)	14*	454	(4.0)	446	(4.3)	8
3	495	(4.3)	482	(4.3)	13*	497	(3.3)	482	(3.3)	14*	498	(3.4)	486	(3.8)	12*
4	537	(4.1)	531	(3.7)	6	535	(3.5)	527	(3.1)	8*	539	(4.0)	535	(3.6)	4
5	592	(4.9)	586	(4.4)	6	585	(3.6)	578	(3.7)	7	587	(4.3)	586	(3.9)	1

* Difference is statistically significant at the 0.05 level.
Note about marks: 1 is unsatisfactory; 2 is satisfactory, 3 is good, 4 is very good, 5 is excellent.

ementary schools. From PISA 2015 data for Slovenia which covers all programs of upper secondary school students, we observe that boys systematically enter less demanding programs in notable higher percentages than girls, that ends with large 60% of female population in gymnasia.

When searching for reasons of differences between school marks and TIMSS achievement we tested the links between both scores and many background factors. Opposite to the case of Sweden, socioeconomic status, measured in TIMSS on scale of home educational resources, including material sources and parental education, in Slovenia is not linked to the inconsistencies in marking. Correlations with this and two main students' attitudes are listed in Table 10.

Table 10 Correlations of school marks and TIMSS achievement by gender, Grade 8.

	Correlation with school marks				**Correlation with TIMSS**			
	Boys		**Girls**		**Boys**		**Girls**	
	Corr.	*SE*	Corr.	*SE*	Corr.	*SE*	Corr.	*SE*
Home educational resources of student	0.33	(0.02)	0.32	(0.03)	0.34	(0.02)	0.34	(0.03)
Student's liking mathematics	0.39	(0.03)	0.42	(0.02)	0.30	(0.03)	0.31	(0.03)
Student's perception of engaging teaching	0.26	(0.03)	0.28	(0.02)	0.21	(0.03)	0.19	(0.02)
TIMSS score	0.65	(0.02)	0.68	(0.02)				
TIMSS score for knowing	0.63	(0.02)	0.67	(0.02)	0.86	(0.01)	0.86	(0.01)
TIMSS score for applying	0.62	(0.02)	0.66	(0.02)	0.87	(0.01)	0.87	(0.01)
TIMSS score for reasoning	0.61	(0.02)	0.63	(0.02)	0.85	(0.01)	0.85	(0.01)

Correlations do not differ by gender. Correlations of liking mathematics and engaging teaching of mathematics with school marks are higher than with TIMSS scores for both genders and confirm the importance of school marks for motivation of learn-

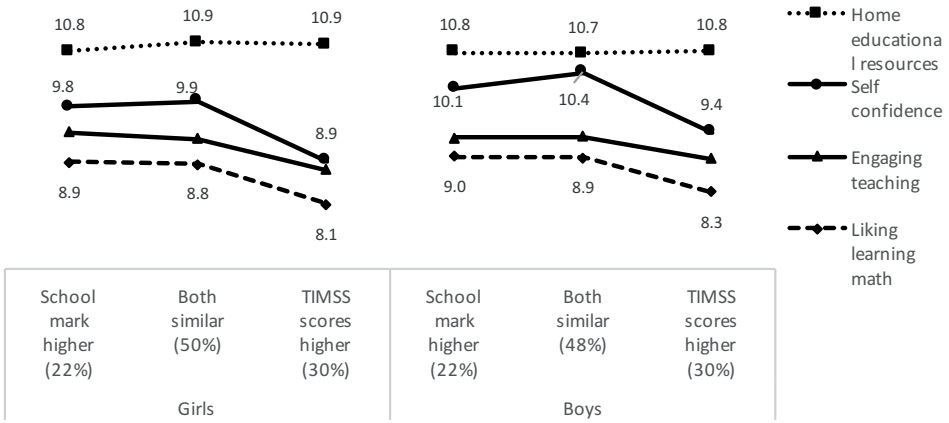


Figure 5 Background factors of learning mathematics with real and estimated marks according to TIMSS by gender, Grade 8.

ing mathematics. The differences in pattern of attitudes among girls and in among students who are marked in school differently as it would be expected from their TIMSS score can be observed in Figure 5.

Across groups of students with higher, similar or lower school mark than expected from TIMSS score, there is almost no difference in their educational home support. We may conclude that socioeconomic status of students does not impact significantly the marking of students in schools. But other three attitudes are dropping from students who have higher school mark than estimated from TIMSS achievement to students with lower school mark than estimated from TIMSS. The decreases are larger among girls, but self-confidence and liking of learning are of significant sizes also for boys. We may not discuss the linkage as causal relations but see that students who were getting lower marks in schools than they demonstrated in TIMSS assessment are less confident and like learning mathematics less than others. Low marks decrease the confidence, but less confident students do not show all their strength in assessments and therefore get lower marks. In Grade 8, there are significant gender differences in valuing mathematics – 8.86 for girls (*SE* 0.04) vs 9.07 (*SE* 0.05) for boys; and in self-confidence – 9.66 (*SE* 0.05) for girls vs 10.04 (*SE* 0.05) for boys, but not in liking learning mathematics (not significant difference of 0.05). Although different attitudes may affect teachers' subjective perceptions (Voyer & Voyer, 2014) it seems that in Slovenia teachers' higher marks for girls are not in accordance to girls' low attitudes. Some information about girls' reaction to tests provide the assessed effort put into the test among participating students in TIMSS. It was measured with five questions developed to help especially in analysing large scale assessments that have no consequences for individual students (Eklof, 2006). The question asked student how much they agree with the following statements: (a) I gave my best effort on this test, (b) I did not give this test my full attention while completing it, (c) I tried less hard on this test than

Figure 6 Effort put in TIMSS test by gender, Grade 8.

I do on other tests we have at school, (d) I worked on each item in the test and persisted even when the task seemed difficult, (e) I was motivated to do my best on this test, (f) While taking this test, I could have worked harder on it; with answer agree a lot, agree a little, disagree a little, disagree a lot. When interpreting statistics, the order of answer categories for (b), (c) and (f) should be turned. From the comparisons of answers per gender (Figure 6) it is evident that in both populations girls admitted that they have put less effort into solving the test than boys, especially among the higher achievers. Although differences are very small, girls in Grade 8 and Grade 13 (Figure 7) admitted they try more for school tests than they did for TIMSS. In general, girls with excellent marks were less motivated for TIMSS than boys with excellent marks. These data therefore support the hypothesis that Grade 8 and Grade 13 girls, especially excellent, try harder for tests with consequences for their promotion in school and boys try as hard as they can even for tests without consequences for their schooling.

In the Grade 13, the most effort was reported by boys of advanced level of national exam who also achieved the highest score and the least effort was reported by girls of basic level of national exam. Results for both populations together support the possible explanation for gender differences. Girls achieved higher scores on national exam because they tried and worked harder than for TIMSS test while middle achieving boys (basic level of NE) did not try for TIMSS tests as well as they do not try very hard for school tests either.

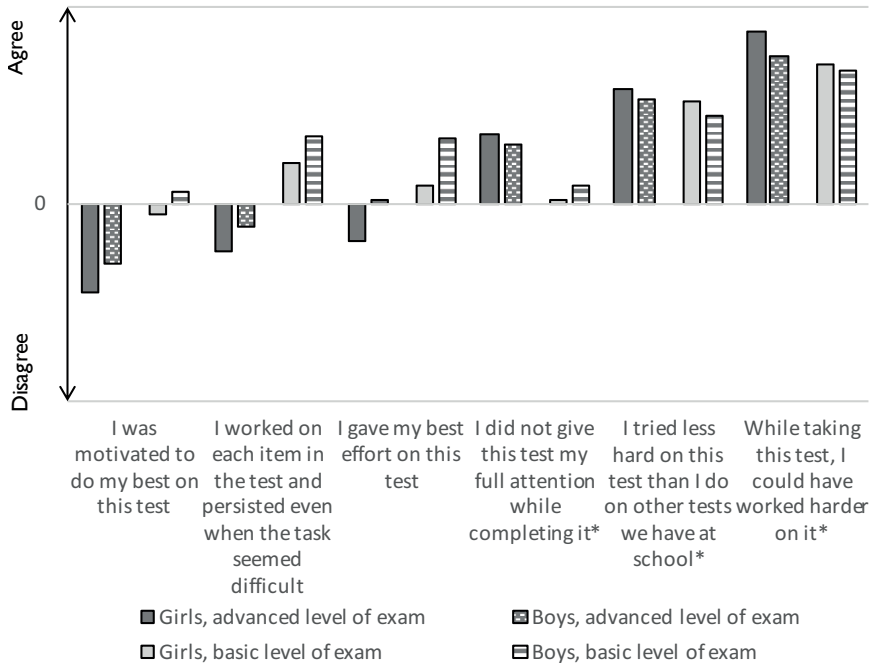


Figure 7 Effort put in TIMSS Advanced test by gender, Grade 13.

6 Conclusion

The results of our study provide some new information about the mathematics achievement of pre-university students in Slovenia and answers to our research questions. The answer to the first question is no, the highest achieving students with regard to the national scores, teacher marks and international assessments outcomes in both observed populations are found to be different groups. In general, Matura scores and school marks were not found to be consistent with TIMSS Advanced achievement, gender gap in Matura is in favour of girls and gender gap in TIMSS Advanced is in favour of boys. The results are consistent with findings in other studies. The Matura math exam in Grade 13 was found to measure similar knowledge between boys and girls who take the AL of exam in domain of knowing and applying but it does not recognize some knowledge of boys in mathematical reasoning. Among lower achievers, the national exam does not recognize and grade part of the mathematics knowledge of boys from all cognitive domains. The main difference between both assessments is oral questioning present in Matura but not in TIMSS. As the results of oral part of Matura significantly favour girls we believe that oral questioning is important factor which is also linked to the fact that girls are given relatively higher school grades. The essential part of school marks in Slovenia

is teacher's oral assessment of each individual student. The finding helps to explain gender difference between international and national assessment but also provides ideas for improvement of the Matura math exam. It suggests that mathematical knowledge of middle achieving boys, especially mathematical reasoning, would need to be better recognized by the national examination as well as by teachers if we want them to be similar to TIMSS. The changing gender gap could be partly described by the reported effort put into solving the TIMSS tests. Girls more often than boys reported that they didn't work as hard on TIMSS test and that they put slightly more effort into the national exams. The mediating effect of effort could explain higher school marks of girls, especially of high achievers, and similar or higher TIMSS achievement of boys. The measurement of effort supports the hypothesis that the larger problem of marking in school is the undermarking of boys (who most likely do not try to show all their knowledge in any test) than overmarking of girls. Next, the marking was found important as a factor of motivation which is a large problem in Slovenia, demonstrated with the very low or lowest mean national scores on any motivation scales in international comparisons and with still declining trends. In particular, in Grade 8, the motivation for learning mathematics is found to be more strongly linked to marks than to the international measurement of knowledge. School marks from mathematics contribute to the opportunities for student's placement in upper secondary school. Therefore, the above mentioned problematic side of marking could also contribute to the lower percentages of boys than girls choosing higher demanding upper secondary school programs and therefore it could be one of the reasons for observed lower overall achieved upper secondary education of boys compared to girls in Slovenia. The evaluation of marking system, especially with regard to middle achieving boys, is clearly needed.

With regard to known results from research literature, gender differences of achievement in Slovenia do not differ much from other countries. They show almost similar achievement in Grade 8 and higher boy's achievement in Grade 13 while school marks were found to be higher for girls, similar to already found pattern in some other countries. By linking TIMSS achievement and school marks we were able to recognise some problems of marking students on national level, after taking into account that differences in students' individual characteristics contribute to a significant extent to gender differences in any school performance.

The data used for this study are limited to the international databases of TIMSS and TIMSS Advanced with few additional answers to the national questions for students. Therefore, they cannot provide all needed information for an extended study of individual or group student characteristics on achievement or on the gender gap. The study found some basic facts which will be studied further, most likely together with teacher characteristics and ways of their assessments of student knowledge in mathematics and science.

In this research girls were found to be somewhat better adapted to today's school environments, as research literature suggests, most likely because of their better verbal intelligence, higher agreeableness, stronger self-discipline, as well as certain

aspects of their motivation (Spinath, Eckert, & Steinmayr, 2014). In light of these specific differences, it could be expected that changing certain aspects of school environments with regard to stimulating boys' motivation and engagement might help boys to better succeed in school and, thus, reduce our national educational inequality.

References

Baye, A. & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-Scale Assessment in Education 4*(1). https://doi.org/10.1186/s40536-015-0015-x

Benko, D., Erker, J., Hvastija, D., Jan, M., Miler, A., Robnik, A., Škof, M., & Žerovnik, J. (2015). *Predmetni izpitni katalog za splošno maturo – matematika* [Standards for the National Examination – Mathematics]. National Examinations Centre, Ljubljana. Retrieved from http://www.ric.si/mma/201720M-MAT-2017/2015083113005248.

