Insights into education

David Aas Correia

Project report submitted as partial requirement for the conferral of STKD6510: Data Analytics: Tools and Techniques for Acquiring Insights from Big Data II

Teacher:

Aiko Yamashita

September 2019

TABLE OF CONTENTS

1. Introduction

Education is considered fundamental for social development and is critical for the well-being and sustainability of a population. It helps individuals making informed decision on what is best for society and helps us to constantly innovate, grow and develop better solutions. Aside from ensuring prosperity, it is also critical for democracies to work and to motivate social cohesion.

Still, with richer societies and a growing global welfare, we are still facing huge differences, injustices and global problems. Education can be considered as one of the most important things the world should focus on in the coming years to solve pressing societal and environmental problems such as the climate crisis and migration flows.

In order to understand education better, I will in this project use the "Programme for International Student Assessment" (PISA) 2015 data to explore what contributes to students' achievement and differences between groups.

1.1 Problem definition

In addition to what was briefly discussed above, student achievement and educational differences has been researched in social sciences through centuries. The complexity of education is not to be underestimated and one should always be careful giving policy recommendations and conclusions. There is a problem of co-founding factors and endogeneity as many explanatory variables are related or are dependent of each other.

Nevertheless, research point out that socio-economic status, parental education and occupation and background explain student achievement, as well as differences between groups (Ammermueller, 2007; Brunello and Rocco, 2013; OECD, 2018).

The true problem of poor education is that it inhibits social mobility. Education is one of the best instruments to lift individuals – families, societies – to a better life in the future, and followingly next generation. Not only does poor education generate increase inequalities into the future, but it also limits democracy and individuals right to affect society. Education can be considered as the fundament for decision making and is understood as a human right. Thus, governments and institutions are expected to ensure quality education for its population. This is also defined as one of United Nations (UN) sustainable development goals (SDG). Lastly, from an economic perspective, educated individuals have better jobs and get more paid. Oppositely will lower education reduce the overall tax base, and on top of that, require significant investment in social policies for unemployed or individuals with low

income. The mentioned effects are important parts of the problem definition of education and reveals the complexity of the problem.

I will therefore in the next subchapter explain what I want to do in this project and how it can help understanding educational complexities better.

1.2 Answering the so-what question

As education undoubtedly are dependent of and affects many dimensions, it is always a topic that needs to be furtherly researched in order to get a better understanding, and thus, to help political and social institutions to know more what they can do to improve educational conditions.

I wanted with this project to look at the data without bias and we therefore used clustering algorithms to students' characteristics in order to label the data. This helped us to improve credibility and reduce arbitrary selection choices on the population. Furtherly, applying simple statistical methods helped us demonstrate differences in various educational outcomes and which characteristics that could create differences between students.

This type of research is critical because it lays the fundament for policy making for politicians and other decision makers. It helps them take more informed choices that reduces the uncertainty of political implementation.

I hope that with this paper to provide more insights into education.

2. Data

PISA aims to measure 15-year-old students' ability to meet the challenges of today's society[1]. It is a triennial survey and started in 2000. The 2018 data was unfortunately not available at the time of this project. Each survey includes three major domains: science, reading and mathematics. Students are also tested in their problem-solving skills and concept understanding.

The 2015 assessment was conducted in 72 countries and economies with approximately 540 000 student responses. Test items were a mix of multiple-choice questions and more elaborated tasks where students had to construct their own responses. OECD standardizes test scores with a mean of 500 and a standard deviation of 100. In addition to the students' scores, there is also a student questionnaire that provides more background information about each student and a school questionnaire that includes information of the school

[1]For more information, see http://www.oecd.org/pisa/.

context and its characteristics[2].

The final dataset, after data preparation, contained 247 237 observations (students), 40 explanatory variables and 3 score variables (science, mathematics and reading). Nonetheless, to reduce complexity, only the following variables were used for analysis purposes and they were split into three variable groups:

1. Student characteristics
    1.1. Gender
    1.2. Ethnic group category
    1.3. Country
    1.4. City (100 000+ inhabitants)
    1.5. Home possessions
    1.6. Socio-economic status
2. Parent characteristics
    2.1. Education
    2.2. Occupational status
3. School characteristics
    3.1. School type (private / public)
    3.2. Music activity at school
    3.3. Who decides course content at the school
    3.4. Fully certified teacher staff

3. Methodology

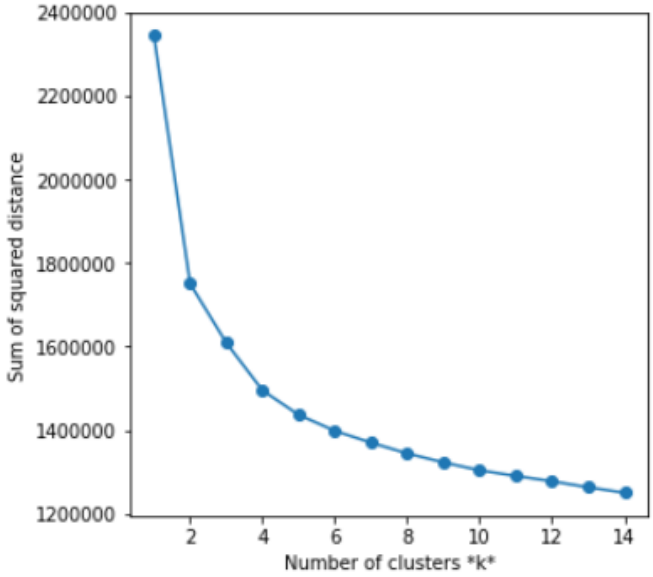
We used three methods in this project to gain insights into education:

a) Clustering
b) Visualizations (Microsoft Power BI)
c) Statistical methods (descriptive statistics and t-tests)

We used the K-means algorithm to label the final dataset. It divides n-observations into k groups where each observation belongs to the group's closest mean. The number of clusters were confirmed using the elbow method followed by the silhouette method. The first method looks at how much of the variance is explained as a function of the number of clusters. The idea is that at a certain point the marginal gain will drop, and thus, giving it an angle looking like an elbow. The figure on the next page shows the function for our dataset.

[2]For more specification or technical information about the data, see technical reports or other implementation tools at http://www.oecd.org/pisa/data/.

Figure 3-1: Elbow method



The elbow method gave us indication that we should use 4 or 5 clusters. To confirm this, we used the silhouette method. The silhouette score used in the interpretation measured how similar an object is to its own cluster compared to other clusters, ranging from -1 to 1, as higher values indicate similarity towards its own cluster. The average silhouette score for 4 clusters was 0,13 and 0,10 for 5 clusters. The combination of these two validations indicated that we should use 4 clusters for our analysis.

For dynamic visualizations we used Microsoft Power BI. We uploaded our dataset into the software and created a heatmap to visualize the proportions of students within each student category / label (high scoring, upper middle, lower middle and low scoring students). The stronger colour, the higher concentration, and vice versa for lower concentration. This allowed us to easily demonstrate strong discrepancies in education across the world. We could also increase the potential value of this analysis by including more information about the students.

Lastly, we used hypothesis testing to see whether some characteristics had positive or adverse effects on students' performance. We decided to research gender, music activity at school and which authority that determines course content. The test allowed us to check if there were significant different mean scores across the three domains between two groups. Based on the mean score, we could also get insights into which characteristic composition that had positive or adverse effect. Some underlying assumptions are present for t-tests and we addressed this by checking whether the score distributions of the two test samples were

normally distributed, as well Levene's Test of Equality to confirm homogeneity of variance between the subsamples.