Bergold, S., Wendt, H., Kasper, D., & Steinmayr, R. (2017). Academic competencies: Their interrelatedness and gender differences at their high end. *Journal of Educational Psychology, 109*(3), 439–449. https://doi.org/10.1037/edu0000140

Burusic, J., Babarovic, T., & Seric, M. (2012). Differences in elementary school achievement between girls and boys: Does the teacher gender play a role? *European Journal of Psychology of Education*, *27*(4), 523–538.

Cankar, G. (2010). PISA 2006 main findings in Slovenia and Simpson's paradox. V: XIV. IOSTE Symposium, International Organization for Science and Technology Education, June 13–18, 2010, Bled, Slovenia. In S. Dolinšek (Ed.), *Socio-cultural and human values in science and technology education: proceedings*. Institute for Innovation and Development of University. Ljubljana.

Cankar, G. (2016). Governing the transition to higher levels of education and differences between achievement and school grades by gender. *The Journal of Educational Research*, *8*(2), 60–86.

Ceci, S. J., Williams, W. M. & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, *135*(2), 218–261.

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, *66* (4), 643–656.

Else-Quest, N. M, Shibley Hyde J., & Linn, M.C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127.

Holman L., Stuart-Fox D., & Hauser C. E. (2018). The gender gap in science: How long until women are equally represented? *PLOS Biology 16*(4), e2004956. https://doi.org/10.1371/journal.pbio.2004956

Houtte, M. V. (2004). Why boys achieve less at school than girls: The difference between boys' and girls' academic culture. *Educational Studies*, *30*(2), 159–173.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581–592.

IEA (2017a). *TIMSS Advanced SPSS Data*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from https://timssandpirls.bc.edu/timss2015/advanced-international-database.

IEA (2017b). *TIMSS SPSS Data*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from https://timssandpirls.bc.edu/timss2015/international-database.

Klapp Lekholm A., & Cliffordson C. (2008). Discrepancies between school grades and test scores at individual and school level: effects of gender and family background, *Educational Research and Evaluation*, *14*(2), 181–199. https://doi.org/10.1080/13803610801956663

Lindberg, S. M., Hyde, J. S., Petersen, J. L. & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135.

Mullis, I. V. S. & Martin, M. O. (Eds.). (2013). *TIMSS 2015 assessment frameworks*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2015/frameworks.html.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016a). *TIMSS Advanced 2015 International Results in Advanced Mathematics and Physics*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/advanced.

Mullis, I. V. S, Martin, M. O., Foy, P. & Hooper, M. (2016b). *TIMSS 2015 International Results in Mathematics*. Boston College, TIMSS & PIRLS International Study Center. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results.

Nosek, A. B., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., … , Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, 106*, 10593-10597. https://doi.org/10.1073/pnas.0809921106

OECD (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. Pisa, OECD Publishing. https://doi.org/10.1787/9789264229945-en

Rutkowski, D., Rutkowski L., & Plucker, J. A. (2012). Trends in education excellence gaps: a 12-year international perspective via the multilevel model for change. *High Ability Studies*, *23*(2), 143–166. https://doi.org/10.1080/13598139.2012.735414

Spinath, B., Eckert, C., & Steinmayr, R. (2014). Gender differences in school success: what are the roles of students' intelligence, personality and motivation? *Educational Research*, *56*(2), 230–243. https://doi.org/10.1080/00131881.2014.898917

Steinmayr, R. & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality*, *22*(3), 185–209.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204.

Barbara Japelj Pavešić
Educational Research Institute
Gerbiceva 62
1000 Ljubljana
Slovenia
Barbara.Japelj@pei.si

Gašper Cankar
National Examinations Centre
Kajuhova ulica 32u
1000 Ljubljana
Slovenia
Gasper.Cankar@ric.si

Factors Explaining the Interest of Czech Students in Reading and Mathematics

Eva Potužníková
Charles University, Faculty of Education

Abstract: The goal of the empirical study is to identify significant predictors of student interest in reading and mathematics using data from international large-scale assessments. According to studies of interest development in educational settings, certain instructional techniques are able to evoke situational interest, whereas personal relevance and active involvement are sources of maintained interest. This study compares the effect of engaging instruction with the effect of student-related characteristics, such as gender, family background, free time preferences and perceived difficulty of the subject. The analyses were performed on PIRLS and TIMSS 2011 data for Grade 4 students from the Czech Republic separately for reading (*N* = 4556) and mathematics (*N* = 4578). In addition, data from a national follow-up study in Grade 6 was used to study interest development (*N* = 2955 for reading, *N* = 2956 for mathematics). Engaging instruction is positively associated with student interest in Grade 4 in both domains. The percentage of students declaring a positive attitude is close to 80% in both domains. A slight decrease in interest levels between Grades 4 and 6 was identified. While the most powerful predictor of interest in reading in Grade 6 is the former interest level, interest in mathematics is best predicted by perceived difficulty. Implications for instructional practice are also discussed.

Keywords: student interest; reading; mathematics; engaging instruction; PIRLS; TIMSS

Czech Republic has been participating in the activities of the International Association for the Evaluation of Educational Achievement (IEA) since the first round of TIMSS in 1995. Our country has also been involved in the assessments conducted by the Organisation for Economic Co-operation and Development (OECD) in student (PISA), teacher (TALIS) and adult (SIALS, PIAAC) populations. Studies carried out among students[1] have brought a wide range of internationally comparable data on their educational outcomes, conditions for learning and classroom activities. Not surprisingly, country results in cognitive tests always raise the greatest attention. On the other hand, results concerning student motivation, attitudes and other non-cognitive outcomes are rather neglected in the Czech Republic (Straková, 2016), although they might provide relevant information on the capacity of the school system to achieve important educational goals according to the Education Act.

This article aims to address student interest in reading and mathematics as educational outcomes that can be supported or inhibited by school instruction.

[1] This article uses the term “students” instead of “pupils” to denote children enrolled in primary and secondary schools regardless of the grade level. This is in line with the terminological conventions applied in the official reports from international large-scale studies.

https://doi.org/10.14712/23363177.2018.295
www.orbisscholae.cz

Chapters about student interest and other motivational aspects are integral part of reports from every international large-scale study. However, student motivation is seen mostly as a precondition that explains why some students achieve better than others, not as a specific outcome that has to be explained. In a different research tradition, educational psychologists devoted immense efforts to identify what aspects of school instruction can promote student interest. This article wants to build a bridge between these two strands of educational research. It will use data from international large-scale assessments to answer questions that are more typical to the research of interest development.

1 Conceptual background

1.1 Role of motivational beliefs in educational achievement and aspirations

Student motivation to learn and perform well at school can be decomposed into different components. One of the most influential theories of motivation in the field of education, expectancy-value theory (Eccles et al., 1983), differentiates two basic sets of motivational beliefs: expectancies of success and task values. Expectancies refer to one's perceived ability to accomplish a given task and are conceptually similar to self-concept and self-efficacy as defined in social cognitive theory of motivation (Bandura, 1977; Pajares & Miller, 1994). Task values are subjective perceptions of how valuable the task is. There are different types of subjective task values: attainment value or importance of doing well on a given task, interest value or enjoyment from doing the task, utility value or usefulness of the task for one's future goals, and costs or subjective assessment of effort necessary to accomplish the task (Wigfield & Eccles, 2000). Interest value is similar to the construct of intrinsic motivation from the self-determination theory (Deci & Ryan, 1985), whereas utility value refers to extrinsic reasons for action.

This article focuses on the interest component of achievement motivation. According to the expectancy-value theory, both expectancies and values are considered as important prerequisites of educational achievement. While the perceptions of one's ability ensure that the goal is experienced as attainable, value-related perceptions support persistence and commitment to the goal (Korhonen et al., 2016). Empirical evidence on the relationship between interest and achievement is, however, not so straightforward.

In a multi-cohort study conducted by the authors of the expectancy-value theory, children's competence beliefs strongly predicted their competence beliefs in the next year as well as their grades. On the contrary, students' interest predicted their next year's interest, but not the grades (Wigfield & Eccles, 2000). Cortright, Lujan and Blumberg (2013) found that interest was associated with higher grades for male students but not for females. A German study on mathematics showed that interest

had no significant effect on achievement between Grade 7 and Grade 10 but more interested students tended to choose advanced courses at upper secondary level. Furthermore, interest in Grade 10 had both direct and indirect (via course selection) effect on achievement in Grade 12 (Köller, Baumert, & Schnabel, 2001). Other more recent studies confirmed the effect of interest on educational choices (Gottfried et al., 2013; Nagy et al., 2006) and aspirations (Korhonen et al., 2016), whereas academic achievement tends to be linked more closely to self-concept than to interest (Nagy et al., 2006).

Even though its effect on student achievement may be lower than one would anticipate, there is a general consensus that interest facilitates learning (Renninger & Hidi, 2011), improves the quality and depth of the learning process (Savelsbergh et al., 2016) and compensates for the lack of skills when solving difficult tasks (Springer, Harris, & Dole, 2017). The importance of student motivation for their achievement might be particularly relevant for young children in the domain of reading (Mullis & Martin, 2015). A positive attitude towards reading is also assumed to be one of the most important attributes of a lifelong reader (Mullis et al., 2009a). The role of interest in reducing achievement gaps and course selection differences between boys and girls was also documented (Gustafsson, Yang Hansen, & Rosén, 2013; Nagy et al., 2006). Interest can even mitigate, although to a more limited extent, the influence of socioeconomic background on student achievement (OECD, 2010). To sum up, interest in school subject matter is an important non-cognitive educational outcome that improves academic achievement, affects career choices and fosters lifelong learning.

1.2 Development of student interest

Numerous studies have identified a general decrease of interest in school subjects as students pass to higher levels of schooling (Krapp, 2002; Renninger & Hidi, 2011). Most of the research has been based on cross-sectional and short-term longitudinal designs, but similar results were reported for longitudinal studies, as well (e.g., Fredricks & Eccles, 2002). The loss of interest can be explained by increased task complexity, higher demands for effort and changes in social relationships during adolescence (Frenzel et al., 2010). Another factor could be a more frequent use of traditional instructional techniques in higher grades (Fredricks & Eccles, 2002).

The loss of interest applies especially to mathematics and science (Frenzel et al., 2010; Gläser-Zikuda, Stuchlíková, & Janík, 2013; Gottfried et al., 2013; Savelsbergh et al., 2016). Low levels of student motivation in STEM (Science, Technology, Engineering and Mathematics) subjects has even become a major concern of educational policy in many countries (Kearney, 2016), as expertise in STEM subjects is seen as a necessary precondition for economic progress. Another consistent finding is that boys are more interested in mathematics than girls (Frenzel et al., 2010; Köller et al., 2001), but the gender gap may not intensify as students grow older (Fredricks & Eccles, 2002; Frenzel et al., 2010).

The mechanism of interest development in learning environments has been extensively studied by educational psychologists. Hidi and Renninger (2006) proposed a four-phase model of interest development with phases of triggered situational, maintained situational, emerging individual and well-developed individual interest. A recent study by Rotgans and Schmidt (2017) demonstrated that, indeed, situational interest led to the development of individual interest. Similarly, Krapp (2002; 2007) distinguished between situational interest and individual interest (as a personal trait) and suggested a three-step ontogenetic transition from the first to the latter with an intermediate step of stabilized situational interest.[2]

Many researchers have tried to identify what aspects of school instruction have the potential to raise student interest. Whereas hands-on activities, group work, novelty and changes in the environment are among the most cited sources of situational interest, personal relevance and active involvement tend to support longer lasting interest (Renninger & Hidi, 2011). As Springer, Harris and Dole (2017) point out, so-called catch activities that sparkle students' situational interest must be followed by something more meaningful that will maintain their interest for a longer time. The teacher's emotional involvement, enthusiasm and his/her personal belief about the value of the learning material were also positively related to students' interest (Frenzel et al., 2010; Gläser-Zikuda et al., 2013). On the other hand, classroom practices like public praise and criticism, public drill or the use of competitive approaches tend to undermine the initial interest of students (Frenzel et al., 2010).

Research carried out in the Czech Republic has confirmed the general pattern of declining interest in reading and mathematics as children grow older. Whereas 93% of girls and 74% of boys aged 8–9 years liked reading, only 67% of girls and 35% of boys aged 14–15 years did so (Ronková, 2015). The author explains the weakening interest in reading among older students, especially boys, by the fact that reading as a time-intensive activity has to compete with other free time activities and loses its appeal when confronted with some less demanding and more tempting entertainments, in particular computer games. Interestingly, internet was not identified as a direct "rival" of reading for children; it competes rather with TV watching.

Moving to the domain of mathematics, Chvál (2013) examined students' attitudes towards mathematics using the method of semantic differential. He found a decrease in liking mathematics, with the most pronounced drop between Grades 5 and 6. The generally decreasing trend continued at the upper secondary level. By contrast, students' attitude towards Czech language declined up to Grades 6 and 7, but then it increased to more positive values. Foreign language was perceived positively, without dramatic changes between different years of schooling. Pavelková and Hrabal (2012) relate low level of interest in mathematics to its perceived difficulty. In their study of attitudes towards school subjects among Czech students at

[2] Although the prototypical trajectory goes from situational to individual interest, an opposite process of arousing situational interest on the basis of a strong individual interest can also be observed (Krapp, 2002). For example, students' interest in reading can be raised when the teacher offers them books on topics they are already interested in (Springer, Harris & Dole, 2017).

lower secondary level (Grades 6 to 9), mathematics was perceived as the most difficult and the third most unpopular subject. The development of students' attitudes towards mathematics was characterized by a decreasing popularity after Grade 6, coinciding with worsening marks and growing perceived difficulty.

1.3 Measurement of student interest in international large-scale assessments

Both the general public and educational research community appreciate international large-scale assessments as a valuable source of information on student achievement in comparison to other countries. Along with the widely followed results on academic achievement, non-cognitive educational outcomes are also assessed. The conceptualization of student motivation builds on prominent psychological theories and provides a solid basis for a detailed investigation of student self-concept, self-efficacy and interest. Secondary analyses of released datasets can profit not only from high-quality data collected on representative student samples, but also from repeated administration of the same items in consecutive data collections to observe the change of student attitudes in time. Another advantage is the possibility to link data on students' motivation with their cognitive achievement, family background and variables related to teaching and instruction.[3]

Student interest is generally measured through students' agreement or disagreement with statements affirming that they like reading, mathematics or science, that they are interested in solving mathematics or science problems, that they would like to have more time for reading, etc. Also included are items expressing a negative attitude, such as "I read only if I have to". Students' answers to individual items are then combined to summary scales, after re-coding of negative items. The scales are part of the final dataset and can be directly used for secondary analyses. Alternatively, individual items can be analysed.

A typical finding on student interest published in international and national reports consists of country comparisons of mean values and gender differences on interest scales. Interest is also routinely correlated with achievement. Generally, the more interested the students are, the higher levels of achievement they show, although the association between self-concept and achievement tends to be stronger than the correlation between interest and achievement (see also Chvál, 2013). Girls are more likely to show higher interest in reading than boys, whereas boys are more interested in mathematics than girls. Interest as outcome variable and its

[3] The measurement of different aspects related to teaching and instruction by means of teacher questionnaires was traditionally a distinctive feature of IEA studies. The OECD PISA study has recently also recognized the importance of teacher variables in explaining student achievement. In 2015, PISA included two optional questionnaires for teachers of science and other subjects. PISA 2018 can be optionally linked to the OECD TALIS study (Teaching and Learning International Survey).

relationship to teaching practices or other variables related to school instruction, while assumed, are typically not examined.

1.4 Aims of the present study

This study aims to explore the potential of instruction-related variables to explain student interest in reading and mathematics. More specifically, it compares the effect of engaging instruction with the effect of student personal characteristics, such as gender, family background, perceived difficulty of the subject and free time preferences.

Following the work of McLaughlin et al. (2005), engaging instruction was introduced as a new measure in TIMSS and PIRLS 2011 to describe the cognitive interaction between the student and instructional content. This measure complements the information on the use of various instructional techniques and strategies and connects the instruction with curriculum (Mullis et al., 2012a), which has been a central category of all IEA studies (Mullis et al., 2009b). The original concept of student content engagement, as defined by McLaughlin and her colleagues, was intended as a general framework for research on teaching quality, i.e. as a tool for defining and organizing teacher characteristics that contribute to better learning and higher achievement levels. Accordingly, engaging instruction was included in TIMSS and PIRLS 2011 as a potential teacher-related predictor of student achievement (Mullis & Martin, 2015). However, its association with student achievement tends to be rather small, at least for the Czech Republic (Mullis et al., 2012a, b). Nevertheless, engaging instruction seems to be a promising construct for analysing the role of the teacher in arousing and maintaining student interest.

This study seeks answers to the following research questions:

1. What is the effect[4] of engaging instruction on student interest in reading and mathematics comparing to the effect of student personal characteristics, such as gender, family background and perceived difficulty of the subject matter?
2. Does the effect of engaging instruction at the primary level endure to the lower secondary level?
3. Is the effect of different variables on student interest comparable for both domains?

[4] The term “effect” is used in the statistical sense as the relationship between a predictor and the outcome variable when all other predictors are held constant. Cross-sectional data collected in one time point, as is the case of all international large-scale assessments, do not allow to draw conclusions about causal effects.

2 Method

2.1 Data

The primary data source is TIMSS and PIRLS 2011 data for Grade 4 students from the Czech Republic. TIMSS is an IEA study of mathematics and science, which is organized every four years. It targets at Grade 4 and Grade 8 students, but only the younger population of fourth-graders participates now in the Czech Republic.[5] PIRLS is an IEA reading literacy study, which repeats every five years and measures reading comprehension skills of Grade 4 students only. In 2011 the cycles of the two studies met, which allowed to optionally include the same students in both of them. The Czech Republic used this option. Therefore, the datasets from the two studies can be combined together via a common (anonymised) student ID code.

The analyses for this study were conducted separately for reading (N = 4556) and mathematics (N = 4578). Slight difference in the numbers of participants is caused by the fact that some students could not attend both administrations. This study uses the link between PIRLS and TIMSS only to merge data from PIRLS parent questionnaire with TIMSS student questionnaire data. Parent questionnaire is a unique source of information on family background, which is normally not administered in TIMSS.

Analyses concerning the transition from primary to lower secondary education use data from the Czech Longitudinal Study in Education (CLoSE). CLoSE is a multi-cohort 7-year research project that focuses on the formation of skills, attitudes and preferences during school attendance and their role at the labour market. One of the cohorts included in the project consists of students who participated in TIMSS and PIRLS 2011 and were later contacted at several points of their educational career. They completed a questionnaire in Grade 5 and a test and questionnaire at the beginning of Grade 6. As some students transited to 8-year academic track after the completion of five years of primary education, their new classmates were added to the sample to collect more information about the differences between the standard and academic tracks. The next follow-up was in Grade 9 in both school types. This article analyses questionnaire data from Grade 6 students with disponible data from Grade 4 (N = 2955 for reading, N = 2956 for mathematics).

2.2 Measures

Student interest in reading/mathematics

Student interest in Grade 4 was measured with summary scales created by the TIMSS and PIRLS international study centre. These scales were included in the respective datasets under variable names ASBGSLR (*Students like reading*) and ASBGSLM (*Students like learning mathematics*). The original English wording of items used to

[5] The inclusion of Grade 8 students in TIMSS has no longer been considered as a political priority after the introduction of the OECD PISA study. PISA targets at the population of 15-year-old students, who are typically enrolled in Grades 9 and 10 (cf. Straková, 2016, p. 32).

construct these scales is given in the Appendix (see Martin & Mullis, 2012 for further details). Czech translation of the items as adopted in the national versions of PIRLS and TIMSS student questionnaires is also reported to increase the transparency of the present study for readers from the Czech Republic.

Student interest in Grade 6 was assessed through questionnaire items administered within the CLoSE study in autumn 2012. For reading, three items were identical as in Grade 4. These three items were selected to create a scale of student interest in reading. The scale was created by means of principal component analysis in SPSS, without rotation. The first principal component explained 77% of the variance, the items were highly inter-correlated (Cronbach's $\alpha = .85$).

Similarly, student interest in mathematics was constructed as the first principal component of four items (explained variance 72%, Cronbach's $\alpha = .87$). None of them was identical to those used in Grade 4. Three items (see the Appendix for their wording in Czech and translation into English) were scored using a 4-point agreement Likert scale. The fourth item assessed the popularity of mathematics among other school subjects on a 5-point scale ranging from most popular to least popular. The items were recoded so that higher values represent higher interest.

Engaging instruction

Engaging instruction was introduced as a new concept in TIMSS and PIRLS 2011 to capture cognitive interaction between the student and instructional content. For each study, two complementary scales were developed to measure student engagement during the classroom instruction. The first one looked at student engagement from the teacher point of view and contained different teaching practices intended to raise student interest and reinforce their learning. The second one represented the students' perceptions of classroom instruction in terms of how interesting and clear it was.

This article uses the student perspective to measure engaging instruction (student variables ASBGERL for reading and ASBGEML for mathematics). Individual items constituting the scales are described in the Appendix together with their Czech equivalents. One reason for the selection of the student-based rather than the teacher-based scale is that it generates greater variability in the student-level data. It also reflects the fact that some teaching methods may work well for some, but not for other students, depending on their learning styles, prior experience, ability and other factors. The student-based scales also had higher internal consistency and explained more variance than the corresponding teacher-based scales (see Martin & Mullis, 2012 for more technical details about psychometric properties of the scales). Engagement in classroom instruction was not measured in Grade 6.

Perceived difficulty of reading/mathematics

Following the work of Pavelková and Hrabal (2012), perceived difficulty was selected as a variable with a possible significant effect on student interest, especially in the domain of mathematics. Pavelková and Hrabal used one item to assess perceived

difficulty of different school subjects on a 5-point Likert scale ranging from very difficult to very easy. The present study uses summary scales derived from several items by means of principal component analysis. It was not possible to use the same items for both domains and both grades, due to different content of the questionnaires.

The scale of perceived difficulty of reading in Grade 4 was derived from four items: I usually do well in reading, Reading is easy for me, Reading is harder for me than for many of my classmates, Reading is harder for me than any other subject, coded such that higher values represent higher difficulty (explained variance 52%, Cronbach's α = .77). In Grade 6, five items were used: I usually do well in reading, Reading is easy for me, I sometimes have troubles to exactly understand what I read, I have to read the text more than once to understand it properly, I understand well and easily what the text says, coded such that higher values represent higher difficulty. This scale had lower internal consistency (Cronbach's α = .42) and explained less variance (34%) than other summary scales created for the purpose of this study. However, I decided to keep it because the items describe quite precisely the typical reading comprehension difficulties of Grade 6 children. An alternative scale containing only the first two items (which were taken from the PIRLS questionnaire) had better psychometric properties, but conceptually could not serve as a measure of perceived difficulty of reading in Grade 6.

The scale of perceived difficulty of mathematics in Grade 4 was derived from four items with similar wording as in the case of reading (explained variance 65%, Cronbach's α = .82). The questionnaire for Grade 6 students did not specifically focus on perceived difficulty of mathematics and contained only three suitable items. Two (I was always good at mathematics, I have good marks in mathematics) were scored using a 4-point Likert agreement scale, one assessed the difficulty of mathematics among other subjects on a 2-point scale difficult vs. easy. The summary scale constructed from these three items explained 72% of the variance and had high internal consistency (Cronbach's α = .79). Both English and Czech wording of items used to measure perceived difficulty is given in the Appendix.

Other variables

Other variables whose effect on student interest was tested in the study included time spent on PC games, time spent on TV (Grade 4) / TV or video (Grade 6), gender and family background. Time spent on PC games and time spent on TV/TV or video were measured along with other free time activities on a 4-point frequency scale ranging from not at all to 4 hours a day or more (in Grade 4) and from no time to more than 3 hours a day (in Grade 6). Family background was measured by the PIRLS *Home resources for learning* scale (ABSGHRL), which synthetizes the information about parents' education, parents' occupation, number of books at home and two additional study supports – internet connection and children's own room (see Martin & Mullis, 2012 for more information). Table 1 presents descriptive statistics for all variables used in the study.

Table 1 Descriptive statistics of the study variables.

Variable (grade)	Mean[a]	*SD*	Reliability[b]	Items in scale	Source
Interest in reading (4)	10.00	2.13	.85	8	PIRLS, original scale
Interest in math. (4)	9.84	1.95	.86	5	TIMSS, original scale
Engag. instr. reading (4)	9.70	1.98	.77	7	PIRLS, original scale
Engag. instr. math. (4)	10.16	2.00	.71	5	TIMSS, original scale
Difficulty of reading (4)	0.00	1.00	.77	4	PIRLS, own calculation
Difficulty of math. (4)	0.00	1.00	.82	4	TIMSS, own calculation
Family background (4)	10.51	1.45	.69	5	PIRLS, original scale
Gender (4)	0.51	0.50	–	–	PIRLS, TIMSS
Time on PC games (4)	2.36	0.85	–	–	PIRLS, TIMSS
Time on TV (4)	2.63	0.72	–	–	PIRLS, TIMSS
Interest in reading (6)	0.00	1.00	.85	3	CLoSE, own calculation
Interest in math. (6)	0.00	1.00	.87	5	CLoSE, own calculation
Difficulty of reading (6)	0.00	1.00	.42	5	CLoSE, own calculation
Difficulty of math. (6)	0.00	1.00	.79	3	CLoSE, own calculation
Time on PC games (6)	2.55	0.93	–	–	CLoSE
Time on TV or video (6)	2.75	0.75	–	–	CLoSE

[a] The original TIMSS and PIRLS scales were standardized to have international mean 10 and standard deviation 2, scales created for the purpose of this study were *z*-standardized.
[b] Cronbach's α of the original TIMSS and PIRLS scales for each participating country is reported in Martin and Mullis (2012).

2.3 Statistical analyses

Several linear regression models were fitted to answer the three research questions. The analyses were conducted in SPSS (version 20) using syntax files created by the IEA IDB Analyzer (version 4.0.21).[6] IDB Analyzer is a software developed by the IEA Research and Analysis Unit in Hamburg for processing of large-scale assessment data. It takes into account the complex sampling and assessment design and computes correct parameter estimates together with correct standard errors.