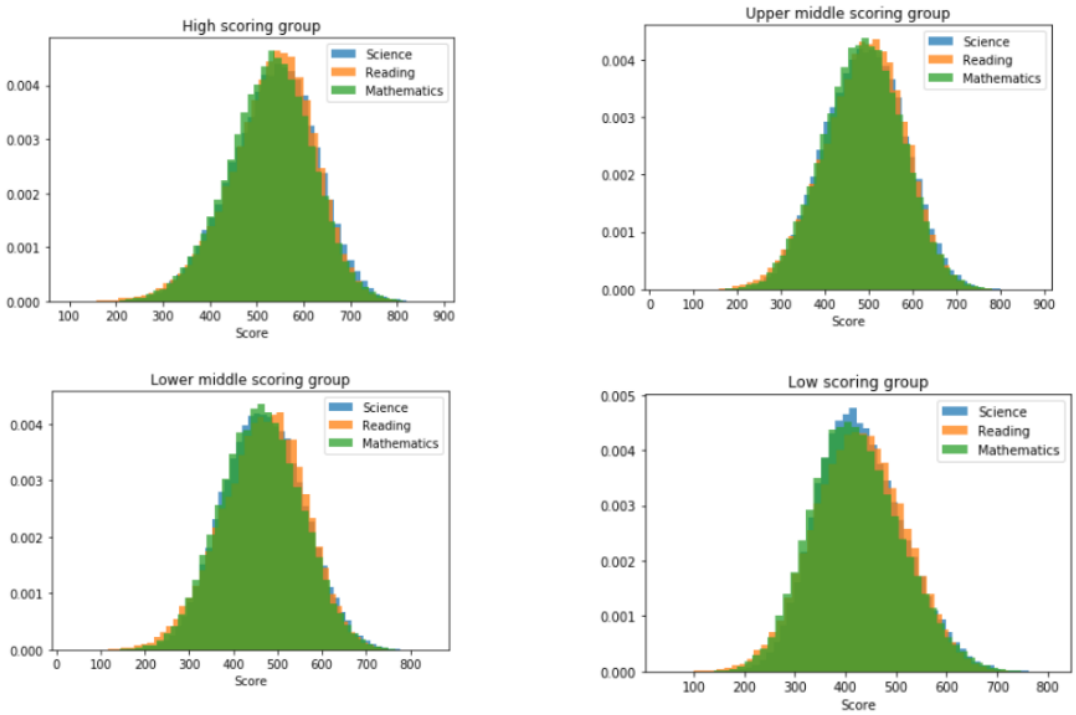
4. Results

We will in this section present and discuss our findings. The results should also be interpreted in the bigger picture and not be taken literally. Education is as pointed out in the beginning a social formation that involves complex human and societal interactions.

4.1 Clustering - K-means

The clustering algorithm helped us label the data without bias and we ended up with 4 score groups of students. Not only did we see clear score differences across these groups, hence the category names, but also different characteristic composition. Some descriptive statistics are presented below; count, mean scores and standard deviation (SD.):

Figure 4-1-1 Score distributions of the groups



Based on figure 4-1-1 and table 4-1-1 we can see that the score distributions within each cluster is approximately normally distributed. The low and high scoring groups are a bit skewed towards their score group positions. The standard deviations for each group are also very similar.

Furtherly, we can see in the tables in the next page that each group has some similar and different characteristic traits. Make notice that students are grouped across the world, so we cannot conclude on national compositions based on these clusters. To narrow down our analysis we emphasized on differences between the high and low scoring group.

Table 4-1-1 Descriptive statistics on scores of the groups

Profile	Count	Science	SD.	Reading	SD.	Math	SD.
1	64 681	535,4	90,7	532,3	90,1	527,3	89,6
2	56 413	491,3	89,9	489,1	91,0	486,1	89,7
3	57 293	469,0	90,0	466,7	93,6	462,2	90,9
4	68 490	431,1	84,8	427,3	89,0	422,0	88,8

Table 4-1-2 Descriptive statistics on characteristics of the groups

Scores	Male	Immigrants	Private school	City	Ec./social status	Centralised course content	Certified teacher staff
High	47,7%	12,1%	31,9%	51,3%	3,7	27,8%	53,3%
Upper middle	45,7%	7,2%	17,4%	35,5%	1,4	32,1%	54,2%
Lower middle	52,2%	15,2%	24,9%	43,7%	2,7	34,0%	46,3%
Low	48,1%	8,9%	13,0%	29,8%	0,4	42,4%	46,3%

Table 4-1-3 Descriptive statistics for parental education and occupation for low and high scoring group

Scores	Occupation				Highest education			
	Father		Mother		Compulsory school		Higher education	
	Low	High	Low	High	Father	Mother	Father	Mother
High	8,9%	63,8%	6,2%	57,9%	3,4%	2,5%	79,1%	82,5%
Low	56,9%	3,3%	48,0%	2,0%	45,9%	46,7%	9,9%	8,9%

The male percentage are approximately the same for both groups. However, the high scoring group has significantly higher concentration of students living in cities, higher socio-economic status and attends private schools. Differences in these characteristics are also thoroughly researched and are generally found to be important factors to explain differences in student achievement (Hanushek and Luque, 2003; OECD, 2018; Wößmann, 2016). We also found that parental education and occupation levels that were extremely different between the low and