Regression parameters were estimated separately for reading and mathematics. In the first step, models for Grade 4 students were run using PIRLS and TIMSS student datasets for the Czech Republic to which a scale of family background was added from the parent questionnaire data. A set of national items including questions about free time activities was part of the original datasets. In the second step, models for Grade 6 students were run using a sub-sample of the CLoSE dataset containing students who had records for both grades. The data was weighted by the appropriate total student weight, which was included in the Grade 4 datasets. Weighting by

[6] http://www.iea.nl/data

a weight calculated for Grade 6 dataset to reflect the changes in the data structure led to similar results.

3 Results

Several linear regression models were fitted to estimate the relative strength of variables related to classroom instruction and student personal characteristics to predict student interest in reading and mathematics. The following sections present standardized regression coefficients and their standard errors. Statistical significance is reported at .05 confidence level.

3.1 Effect of engaging instruction on student interest in Grade 4

The initial step to answer the first research question consisted in performing several analyses whose results were then compared. Table 2 shows results of two models estimated for reading. Model 1 contains two conceptually relevant variables, namely engaging instruction and perceived difficulty of reading, and two other student-related variables that served as controls (gender and family background). Model 2 adds time spent on PC games and TV watching as two typical free time entertainments that potentially compete with reading.

All predictors are statistically significant at .05 confidence level. Engaging instruction is the most powerful predictor, suggesting that certain instructional activities, such as bringing attractive texts to the classroom, setting clear and interesting tasks or explaining things clearly (the exact description of these activities is given in the Appendix), are strongly associated with higher interest in reading among students. Perceived difficulty of reading is inversely related to interest, but the relationship is only half as strong. This suggests that engaging instruction can stimulate interest in reading even among children with reading difficulties. Both playing computer games and TV watching have a small negative effect on reading interest above the effect of other variables. They also partially explain the role of family background and gender in the sense that lower interest in reading among boys and children from disadvantaged families can be partly attributed to their free time preferences. An additional (unpublished) model tested also the role of watching videos or DVDs, with an insignificant effect.

Similar models were specified for the interest in mathematics (Table 3). It was not supposed that playing computer games or TV watching would be related to interest in mathematics, but these variables were included for comparative purposes. The results for interest in mathematics differ mainly in that perceived difficulty has approximately the same effect (in absolute values) as engaging instruction. This means that teaching activities intended to engage students are associated with higher interest in mathematics, but only for students who do not perceive it as difficult. Boys and girls have approximately the same interest in mathematics when controlled for

other variables. This is not surprising given that gender difference was not significant already without other controls in TIMSS 2011.

The correlations between individual predictors were also calculated to see whether multicollinearity could be a problem. Multicollinearity occurs when an independent variable is highly correlated with one or more of the other independent variables in a multiple regression model. Multicollinearity is a problem because it increases the sensitivity of the regression coefficients to small changes in the model specification and complicates the interpretation of the results. In the case of models presented in this section the intercorrelations between independent variables were not high. The highest correlations were found between engaging instruction and perceived difficulty of mathematics (−.36) and time spent on PC games and time spent on TV watching (.34).

Table 2 Linear regression models predicting student interest in reading − Grade 4.

	Model 1		**Model 2**	
	Beta	*SE*	*Beta*	*SE*
Engaging instruction	.40*	.02	.38*	.02
Perceived difficulty of reading	−.22*	.02	−.23*	.02
Family background	.15*	.02	.13*	.02
Gender (boy)	−.19*	.01	−.16*	.02
Time spent on PC games			−.08*	.02
Time spent on TV watching			−.07*	.02
N (listwise)	4223		4017	
R^2	.34		.36	

* $p < .05$

Table 3 Linear regression models predicting student interest in mathematics − Grade 4.

	Model 1		**Model 2**	
	Beta	*SE*	*Beta*	*SE*
Engaging instruction	.42*	.02	.42*	.02
Perceived difficulty of mathematics	−.44*	.02	−.45*	.02
Family background	−.06*	.01	−.06*	.01
Gender (boy)	−.01	.01	.00	.02
Time spent on PC games			−.03	.02
Time spent on TV watching			−.01	.02
N (listwise)	4148		4001	
R^2	.51		.51	

* $p < .05$

3.2 Changes in student interest at the transition from primary to lower secondary education

First of all, a series of descriptive comparisons between Grade 4 and Grade 6 was performed to describe the change of student interest in time. For reading, three identical items were used in both grades. The percentages of students declaring positive attitudes to reading (strong or little agreement with positively formulated statements and strong or little disagreement with a negatively formulated statement) decreased by 10 to 20%. For example, student agreement with a statement "I enjoy reading" dropped from 80 to 69%. For mathematics, none of the items used in Grade 6 corresponded exactly to items administered in Grade 4. However, a rough comparison of student agreement with items "I enjoy learning mathematics" (Grade 4) and "I don't want to give over mathematics because I enjoy it" (Grade 6) showed a similar decrease in student interest from 77 to 68%.

In the next step, regression models were specified to see what factors can be held responsible for the interest decrease. Only the coefficients of full models are presented here (Table 4). The analyses were performed on longitudinal data of students for whom the answers from both Grades 4 and 6 were available. This allowed to include prior student interest as an additional predictor. Unfortunately, the measure of engaging instruction was available only for Grade 4. It was therefore not possible to estimate the association between the momentary engagement in classroom instruction and student interest in Grade 6. Instead, an enduring effect of previous instruction was analysed.

As in the previous section, correlations between individual predictors were checked to control a possible occurrence of multicollinearity. Given the association between the students' engagement in classroom instruction and their momentary interest in the subject, which was confirmed by the models for Grade 4 (Tables 2 and 3), high intercorrelations between these two variables were expected. For reading, the correlation between engaging instruction and prior interest was .45, for mathematics it was .60. Nevertheless, the variance inflation factor and tolerance, which are commonly used to estimate the magnitude of multicollinearity (O'Brien, 2007), had acceptable values.

It is evident that classroom engagement in Grade 4 is not associated with student interest in Grade 6 in any of the domains. When tested without other controls, previous engagement had a significant effect .18 for reading and .24 for mathematics. When other variables are included in the model, the net effect of previous classroom instruction on student interest is no longer significant. Its impact is most likely mediated through previously aroused interest, which is a significant predictor of student interest in Grade 6 for both reading and mathematics. There are, however, notable differences between the two domains: whereas prior interest tends to be a dominant predictor of future interest in reading, the role of prior interest in mathematics is relatively less important while perceived difficulty is much closely connected with (low) interest in Grade 6.

Table 4 Linear regression models predicting student interest in Grade 6.

	Model 1 – Reading		Model 2 – Mathematics	
	Beta	*SE*	*Beta*	*SE*
Engaging instruction (Grade 4)	−.04	.02	.01	.02
Perceived difficulty (Grade 6)	−.14*	.02	−.68*	.02
Prior interest (Grade 4)	.36*	.03	.14*	.02
Family background	.06*	.02	−.04*	.02
Gender (boy)	−.13*	.03	.02	.02
Time spent on PC games (Grade 6)	−.13*	.03	−.07*	.04
Time spent on TV or video (Grade 6)	−.10*	.03	−.01	.02
N (listwise)	2727		2402	
R^2	.30		.56	

* $p < .05$

The decrease of the interest in reading can be explained by the fact that reading as a time-intensive and cognitively demanding activity has to compete with other less demanding and more alluring free time activities, whose effect tends to be stronger than in Grade 4. Moreover, the percentage of students who spend more than one hour with watching TV and playing computer games increased between Grades 4 and 6 from 55 to 64% and from 37 to 50%, respectively. Perceived difficulty also plays a role, but its effect is weaker than in Grade 4. Interestingly, the percentages of students indicating that reading is easy and that they usually do well in reading are almost identical in Grades 4 and 6 (approximately 50% of students declare strong agreement and around 35% little agreement with the two statements). However, 40% of Grade 6 students admit that they sometimes have troubles to exactly understand what they read.

Contrary to reading, increasing difficulty of mathematics can be regarded as the main reason why students lose their interest as they pass to higher grades. The effect of perceived difficulty is stronger than it was in Grade 4 and older children also tend to assess mathematics as more difficult. Although a direct comparison is not possible, the percentage of Grade 6 students who disagreed with the statement "I was always good at mathematics" (29%) was more than twice higher than the percentage of students who rejected a similar statement "I usually do well in mathematics" in Grade 4 (13%). In general, 38% of Grade 6 students regard mathematics as difficult rather than easy. Gender does not have a significant effect on student interest in mathematics when other predictors are accounted for. This is in line with the results for Grade 4. The small, but significant effect of time spent on PC games is difficult to interpret, but it can signalize a differential identity building during adolescence as outlined by Frenzel et al. (2010).

4 Conclusion and discussion

This study contributes to the discussion of interest development in educational settings through a secondary analysis of data from international large-scale assessments. To my best knowledge, this has not yet been done. In this study, I used PIRLS and TIMSS 2011 data to address three research questions. First, I estimated the role of classroom instruction in arousing student interest in reading and mathematics as compared to the role of student personal characteristics. Second, I examined the enduring effect of classroom instruction on student interest. Third, I analysed the commonalities and differences between reading and mathematics. Data from PIRLS and TIMSS 2011 appeared to be suitable for these purposes. In 2011, the same students were administered both reading and mathematics/science assessments, which facilitated the comparison of reading and mathematics. Further, a new scale of engaging instruction was introduced, which allowed to include a promising teacher-related variable in the analyses. And finally, student interest and its development could be studied on longitudinal data thanks to the CLoSE project that followed up the respective cohort of students.

As regards the first research question, engaging classroom instruction in Grade 4 is closely (with a net effect of around .40) associated with higher interest in both domains. Its estimated effect on student interest in reading is markedly higher than the effect of any of the student variables included in the models. In mathematics, however, the relative position of engaging instruction among a set of different predictors is not as dominant as in the domain of reading. Rather, it tends to be comparable to the position of perceived difficulty of mathematics, which has a similar effect but in the opposite direction. The effect of engaging instruction on momentary student interest in Grade 6 could not be tested with the available data.

With regard to the enduring effect of classroom instruction on student interest in higher grades, which was the subject of the second research question, it is clear that engaging instruction in Grade 4 does not have an independent effect on student interest in Grade 6 when the interest level in Grade 4 is accounted for. Rather, its effect is mediated through previously evoked interest in the subject, which was the most powerful predictor of future interest in reading and significantly related to future interest in mathematics.

Concerning the third research question, the results show that despite the general similarities related to the role of different factors in explaining the level of student interest in both domains, there are also some important distinctions. Most remarkably, perceived difficulty of the subject is a crucial predictor of (low) interest in mathematics with a strengthening effect from Grade 4 to Grade 6. Prior interest partly counter-balances the negative effect of perceived difficulty but only to a limited extent. Relatively high values of explained variance in the models for mathematics indicate that perceived difficulty is an essential variable that has to be considered when thinking about practical measures to raise student interest in mathematics. By contrast, the net effect of perceived difficulty of reading in Grade 4 was

comparatively weaker than in mathematics and further decreased in Grade 6. On the other hand, the effect of free time entertainments, such as watching TV and playing computer games, on the interest in reading increased between Grades 4 and 6. Time spend by these free time activities is practically negligible when it comes to the interest in mathematics.

The present study has confirmed a general decrease of interest in both domains as students grow older. On the other hand, the results do not show a dramatically low interest in mathematics. Although Czech students tend to have lower interest levels than their peers from other countries (Chvál, 2013; Mullis et al., 2012b), the majority of them still likes mathematics. Almost 80% of students liked mathematics in Grade 4, which was similar to the percentage of students who liked reading. Moreover, similar interest decreases by approximately 10% were observed in both domains between Grades 4 and 6. Based on the previous research (Chvál, 2013; Pavelková & Hrabal, 2012), a more substantial drop of interest in mathematics is to be expected in the next period.[7] A parallel steep decrease of interest in reading was registered only among boys (Ronková, 2015). The trajectories of interest development during Grade 6 and after were outside the scope of this study and remain to be analysed in the future, for example using the data from Grade 9 students collected within the CLoSE project.

An important contribution of this study consists in the inclusion of variables related to student engagement in classroom instruction. When the results from TIMSS 2007 were published, the decline of Czech students' mathematics achievement attracted wide attention of policy makers, experts on education, teachers and the general public. Low student interest in mathematics as compared to other countries has also been discussed (Chvál, 2013) and related to student, teacher and school characteristics (Federičová & Münich, 2015). However, teacher variables that were selected as possible predictors (gender, age and length of teaching experience) explained only a low proportion of variance in interest levels. The present study, by contrast, suggests that certain classroom activities, such as an easy-to-understand instruction, a clear task formulation, working on interesting tasks and with attractive materials, can effectively arouse student interest. However, it has to be emphasised that cross-sectional data do not allow to draw causal conclusions. Another possible interpretation of the association between the two variables could be that students who are *a priori* more interested also feel more engaged during lessons. Most probably, both processes occur, reinforcing each other. The longitudinal extension of the dataset does not allow to decide which one is predominant, as Grade 6 students were not asked about their momentary classroom engagement and the effect of prior interest on future engagement could not be tested. The exploration of student engagement in classroom instruction in higher grades including best ways of its measurement are open for future research.