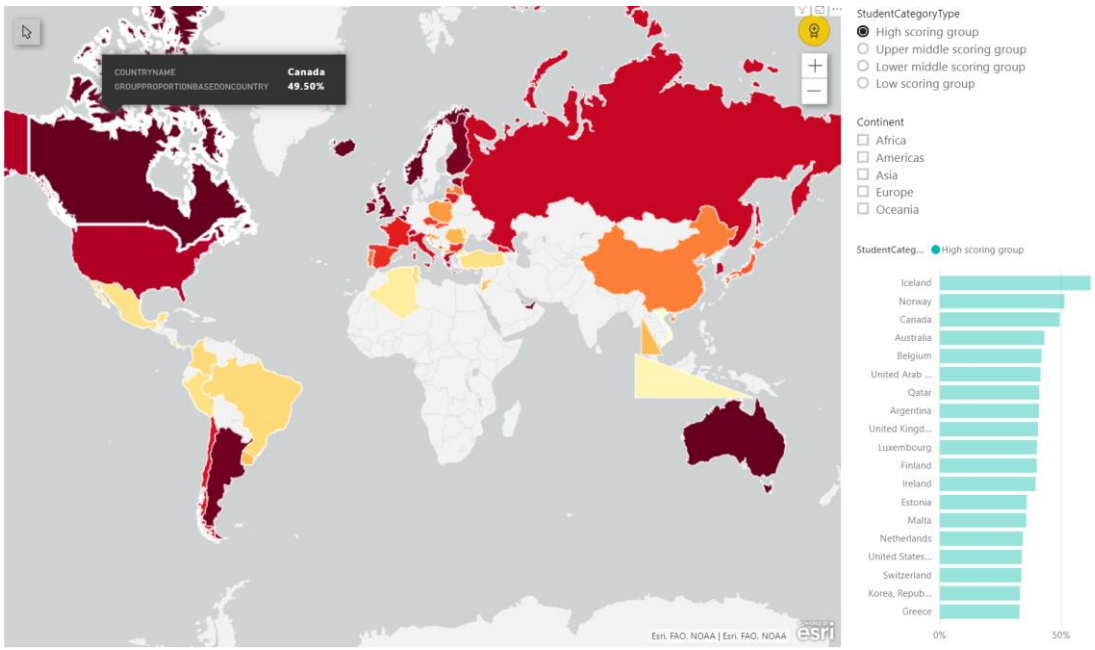
high scoring where about 80% of the high scorings group's parents had higher education against striking 10% for the lower scoring group. In terms of occupation we also observe that about 60% of the parents in the high scoring group enjoy "good" jobs, while about 50% of the parents the low scoring group have "low" score occupations (Ganzeboom, 1992).

Lastly, we can see that there is a higher concentration of high scoring students in situations where principals and teachers enjoy more autonomy to decide course content in their school. This is a very interesting insight and we will elaborate more on later as we perform hypothesis testing.

4.2 Visualization

We decided to use Power BI to create a heatmap of the world and to visualize the proportions of students in each score group. The BI tool did not only prove useful to easily visualize the data, but it also gave us the possibility to distribute and provide a self-service tool that others can use to gain insights into education. Furtherly, it allowed us to combine multiple graphical visualizations with dynamic filter options.

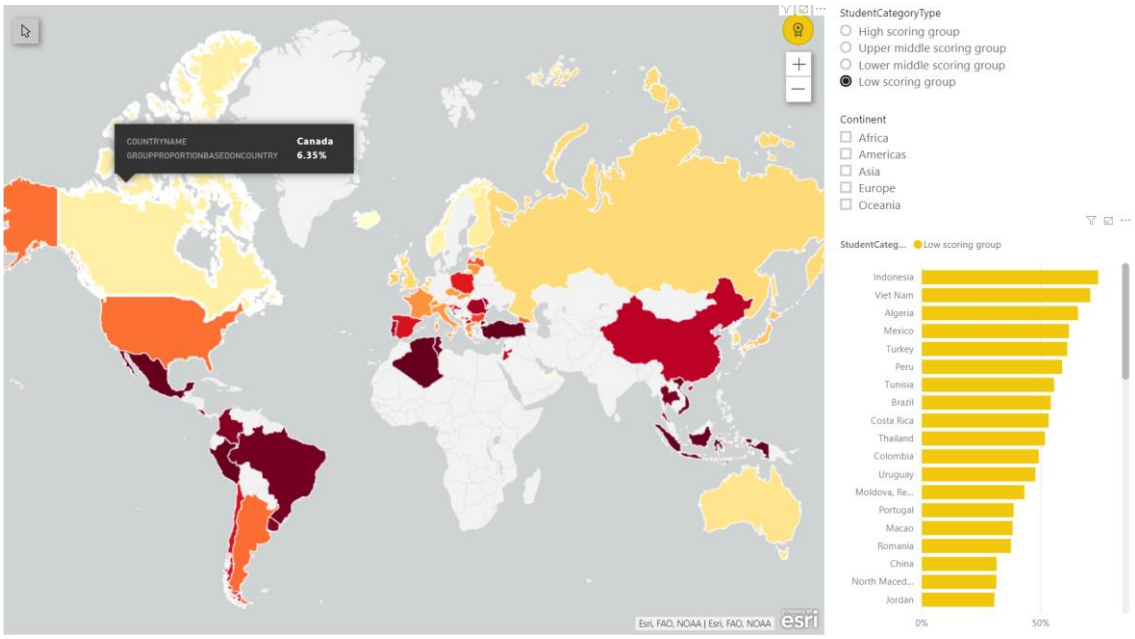
Figure 4-2-1 World map in Power BI, concentration of students in high scoring group



In both figure 4-2-1 and figure 4-2-2 we can see that many countries don't participate in the PISA test which furtherly increase the risk and complexity of drawing conclusions from the data. Nevertheless, we can see that there are clear tendencies that countries Europe, Canada, Australia and some countries in South America have high proportions in the high

scoring group while Latin-America, Africa and South-East Asian countries have higher proportions of students in the low scoring group.

Figure 4-2-2 World map in Power BI, concentration of students in low scoring group



The file can be distributed on request and the notebook used to provide the results are available in a GitHub repository[3].

4.3 Hypothesis testing

We tested several hypothesis and applied t-tests on each case to see whether the subsamples had significant mean score differences. Initially we tested the mean score differences between each score group and all of them were found significantly different for the three domains (science, reading and mathematics) on a 1% significance level.

4.3.1 Music activity in school

Our hypothesis was that students engaging in musical activity at school more frequently activated the creative part of the brain, and that this furtherly could have a positive effect on their performance. We firstly distinguished students that were offered this activity at school and not, and found following mean score differences (having – not having):

- Science: 33
- Mathematics: 32
- Reading: 32

Students with this activity seems to generally have better score, but we can't conclude that

[3] The GitHub repository is publicly available here: https://github.com/davidaascorreia/pr-oslomet-education.

it is because of that characteristic. We also did a t-test and found that the mean scores were significantly different. The same differences and results were found when we looked at one gender at a time.

4.3.2 Central authority decides course content

Our hypothesis here was that students attending schools were central authority decided local course content performed worse. We believe autonomy for schools and allowing principal and teachers to decide course content creates positive learning environments for students.
The mean score differences we found were:

- Science: 27
- Mathematics: 31
- Reading: 25

The t-tests also found the mean scores to be significant different, even when separating for gender. This school / policy setting is interesting. We think that some of the effect may come from differences in educational systems and more democratic systems, but we don't want to limit the effect to that. The effects of a policy implementation of more autonomy in deciding school autonomy for course content could be tested to control for its effects.

5. Conclusions

The aim of this project was to get insights into education. The clustering strategy gave us a good starting point for the data as we had used a data-driven approach to label the students. This also reinforced our results as we didn't make any biased assumptions or choices in the data preparation. It was also very interesting to see that unsupervised learning neatly split the students into four different score-groups. This exceeded out initial motivation to use a clustering algorithm on the data and it proved to work well on the dataset.

Followingly, we learned that Microsoft Power BI allowed us to explore the dataset in an easy way. We chose to make a visualization with the proportion of students within the different score-level groups as it was most aligned with our project motivation. Still, we had much more data available, so, if we had more time we would expand the data exploration and dive further into the possibilities here.

Lastly, we saw that students with either musical activity at school or schools where teacher and/or principals decided course content tended to perform better. It could confirm that local autonomy enhances performance, but this needs to be furtherly tested. A possibility could be to do an experiment between some schools and see whether these initiatives had expected effects or not.