[7] Data collection in Grade 6 was at the beginning of the school year.

The relationship between student interest in Grades 4 and 6 is generally consistent with the theory of ontogenetic interest development from situational to personal interest (Hidi & Renninger, 2006; Krapp, 2002, 2007). This study could not prove the validity of this theory, but it drew attention to the fact that the development and stabilisation of student interest evoked by a favourable classroom instruction might be by inhibited by other factors, most importantly by seductive free time activities that divert the children from reading and by perceived difficulty that counteracts the effect of prior interest in mathematics. As noted before, the role of continuously engaging classroom instruction in this development could not be examined due to the lack of appropriate data and deserves further investigation. A proper understanding of the complex process of interest development will most likely need not only better measures, but also more sophisticated analytical methods, such as structural equation modelling. A further limitation of this study is a problematic scale of perceived difficulty of reading in Grade 6, which has a low reliability. Possibilities of improving the scale should be investigated in future studies.

Several implications for educational practice could be drawn from this study. As for reading, targeted use of engaging instructional methods in primary education and offering interesting reading materials that would motivate children to limit their time spent on computer and TV in favour of reading in lower secondary education could lead to interest increase. An unresolved question is how to motivate boys who demonstrate a significantly lower interest in reading when all other variables are controlled. It is important to note that boys in Grade 6 show lower interest in reading even when controlled for prior level of interest, which means that they lose their interest more easily than girls. One possible option could be to offer a wider selection of reading materials including non-fiction texts dealing with topics that could attract boys' attention.

In mathematics, the use of engaging instructional methods seems to be less important than targeted efforts to convince students that mathematics is not as difficult as they may perceive it. It would be very useful to find out which classroom practices can potentially reduce the fear from mathematics and to share them as examples of best practice. It needs to be recognized, however, that this study did not analyse other factors that might be responsible for the decline of student interest in mathematics. For example, lower secondary school students might develop a deeper interest in another subject (physics, biology, history, foreign language …), which leads to changes in their relative interest in mathematics compared to other domains.

This study has also broader implications for educational policy and research. It showed that secondary analyses of data from international large-scale assessments can be used to gain a better insight into questions related to the development of student interest in core school subjects. Although it is not always easy to connect variables from international studies to specific characteristics of national educational systems and their particular problems, student attitudes are obviously one of the research fields that can benefit from a more extensive use of large-scale

assessments data. It would be more than welcome if local researchers proposed and national educational authorities and granting agencies supported more projects that relate findings from international large-scale assessments to issues relevant for the local context.

Although the Czech School Inspectorate, which is responsible for the implementation of international large-scale assessments in the Czech Republic, has recently made significant progress in building bridges between international large-scale assessments and local educational research, as exemplified for instance by a series of publications following TALIS 2013, there is still a lot of gaps to be filled in. Especially, advanced secondary analyses of the data collected in international studies are still exceptions performed by a few researchers. There are at least three reasons why international assessments can serve as valuable source of data even for one-country studies. First, the test and questionnaire items are grounded in solid assessment frameworks that incorporate latest theoretical and empirical production. Second, the wording of the test and questionnaire items are thoroughly piloted before real administration. Third, the data are collected on representative samples under standardized conditions and carefully cleaned. Very few national studies yield quantitative data of such quality.

Secondary data analyses can contribute to an effective exploitation of resources invested in international large-scale assessments, and they might be worthwhile in at least two other ways: they can not only help focus further research on important questions that cannot be answered solely by international large-scale assessments, but they can also help formulate proposals for national adaptations and amendments of international instruments so that they better reflect specific issues that need to be investigated. Hopefully, more researchers will use the opportunities to publish their analyses in the future.

Acknowledgements

The study was supported by a grant by the Czech Science Foundation through the project “The relationships between skills, schooling and labor market outcomes: a longitudinal study” (P402/12/G130), and by the Charles University, projects GA UK No 638218 and PRIMUS/17/HUM/11 “Center for Educational Measurement and Psychometrics (CEMP)”.

The author would like to thank Patrícia Martinková and two anonymous reviewers for their helpful comments on previous versions of this text. She would also like to thank Charles University for financial support.

References

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioural change. *Psychological Review, 84*(2), 191–215.

Chvál, M. (2013). Změna postojů českých žáků k matematice během školní docházky [Change of attitudes of Czech pupils towards mathematics during school attendance]. *Orbis Scholae, 7*(3), 49–71.

Cortright, R. N., Lujan, H. L., & Blumberg, A. J. (2013). Higher levels of intrinsic motivation are related to higher levels of class performance for male but not female students. *Advances in Physiology Education, 37*(3), 227–232.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York: Plenum.

Eccles (Parsons), J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation: Psychological and sociological approaches* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Federičová, M., & Münich, D. (2015). Srovnání žákovské obliby školy a matematiky pohledem mezinárodních šetření [A comparison of satisfaction with school and mathematics from the perspective of international testing programs]. *Pedagogická orientace, 25*(4), 557–582.

Fredricks, J. A., & Eccles, J. S. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology, 38*(4), 519–533.

Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. G. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence, 20*(2), 507–537

Gläser-Zikuda, M., Stuchlíková, I., & Janík, T. (2013). Emotional aspects of learning and teaching: Reviewing the field – discussing the issues. *Orbis scholae, 7*(2), 7–22.

Gottfried, A. E., Marcoulides, G. A., Gottfried, A. W. & Oliver, P. H. (2013). Longitudinal pathways from math intrinsic motivation and achievement to math course accomplishments and educational attainment. *Journal of Research on Educational Effectiveness, 6*(1), 68–92.

Gustafsson, J.-E., Yang Hansen, K., & Rosén, M. (2013). Effects of home background on student achievement in reading, mathematics and science at the fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade – implications for early learning.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*(2), 111–127.

Kearney, C. (2016). *Efforts to Increase Students' Interest in Pursuing Mathematics, Science and Technology Studies and Careers. National Measures taken by 30 Countries – 2015 Report*. Brussels: European Schoolnet.

Köller, O., Baumert, J., & Schnabel, K. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education, 32*(5), 448–470.

Korhonen, J., Tapola, A., Linnanmäki, K., & Aunio, P. (2016). Gendered pathways to educational aspirations: The role of academic self-concept, school burnout, achievement and interest in mathematics and reading. *Learning and Instruction, 46*, 21–33.

Krapp, A. (2002). Structural and dynamic aspects of interest development: Theoretical considerations from an ontogenetic perspective. *Learning and Instruction, 12*(4), 383–409.

Krapp, A. (2007). An educational-psychological conceptualization of interest. *International Journal for Educational and Vocational Guidance, 7*(1), 5–21.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McLaughlin, M., McGrath, D. J., Burian-Fitzgerald, M. A., Lanahan, L., Scotchmer, M., Enyeart, C., & Salganik, L. (2005). *Student content engagement as a construct for the measurement*

of effective classroom instruction and teacher knowledge. Washington, D.C.: American Institutes for Research.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. T, & Sainsbury, M. (2009a). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009b). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012a). *PIRLS international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012b). *TIMSS international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework* (2nd ed.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nagy, G., Trautwein, U., Baumert, J., Köller, O., & Garrett, J. (2006). Gender and course selection in upper secondary education: Effects of academic self-concept and intrinsic value. *Educational Research and Evaluation, 12*(4), 323–345.

O'Brien, R. M (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673–690.

OECD (2010). *PISA 2009 results: Learning to learn – student engagement, strategies and practices (Volume III)*. Paris: OECD.

Pajares, F., & Miller, D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology, 86*(2), 193–203.

Pavelková, I., & Hrabal, V. (2012). Mathematics in perception of pupils and teachers. *Orbis scholae, 6*(2), 119–132.

Renninger, K. A., & Hidi, S. (2011). Revisiting the conceptualization, measurement, and generation of interest. *Educational Psychologist, 46*(3), 168–184.

Ronková, J. (2015). *Rozvoj* čtenářské *gramotnosti: Edukační model na bázi metody Podvojného zápisu* [Disertační práce] [Development of reading literacy: Educational model based on the double-entry diary method (Doctoral thesis)]. Praha: Univerzita Karlova, Pedagogická fakulta.

Rotgans, J. I., & Schmidt, H. G. (2017). Interest development: Arousing situational interest affects growth trajectory of individual interest. *Contemporary Educational Psychology, 49*, 175–184.

Savelsbergh, E. R., Prins, G. T., Rietbergen, C., Fechner, S., Vaessen, B. E., Draijer, J. M., & Bakker, A. (2016). Effects on innovative science and mathematics teaching on student attitudes and achievement: A meta-analytic study. *Educational Research Review, 19*, 158–172.

Springer, S. E., Harris, S., & Dole, J. A. (2017). From surviving to thriving: Four research-based principles to build students' reading interest. *The Reading Teacher, 71*(1), 43–50.

Straková, J. (2016). *Mezinárodní výzkumy výsledků vzdělávání. Metodologie, přínosy, rizika a příležitosti* [International large-scale assessment surveys: Methodology, benefits, risks, and opportunities]. Praha: Univerzita Karlova, Pedagogická fakulta.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68–81.

Mgr. Eva Potužníková
Institute for Research and Development of Education
Faculty of Education, Charles University
Myslíkova 7, 110 00 Prague 1, Czechia
eva.potuznikova@pedf.cuni.cz

Appendix – Items used to construct the scales and their Czech equivalents

Students interest in reading (ASBGSLR) – Grade 4

The scale was formed of six items scored on a 4-point Likert agreement scale ranging from "disagree a lot" to "agree a lot" and two additional items scored on a 4-point frequency scale ranging from "never or almost never" to "every day or almost every day". Items indicating negative statements about reading were recoded so that higher values represent higher interest.

English source (PIRLS)	Czech wording in PIRLS questionnaire
I read only if I have to (reverse coded)	Čtu, jen když musím
I like talking about what I read with other people	Rád/a si s ostatními lidmi povídám o tom, co čtu
I would be happy if someone gave me a book as a present	Měl/a bych radost, kdyby mi někdo dal knihu jako dárek
I thing reading is boring (reverse coded)	Myslím si, že čtení je nuda
I would like to have more time for reading	Chtěl/a bych mít na čtení více času
I enjoy reading	Čtení mě baví
I read for fun	Čtu si pro radost
I read things that I choose myself	Čtu to, co si sám/sama vyberu

Students like learning mathematics (ASBGSLM)– Grade 4

The scale was formed of five items scored on a 4-point Likert agreement scale ranging from "disagree a lot" to "agree a lot".

English source (TIMSS)	Czech wording in TIMSS questionnaire
I enjoy learning mathematics	Baví mě učit se matematiku
I wish I did not have to study mathematics (reverse coded)	Nejraději bych se matematiku neučil/a
Mathematics is boring (reverse coded)	Matematika je nudná
I learn many interesting things in mathematics	V matematice se naučím mnoho zajímavého
I like mathematics	Matematiku mám rád/a

Students engaged in reading lessons (ASBGERL)– Grade 4

The scale was formed of seven items scored on a 4-point Likert agreement scale ranging from "disagree a lot" to "agree a lot". The introductory part of the question directed the students to think about reading in school.

English source (PIRLS)	Czech wording in PIRLS questionnaire
I like what I read about in school	Líbí se mi, o čem ve škole čteme
My teacher gives me interesting things to read	Učitel mi dává číst zajímavé věci
I know what my teacher expects me to do	Vím, co učitel chce, abych dělal/a
I think of things not related to the lesson (reverse coded)	Při čtení myslím na něco jiného
My teacher is easy to understand	Učitel vysvětluje srozumitelně
I am interested in what my teacher says	Zajímá mě, co učitel říká
My teacher gives me interesting things to do	Učitel mi dává zajímavé úkoly

Students engaged in mathematics lessons (ASBGEML) – Grade 4

The scale was formed of five items scored on a 4-point Likert agreement scale ranging from "disagree a lot" to "agree a lot". The introductory part of the question explained that the statements relate to mathematics lessons.

English source (TIMSS)	Czech wording in TIMSS questionnaire
I know what my teacher expects me to do	Vím, co učitel chce, abych dělal/a
I think of things not related to the lesson (reverse coded)	Při matematice myslím na něco jiného
My teacher is easy to understand	Učitel vysvětluje srozumitelně
I am interested in what my teacher says	Zajímá mě, co učitel říká
My teacher gives me interesting things to do	Učitel mi dává zajímavé úkoly

Perceived difficulty of reading – Grade 4

The scale was formed of four items scored on a 4-point Likert agreement scale ranging from "disagree a lot" to "agree a lot".

English source (PIRLS)	Czech wording in PIRLS questionnaire
I usually do well in reading (reverse coded)	Čtení mi většinou jde
Reading is easy for me (reverse coded)	Čtení je pro mě snadné
Reading is harder for me than for many of my classmates	Čtení je pro mě těžší než pro spoustu mých spolužáků
Reading is harder for me than any other subject	Čtení je pro mě těžší než ostatní předměty

Perceived difficulty of mathematics – Grade 4

The scale was formed of four items scored on a 4-point Likert agreement scale ranging from "disagree a lot" to "agree a lot".

English source (TIMSS)	Czech wording in TIMSS questionnaire
I usually do well in mathematics (reverse coded)	Matematika mi většinou jde
I am just not good at mathematics	Matematika mi moc nejde
Mathematics is harder for me than for many of my classmates	Matematika je pro mě těžší než pro spoustu mých spolužáků
Mathematics is harder for me than any other subject	Matematika je pro mě těžší než ostatní předměty

Student interest in reading – Grade 6

The scale was formed of three items scored on a 4-point Likert agreement scale ranging from "disagree a lot" to "agree a lot".

English translation	Czech wording in CLoSE questionnaire
I enjoy reading	Čtení mě baví
I would like to have more time for reading	Chtěl/a bych mít na čtení více času
I thing reading is boring (reverse coded)	Myslím si, že čtení je nuda

Student interest in mathematics – Grade 6

The scale was formed of one item assessing the popularity of mathematics on a 5-point scale and the following three items scored on a 4-point Likert agreement scale ranging from "disagree a lot" to "agree a lot".

English translation	Czech wording in CLoSE questionnaire
I sometimes get so engaged in solving mathematics problems that I don't notice the world around me	Někdy se tak zaberu do řešení matematických úloh, že nevnímám svět kolem sebe
I don't want to give over mathematics because I enjoy it	Nechtěl/a bych nechat matematiky, protože mě matematika baví
Mathematics is one of my favourite subjects	Matematika je pro mě jedním z nejlepších předmětů

Perceived difficulty of reading – Grade 6

The scale was formed of five items scored on a 4-point Likert agreement scale ranging from “disagree a lot” to “agree a lot”.

English translation	Czech wording in CLoSE questionnaire
I usually do well in reading (reverse coded)	Čtení mi většinou jde
Reading is easy for me (reverse coded)	Čtení je pro mě snadné
I sometimes have troubles to exactly understand what I read	Někdy mám problem přesně porozumět tomu, co čtu
I have to read the text more than once to understand it properly	Musím si text přečíst vícekrát, abych mu pořádně porozuměl/a
I understand well and easily what the text says (reverse coded)	Dobře a snadno rozumím tomu, co se v textu říká

Perceived difficulty of mathematics – Grade 6

The scale was formed of one item assessing the difficulty of mathematics on a 2-point scale (difficult vs. easy) and the following two items scored on a 4-point Likert agreement scale ranging from “disagree a lot” to “agree a lot”.

English translation	Czech wording in CLoSE questionnaire
I was always good at mathematics	Matematika mi vždycky šla
I have good marks in mathematics	Mám dobré známky z matematiky

Demonstration of Simpson's Paradox in PISA 2015 Data: Confusing Differences between Boys and Girls

Gašper Cankar

National Examinations Centre, Slovenia

Abstract: This paper explores the occurrence of a Simpson's paradox in PISA 2015 science literacy data. Simpson's paradox, a case of contradicting interpretations when results are analysed by groups or aggregated as a whole, has both a practical and an academic significance. It is an interesting phenomenon that is far from theoretical and when it happens, it has profound effects on the interpretation and if left unidentified can cause confusion and misunderstanding. This paper demonstrates best ways to detect Simpson's paradox through appropriate tables and graphs. Actual occurrences of a Simpson's paradox and conditions leading to them are explored using PISA 2015 gender differences in science literacy data in five central European countries – Austria, Croatia, Czech Republic, Slovakia and Slovenia. In countries where the occurrence of a Simpson's paradox was detected, we provide correct interpretation of the results. Beside creating problems with interpretation an occurrence of a Simpson's paradox also provides new insight – it signifies that there is very different gender composition in different educational tracks which has important implications for the educational governance. We will discuss implications of these findings in context of Slovenian educational system.

Keywords: PISA; Simpson's paradox; gender differences; educational tracks; governance

Statistical paradoxes are usually not important when interpreting data from the international large scale assessments (ILSA). As Gardner (1982) points out, they are an interesting topic in itself, but they are more commonly viewed as a hobby of a retired statistician, a relaxing pursuit of students of statistics or as brainteasers intended to rouse curiosity and interest in the mathematics. Sometimes, however they also have profound implications on interpretation of a real data. In this paper we will focus on a Simpson's paradox, it's real life occurrences and implications for use and interpretation of data. As it turns out, the knowledge about a Simpson's paradox can be useful when interpreting results from the large scale assessments.

A Simpson's paradox is a situation where we get conflicting interpretations when same results are analysed at different levels of grouping. Or as Lesser (2001) puts it: "Simpson's paradox can be concisely defined as the reversal of a comparison when data are grouped." It was named a Simpson's paradox by Blythe (1972) after Edward Simpson, a British statistician who first wrote about it when he was still a post-graduate student (Simpson, 1951). Blythe neglected that another British statistician Udny Yule wrote about same paradox already in 1903 (Yule, 1903). To acknowledge this some authors nowadays also call it Yule-Simpson effect (Demers & Rossmo, 2015). We will use a shorter name throughout the paper.

https://doi.org/10.14712/23363177.2018.296
www.orbisscholae.cz

The paradox can be best explained through an example. Imagine two classes of students (Class A & Class B) learning same course on Mathematics and taking same test at the end. Both classes would consist of 30 students and Table 1 presents their average points achieved on test reported by gender.

Table 1 Results on Mathematics achievement test for Class A and B.

	Average (boys)	Average (girls)	Difference (girls−boys)
Class A	23.6	20.3	−3.3
Class B	13.7	10.4	−3.3
Total	16.0	18.0	+2.0

If we would compare boys and girls in Class A alone, we would conclude from difference that the boys on average perform better. Same conclusion would follow from the difference in Class B (3.3 points in favour of boys). But when we combine data from both classes, girls outperform boys for 2 points! This is called a Simpson's paradox and it is not an error in calculations. The reason for the observed phenomena is in the distribution of boys and girls in both classes as seen in Table 2.

Table 2 Number of boys and girls in Classes A and B.

	Number (boys)	**Number (girls)**
Class A	7	23
Class B	23	7
Total	30	30

From Table 1 it was obvious that students in Class A on average performed much better then students in Class B. Therefore, the grouping of students into classes with regard to their Mathematics achievement was not random. The unequal proportions of boys and girls (7:23) combined with non-random grouping resulted in an observed paradox. In other words: in Class A the small number of high performing boys outperformed more numerous female peers. In Class B larger number of boys again outperformed smaller number of girls. Only when we join classes we discover the actual difference where girls on average performed better on the Mathematics test then boys. If we make conclusions only on averages from each class, we miss the real picture.

The example above is artificially constructed to explain the paradox. What about in real life? Is the paradox in practice really common or is it a rare finding that occurs only seldom? Judging from the amount of research literature the occurrence is certainly not uncommon. If we focus only on the recent research literature it can be found in different areas of science and life in general: medicine (Baker & Kramer, 2001; Rücker & Schumacher, 2008), administration (Demers & Rossmo, 2015) and even sports (Wright, 2012). In this paper we will explore its occurrence in large scale assessments in education.

1 State-of-the-art

Before we start with an analysis we will explore different ways to represent a Simpson's paradox as such methods can help researchers to detect it and act accordingly.

To detect the Simpson's paradox we can always calculate differences of averages in all subgroups and in a sample as a whole and see if it occurs as we did in Table 1 of our example. This however misses the point that there are many situations when we don't get an actual Simpson's paradox (reversal of difference between averages) but we get a substantial increase or decrease in the difference. Checking actual tables of averages may be a robust and concise way but it might be less visually appealing as a lot of tables makes results hard to read.

The best methods to spot a Simpson's paradox in practice are graphical. This is due to the fact that a proper graphical representation accounts for different proportions of students in subgroups and difference in averages at the same time.

We will explore three ways to represent data: Bar-plot representation, Square representation, and Trapezoidal representation.

1.1 Bar-plot representation

This is a simple example trying to demonstrate on the same picture proportions of students and their average scores. Figure 1 shows a bar-plot of Class A and B students from our example.

Bar plot in Figure 1 fairly well shows differences in proportions but not differences in averages. It is simple to construct but it doesn't warn us about a Simpson's paradox on the first glance as there is no difference calculated. The reader must infer the inversion from comparison of averages as it is not readily visible.

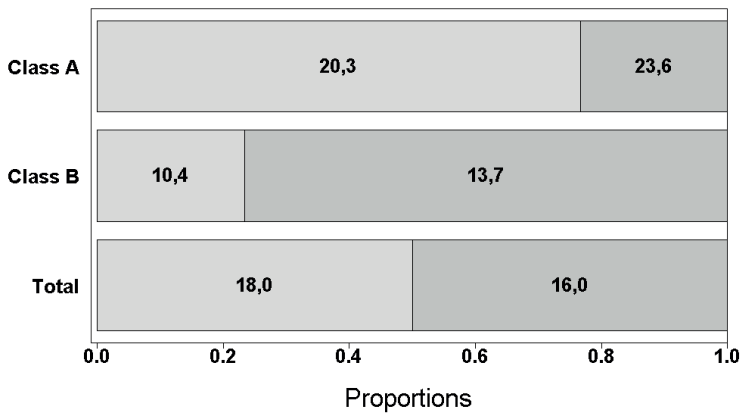
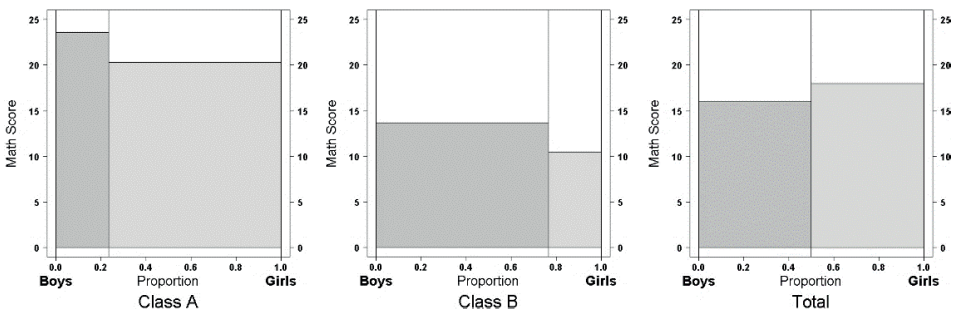


Figure 1 Bar plot of proportions of girls (light) and boys (dark) in classes A and B with averages printed inside bars.

1.2 Square representation

This representation tries to capture differences in proportions and differences in averages in the same figure. It is adapted from unit square representation described by Lesser (2001). For each comparison (Class A, Class B and Total) we construct a square where one dimension represents proportions and the other dimension represents average scores. From series of three figures (for Class A, Class B and Total) we can observe what happened to average scores in subgroups and in total. When drawing the figure we first divide the square according to the proportions (in our example of boys and girls). Then we draw averages for each gender and shade each area respectively. Figures 2 to 4 show graphs for our example.

Square representation allows us to compare graphs for subgroups with the last graph that shows all subgroups together. The inversion of difference in the last graph (Figure 4) is now evident and it's easier to understand what happened. The downside is that you can't represent all information in just one graph but you have to compare several figures simultaneously.



Figures 2–4 Square representations of proportions and average scores for Boys and Girls in classes A, B and in Total respectively.

1.3 Trapezoidal representation

Trapezoidal representation of a Simpson's paradox was first proposed by Tan (1986) who observed that "the length of any line segment which is parallel to the two bases and has its endpoints on the nonparallel sides of a trapezoid is the weighted mean of the lengths of the two bases". What this actually means is that we can plot all information on the same graph following this procedure:

- We start with square plot where x axis represents Proportions, left y axis represents Class A math score and right y axis represents Class B math score.
- On left y axis we mark Class A average score for boys. On right y axis we mark Class B average score for boys.
- We draw the line segment connecting both points (Class A and B boys' average score).
- On the x axis we mark the proportions of boys in Class A and Class B (from all the boys in Total).

– The vertical line delineating those two proportions actually intersects the line connecting both average scores right at the point of total average score for boys. Example is shown in Figure 5.

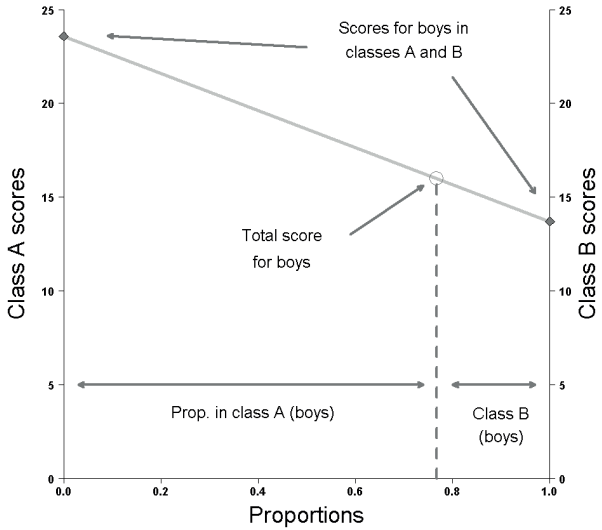


Figure 5 Example for construction of trapezoidal representation for boys.