5.1 Learning points / what could have been done differently

Education is complex. We learned that ideological approaches disturb recommended perceptions and motivations when looking at data. Despite using data-driven approaches, it is still challenging to look at the data without any assumptions or biases.

Even more important; what is the correct measurement for student performance? PISA uses science, mathematics and reading as principal domains to measure students' abilities. These are undoubtedly important subjects that determine students progress, but is it a universal measurement? After presenting this project at OsloMet, one of our professors asked us to give our opinion on how to construct a fair and sustainable educational system. Our answer was that creating school environments that involves constant and equal factors for the students could be one option. However, the real question (and answer) was more about how we define students' achievement. We answered that using students' welfare and fulfilment could measure something more universal. Measures on how meaningful school feels and subjective perceptions on anxiety, satisfaction, teamwork and more. These variables were also available in the PISA dataset and would be interesting looking into.

Anyhow, the discussion on measurement of student achievement and analysis on individuals pose, and will always pose, challenges. This is one of the most important lessons we got from our analysis. Despite having an extensive dataset, it was very likely that contained endogenous variables. The variables are interconnected and affects each other, so observing an effect on one variable, might be driven by a change in another one. It is difficult to control for all these effects. And because of this we want to underline that the results of this project may serve has pointers on what to pursue next, not as pure policy recommendations.

Lastly, we thought that looking at students as part of an "educational production function" to be wrong. Everybody has different motivations and personalities. Education and human interactions can not be simplified into simple input elements that provide output. Same here, a set of characteristic and environment variables, can not be applied universally and solve educational inequality and make every student score well on tests like this. Each student - each family – each school – each region – each country – has its own definition of success and what it important in life.

May this paper serve as insights into education for policy and decision makers and support them in the maze policy making.

5.2 Moving forward / what next

We only used the PISA 2015 in this analysis. But it exists datasets all the way back to 2000, and 2018 will soon be available, so it would enrich our results if we could consolidate it across

more years. It would also establish a time series so that we could make more advanced predictions on how effects have evolved and will evolve into the future.

School systems in the developed countries are going through dramatic changes with the use of digital and interactive learning. This may also increase the gap towards developing countries that still uses traditional learning methods. This and other institutional examples should be addressed in the future to increase the validity of analysis. Institutional settings are important to identify and isolate in order to look at similar peers as unit of analysis.

Furtherly, due to complexity, we would want to have more time to develop stronger statistical models and machine learning algorithms more tailored for the purpose. The results are more descriptive than predictive, so it would be very interesting to use more time researching how the different explanatory variables were correlated, to create interaction variables and enrich the data with more external data. We could see how performance behaves with changes in GDP, governments and other relevant settings.

This paper only addresses the "global" student, and therefore, splitting the data into more local subsets could improve the validity significantly. The demographic hierarchy is continent, country, region and school. Analysis could be performed on each of these levels. More local contexts enhance the applicability of the results greatly for policy makers. This, combined with our discussion on local autonomy on course content, could provide strong foundations for policy makers.

Finally, students' endowments are in many ways inherited thought their parents and we saw in our analysis the huge differences in parental education and occupation. This highlights that social mobility is highly dependent on education. The idea of looking at instruments that improves students that perform poorly today, may have great effects in the next generations. The generation multiplier is important to consider in policy making and educational implementation should emphasise actions that both help today but also considers that some effects only come for later generations. It could therefore be more helpful to make small changes to improve the situation a bit, rather than trying to solve the greater problems at first approach.

Bibliography

Ammermueller, A. (2007). Poor background or low returns? Why immigrant students in Germany perform so poorly in the programme for international student assessment. *Education Economics*, 15(2):215–230.

Brunello, G. and Rocco, L. (2013). The effect of immigration on the school performance of natives: Cross country evidence using PISA test scores. *Economics of Education Review*, 32:234–246.

Ganzeboom, H. B. G., De Graaf, P. M., and Treiman, D. J. (1992). A standard international socioeconomic index of occupational status. Social Science Research, 21:1–56

Hanushek, E. A. and Luque, J. A. (2003). Efficiency and equity in schools around the world. Economics of Education Review, 22:481–502.

OECD (2018). *The resilience of students with an immigrant background: Factors that shape wellbeing*. OECD Publishing, Paris.

Wößmann, L. (2016). The importance of school systems: Evidence from international differences in student achievement. *Journal of Economic Perspectives*, 30(3):3–32.