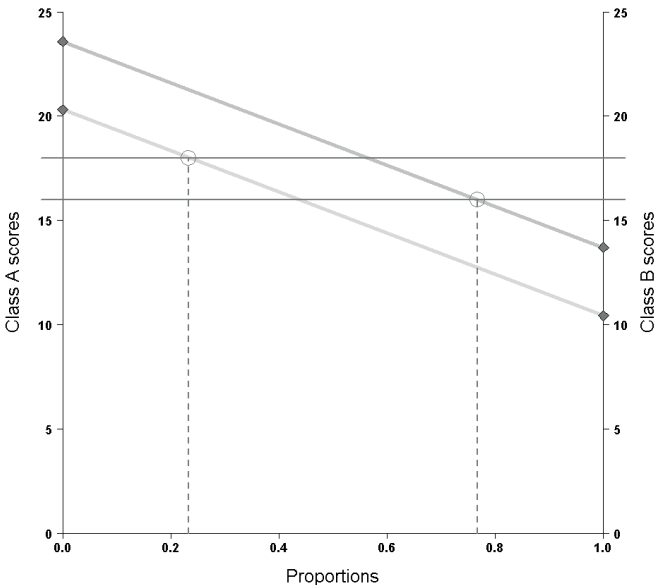


Figure 6 Trapezoidal representation of our example of classes A and B. Left circle represents girls' average, right circle boys' average.

If we repeat same procedure for girls we can draw on the same graph another set of lines for girls. Then we can compare on the same graph differences in lines connecting averages and differences in heights at intersections (where the averages of all the boys and all the girls can be found). Our example of a Simpson's paradox can be seen in Figure 6.

Figure 6 shows very clearly that girls have lower average in both classes A and B. At the same time we also see that the average from both classes together is higher for girls than for boys. With trapezoidal representation we can show a Simpson's paradox in only one graph. There is a downside though – the method is suitable only when we have two subgroups like Class A and B in our example. If we would have three classes, the trapezoidal representation couldn't be applied.

We presented three graphical ways to explore the relationship between differences in subpopulations and in the total population and we mentioned their strengths and shortcomings. Trapezoidal representation seems most prudent as it clearly shows all information in just one graph, but it will be unusable for our purpose in this paper since we will be exploring occurrence of a Simpson's paradox between boys and girls in educational tracks. Most countries have their 15-year-old students in more than two educational tracks of formal education which suggests we should use graphical method that can accommodate more than two groups. One option would be to proceed with a Square representation but educational tracks present quite a challenge since they are a) numerous, which means a lot of graphs for each country; and b) not equal in size. Some educational tracks cover large portions of population of 15-year-olds other educational tracks include only small subgroups. Making them visually equal might again skew the interpretation.

To address this issue we will modify the Square representation by joining all educational tracks in the same graph and defining their widths according to the size of population in each track. Overall averages can be drawn as horizontal lines across whole graph. Examples are shown in the results section below.

1.4 Hypothesis

To focus our research, we state following two null hypotheses about differences between boys and girls in total and in subpopulations of each educational track (for each country):

H01: Differences between boys and girls in PISA science results within educational tracks are equal to overall difference between boys and girls in each country.

We also state stricter hypothesis that explicitly involves a Simpson's paradox (for each country):

H02: Differences between boys and girls in PISA science results within educational tracks and in total don't show the pattern of Simpson's paradox (reversed difference) in each country.

2 Method

This research draws data from the Programme for International Student Assessment (PISA) from 2015 cycle. Participants are students who were at the time of PISA main study 15 years old and still in formal education. To limit our exploration, we selected data from following countries: Austria, Croatia, Czech Republic, Slovakia and Slovenia. On this data we performed secondary data analysis to find out proportions of boys and girls in each educational track and their score on Science literacy.

As Smith (2008) points out secondary data analysis can be full of errors if it's not done correctly. In case of ILSA we therefore consulted Technical report (OECD, 2017) where appropriate. All secondary data analyses were made using software IDB Analyzer 4.0.21 (IEA, 2018), using all 10 plausible values for Science literacy (PVSCIE) and the Final trimmed nonresponse adjusted student weight (W_FSTUWT). Plausible values are student's results (in our case for Science literacy) prepared in such a way that researchers can calculate standard errors of any statistical parameter they estimate from them. This is very important since it helps us to interpret the data better and puts findings into a perspective. Student weights (W_FSTUWT) are ponders that reflect sampling procedure and enable us to calculate representative estimates for a whole population of 15-year-olds in a country even if only a sample participated in a study.

Proportions by an educational track and gender and PVSCIE averages as well as standard errors (for significance testing) were calculated using the module 'Percentages and means'. Missing values were excluded from analyses by default. Educational tracks were captured in a PISA variable PROGN and names of educational tracks for each country are taken from that variable. Graphical representations were made using a statistical environment R (R Core Team, 2017).

3 Results and interpretations

For each country's results we will present PISA 2015 science results (PVSCIE) grouped by gender and educational tracks as noted in a variable PROGN. Students that participate in PISA can be in very different educational tracks; some are still in a comprehensive basic education, others already started in educational programmes leading to different secondary education outcomes. Educational tracks also differ widely in frequency – some are very popular and include large proportions of a whole population, others include only handful of students. Tables for each country are therefore not directly comparable. Educational tracks within the tables are ordered ascending according to average science score for each track.

To better understand proportions by gender and educational track each table also includes percentages of girls and boys and sums of student weights – they denote the size of a population captured in each statistic. Last column in each table presents a difference in science score between girls and boys in each educational track and

in total (the last line). Positive difference means girls have higher average PISA 2015 science score than boys.

Table 3 PISA 2015 science results by gender and educational track for Austria.

	National Study Programme	N_{GIRLS} (W_FSTUWT)	N_{BOYS} (W_FSTUWT)	% (GIRLS)	% (BOYS)	PVSCIE (GIRLS)	PVSCIE (BOYS)	Difference (GIRLS-BOYS)
Pr.1	Compulsory school	1925	2553	42.99	57.01	366.49	395.01	−28.52**
Pr.2	Voc. sch. for apprentices	4268	8782	32.71	67.29	417.38	442.13	−24.75**
Pr.3	Intermed. tech. and voc. schools	6048	5224	53.66	46.34	428.29	451.63	−23.34**
Pr.4	Higher tech. and voc. college	13011	11980	52.06	47.94	501.80	547.84	−46.04**
Pr.5	Academic secondary school	11091	8497	56.62	43.38	544.53	572.68	−28.15**
	Total	36345	37034	49.53	50.47	485.53	504.37	−18.84**

** Differences are statistically significant at $p < 0.05$.

PISA 2015 science results for Austria in Table 3 on first glance present uniform picture – boys outperformed girls within every educational track and also on a country's level. We can note, however that overall difference is smaller than any difference within educational tracks. A Simpson's paradox didn't happen, but the data on a whole and grouped by educational tracks suggest slightly different conclusions. While differences within educational tracks suggest that boys outperform girls for more than 23 points and in case of most numerous educational programme for more than 46 points the total difference is actually only 18.84 points.

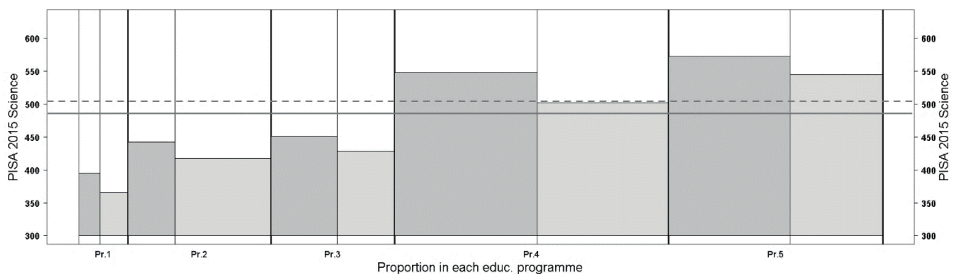


Figure 7 PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Austria. Width of each programme corresponds to proportion of programme in a whole population. Lines show total average (dashed – boys, solid – girls).

Figure 7 shows the same trend of bigger differences in each educational track and smaller overall difference for Austrian data. A Simpson's paradox didn't occur but conclusions about the size of difference when examining data per country and within educational tracks are different.

Table 4 PISA 2015 science results by gender and educational track for Croatia.

	National Study Programme	N_{GIRLS} (W_FSTUWT)	N_{BOYS} (W_FSTUWT)	% (GIRLS)	% (BOYS)	PVSCIE (GIRLS)	PVSCIE (BOYS)	Difference (GIRLS-BOYS)
Pr.1	Primary school – lower sed.+	54	34	61.89	38.11	339.88	402.30	−62.42**
Pr.2	Lower qualification voc. prog.	40	37	51.58	48.42	340.00	344.23	−4.23
Pr.3	Vocational prog. for crafts	2492	4091	37.85	62.15	381.54	399.14	−17.60**
Pr.4	Vocational prog. for industry	654	1638	28.52	71.48	382.80	403.85	−21.05**
Pr.5	Art programmes	285	51	84.88	15.12	451.04	489.66	−38.62
Pr.6	Four year vocational prog.	9214	9039	50.48	49.52	454.76	483.49	−28.73**
Pr.7	Gymnasium	8487	4783	63.96	36.04	527.78	563.63	−35.85**
	Total	21226	19673	51.90	48.10	472.59	478.42	−5.83

sed+ – secondary education; ** differences are statistically significant at $p < 0.05$.

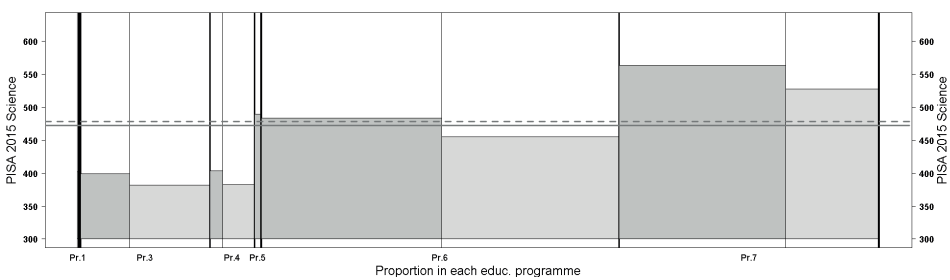


Figure 8 PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Croatia. Width of each programme corresponds to proportion of whole population. Lines show total average (dashed – boys, solid – girls).

PISA 2015 science results for Croatia in Table 4 show similar trend than in Austria. Although boys outperform girls on whole and within every educational track we can still note that overall difference is rather low (5.83) compared to differences in most numerous educational tracks where boys outperform girls on average between 17 and 35 points! This is also evident in statistical significance results – overall difference is within the margins of ±1.96 standard errors while differences in most educational tracks are much bigger and statistically significant.

Figures of differences for educational tracks in Croatia give similar conclusion as Table 4 – reversal of differences didn't occur but it is much smaller on a whole compared to major educational tracks within the country.

Table 5 PISA 2015 science results by gender and educational track for Czech Republic.

	National Study Programme	N_{GIRLS} (W_FSTUWT)	N_{BOYS} (W_FSTUWT)	% (GIRLS)	% (BOYS)	PVSCIE (GIRLS)	PVSCIE (BOYS)	Difference (GIRLS-BOYS)
Pr.1	Basic special schools	680	813	45.55	54.45	361.18	348.96	12.22
Pr.2	Secondary special schools	226	248	47.62	52.38	403.92	405.95	−2.03
Pr.3	Voc\tech sed+ without maturate	2850	4618	38.17	61.83	400.53	420.63	−20.10**
Pr.4	Basic school	17140	21852	43.96	56.04	464.64	471.25	−6.61
Pr.5	Voc\tech sed+ with maturate	11532	8636	57.18	42.82	486.79	525.64	−38.85**
Pr.6	4-year gymnasium	4031	2157	65.15	34.85	567.80	595.70	−27.90**
Pr.7	6, 8-year gymnasium and 8-year conservatory (lower secondary)	2268	2717	45.49	54.51	581.31	605.98	−24.67**
Pr.8	6, 8-year gymnasium (upper secondary)	2400	2351	50.51	49.49	593.02	626.00	−32.98**
	Total	4112	43392	48.66	51.34	488.40	497.03	−8.63**

sed+ – secondary education; ** Differences are statistically significant at $p < 0.05$.

In Table 5 we present PISA 2015 science results by gender and educational track for the Czech Republic. Gender difference on country level (8.63) are similar to

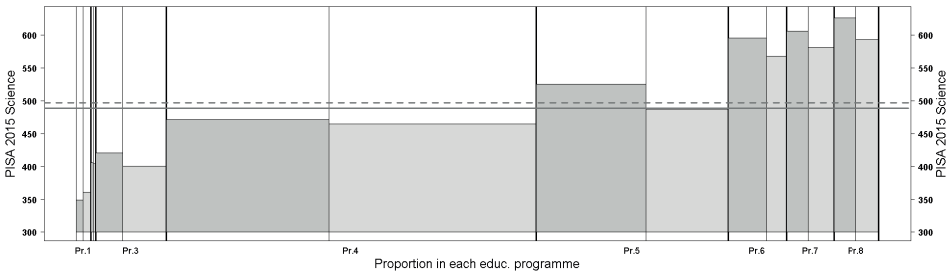


Figure 9 PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Czech Republic. Width of each programme corresponds to proportion of whole population. Lines show total average (dashed – boys, solid – girls).

difference between students still in Basic schools. This makes sense since those students are still in comprehensive part of educational system. Differences increase drastically in secondary education where students choose educational track according to their abilities and preferences.

Figure 9 and Table 5 show that Simpson's paradox didn't occur in case of PISA 2015 data for Czech Republic but they also show that secondary education tracks show much larger differences than Basic schools and all tracks together.

Table 6 PISA 2015 science results by gender and educational track for Slovakia.

	National Study Programme	N_{GIRLS} (W_FSTUWT)	N_{BOYS} (W_FSTUWT)	% (GIRLS)	% (BOYS)	PVSCIE (GIRLS)	PVSCIE (BOYS)	Difference (GIRLS-BOYS)
Pr.1	Vocational basic school	580	683	45.94	54.06	306.16	306.92	−0.76
Pr.2	Secondary vollege – without SLE	1014	1807	35.94	64.06	355.58	377.77	−22.19**
Pr.3	Basic school	9655	11518	45.60	54.40	431.51	440.72	−9.21**
Pr.4	Secondary college – with SLE	6122	7237	45.83	54.17	453.48	466.97	−13.49**
Pr.5	High school	5415	3293	62.19	37.81	538.46	559.27	−20.81**
Pr.6	Secondary school (ISCED2)	603	494	54.96	45.04	540.15	558.69	−18.54
Pr.7	Secondary school (ISCED3)	682	549	55.40	44.60	557.04	566.11	−9.07
	Total	24072	25582	48.48	51.52	461.22	460.36	0.86

SLE – school leaving examination; ** differences are statistically significant at $p < 0.05$.

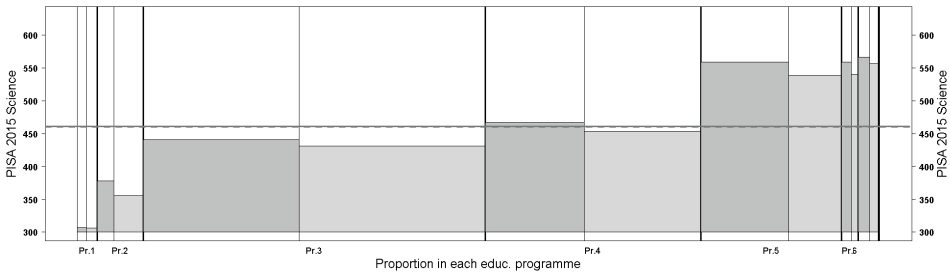


Figure 10 PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Slovakia. Width of each programme corresponds to proportion of a whole population. Lines show total average (dashed – boys, solid – girls).

Pisa 2015 results for Slovakia in Table 6 are an example of a Simpson's paradox in real life data. While all educational tracks suggest that boys outperform girls, on whole results suggest otherwise.

Graphically Figure 10 clearly shows that great differences in each educational track (most of them are statistically significant at the *p*-value 0.05 and less) don't translate to overall difference. Here results between boys and girls are practically identical as they are well within margins of standard error (SE_{GIRLS} = 3.31; SE_{BOYS} = 2.98).

Table 7 PISA 2015 science results by gender and educational track for Slovenia.

	National Study Programme	N_{GIRLS} (W_FSTUWT)	N_{BOYS} (W_FSTUWT)	% (GIRLS)	% (BOYS)	PVSCIE (GIRLS)	PVSCIE (BOYS)	Difference (GIRLS-BOYS)
Pr.1	Voc. ed. short duration	42	121	25.58	74.42	356.10	380.83	−24.73**
Pr.2	Voc. ed. medium duration	737	1786	29.20	70.80	403.94	423.81	−19.87**
Pr.3	Basic (elementary) education	347	510	40.53	59.47	440.68	446.90	−6.22
Pr.4	Technical ed.	3207	3729	46.24	53.76	486.13	510.41	−24.28**
Pr.5	Sed+ – technical gymnasiums	512	524	49.38	50.62	537.36	566.13	−28.77**
Pr.6	Sed+ – general gymnasiums	3264	1993	62.09	37.91	576.78	596.31	−19.53**
	Total	8109	8664	48.34	51.66	515.77	510.14	5.63**

sed+ – secondary education; ** differences are statistically significant at $p < 0.05$.

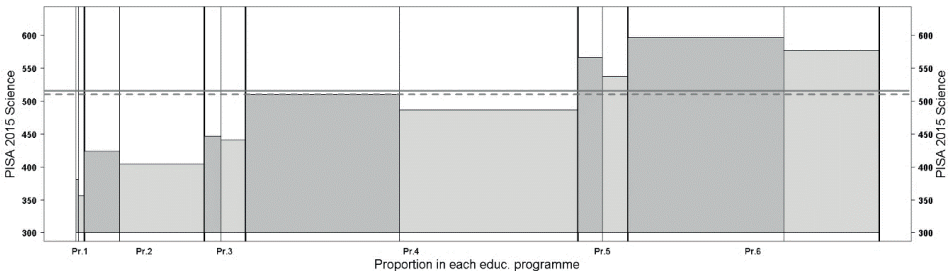


Figure 11 PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Slovenia. Width of each programme corresponds to proportion of a whole population. Lines show total average (dashed – boys, solid – girls).

PISA 2015 science data by gender and educational tracks for Slovenia demonstrates a Simpson's paradox. Since for Slovenia standard errors are quite small (SE_{GIRLS} = 1.88; SE_{BOYS} = 1.92) the difference of 5.63 points is statistically significant and shows that on average girls outperformed boys, while results in every educational track suggest otherwise.

Square representation graphs for Slovenia in Figure 11 show the remarkable shift of a Simpson's paradox. While generalizations from every educational track would implicate that boys outperform girls in PISA 2015 science literacy in fact the opposite is true!

We can summarize our findings with regards to our hypotheses as following:

H01: Differences between boys and girls in PISA science results within educational tracks are equal to overall difference between boys and girls.

Austria *CONFIRMED* – Overall difference and differences within educational tracks show same trend and are both statistically significant.

Croatia *NOT CONFIRMED* – Overall difference and differences within educational tracks show same trend but overall difference is not statistically significant.

Czechia *CONFIRMED* – Overall difference and differences within educational tracks show same trend and are both statistically significant.

Slovakia *NOT CONFIRMED* – Overall difference and differences within educational tracks don't show same trend and overall difference is not statistically significant.

Slovenia *NOT CONFIRMED* – Overall difference and differences within educational tracks don't show same trend and both are statistically significant in different directions!

H02: Differences between boys and girls in PISA science results within educational tracks and in total don't show the pattern of Simpson's paradox (reversed difference).

Austria	Croatia	Czechia	Slovakia	Slovenia
CONFIRMED	*CONFIRMED*	*CONFIRMED*	*NOT CONFIRMED*	*NOT CONFIRMED*

4 Discussion

A simple analyses of differences by gender or other characteristics are very common. Furthermore, due to the simplicity of calculating averages they are often not done and interpreted by statisticians alone but by people with wide variety of statistical knowledge. As Smith (2008) notes, secondary data analysis in general is often seen with scepticism because data, gathered for one reason is being used for another and this opens doors to errors. But even Smith (2008) recognizes great opportunities in using large scale data coming from well conducted research with good technical documentation. To avoid the pitfalls we must empower the researchers that use data. We demonstrated that researchers must be aware of possibilities for occurrence of Simpson's paradox and must pay attention against its effect on results and interpretations. This paper should empower researchers to keep guard and discover Simpson's paradox during analyses and thus provide correct interpretation of the findings.

Simpsons paradox can easily influence results of modern statistical analyses when we combine data sets from different sources and produce meta-analyses. Cohen and Moch (2017) warn researchers to be on guard and look for occurrences of Simpson's paradox when combining datasets. They provide examples from medicine, where samples are often small and the paradox occurs because different datasets are of different sizes. They cite cases where the results were different when datasets were analysed separately as when combined and conclude that only when researchers are prepared for the phenomenon of Simpson's paradox in advance can we avoid erroneous results and interpretations. Their results can be easily generalized outside medicine.

We should be aware that in case of Simpson's paradox it is not always straightforward which of the results is erroneous. In our PISA 2015 data the differences within educational tracks were misleading and difference in total dataset showed the real difference but it could easily be the case that total difference would be wrong and differences by subgroups would be correct. Baker and Kramer (2001) explored generalizations from studies of another set of medical interventions. They report on example where the treatment was better for males and females but when datasets were combined it appeared to be harmful to everyone!

The real examples from PISA 2015 data also provides several lessons. First lesson would be that it is important to follow proportions of boys and girls in different educational tracks. The proportions widely differ and the effects on educational systems in the long run can be profound.

Differences within educational tracks are interesting as they are heavily weighted by the proportions of boys and girls in each track and even more importantly by their preference for certain educational track. Boys and girls aren't allocated to educational tracks randomly but they rather select them according to their abilities and preferences. Some vocational and technical tracks can be more appealing to boys than girls and in other tracks situation might be reversed. From the point

of educational governance, it is important to wonder if observed proportions are a reason to worry or not. Educational systems around the world are often aware of such differences and try to act upon them and govern their educational systems to address this mostly through the questions of equity. One good example are initiatives to attract more girls into STEM. Such initiatives can be found globally and are among others supported by UNESCO (2014) and EU (2016).

Another lesson from our analysis is also that we shouldn't generalize findings from one educational track to others (or to educational system as a whole). It is often the case that countries have only data for one educational track (like specific leaving examinations that isn't available in other educational tracks). Findings from secondary analyses of such data shouldn't be generalized to the educational tracks where similar data doesn't exist or to whole educational system. As shown on example of PISA 2015 data we should consider the analysis carefully to avoid misleading interpretations.

Situations where differences in proportions have substantial influence on results are important to note regardless of the fact if there was an actual case of Simpson's paradox. In our PISA 2015 data Simpson's paradox occurred only in Slovakia and Slovenia, but similar underlying tendencies of smaller overall difference were also detected in all other countries. This is important for interpretation as it reveals that boys and girls in same educational track are not directly comparable. In case of Slovenia data shows great differences in gender composition in different educational tracks and this finding should serve as basis for raising the awareness about the issue and future steps that would address it. Since effects of education are often very long term and profound such warning signs should not be neglected.

The topic of this paper focuses on two main parts revolving around Simpson's paradox: theoretical and empirical one. Theoretical part warns the researchers to keep guard and spot Simpson's paradox when it occurs so the interpretations of the data are valid. We have demonstrated that Simpson's paradox isn't a statistical amusement, it's a real threat to validity of conclusions based on data and it's a clear signal of neglected and overlooked factors influencing the data. This brings us to our empirical part where we use PISA 2015 Science data to demonstrate Simpson's paradox but in the process we also uncover new insights. When gender of students is compared, many countries show differences in allocation of boys and girls to educational tracks, differences that raise questions of equity and fairness of each educational system, differences that can have long lasting effects in each country.

References

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366.

Baker, S. G., & Kramer, B. S. (2001). Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies. *Journal of Women's Health and Gender-Based Medicine*, *10*(9), 867–872.

Cohen, B. S., & Moch, P. L. (2017). Guarding against Simpson's paradox when combining data sets. *Curriculum and Teaching Dialogue*, *19*(1 & 2), 153.

Demers, S., & Rossmo, D. K. (2015). Simpson's paradox in Canadian police clearance rates. *Canadian Journal of Criminology and Criminal Justice*, *57*(3), 424–434.

European Commission. (2016). *She figures 2015*. Luxembourg: Publications Office of the European Union. Retrieved from http://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015-final.pdf.

Gardner, M. (1982). *Aha! Gotcha: paradoxes to puzzle and delight*. San Francisco: Freeman.

Lesser, L. M. (2001), Representations of reversal: An exploration of Simpson's paradox. In A. A. Cuoco, & F. R. Curcio (Eds.), *The roles of representation in school mathematics* (pp. 129–145). Reston, Virginia: National Council of Teachers of Mathematics.

OECD. (2017). *PISA 2015 technical report*. Paris: OECD Publishing.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. On-line: https://www.R-project.org.

Rücker, G., & Schumacher, M. (2008). Simpson's paradox visualized: The example of the Rosiglitazone meta-analysis. *BMC Medical Research Methodology*, *8*(1).

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, *13*(2), 238–241.

Smith, E. (2008). *Using secondary data in educational and social research*. New York, NY: McGraw Hill/Open University Press.

Tan, A. (1986). A geometric interpretation of Simpson's paradox. *College Mathematics Journal*, *17*, 340-341.

UNESCO (2014). *UNESCO's promise: Gender equality, a global priority.* Paris: UNESCO. Retrieved from http://unesdoc.unesco.org/images/0022/002269/226923m.pdf.

Wright, B. (2012). *Best of N contests: Implications of Simpson's paradox in tennis [Masters Thesis]*. The Florida State University.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, *2*(2), 121–134.

Dr. Gasper Cankar
Kajuhova 32 U
SI-1000 Ljubljana
Slovenia
gasper.cankar@ric.si