[1] **Yi Huang**
[2] ***Yan Zhang**
[3] **Haijian Hu**

Big Data on Exploring Influencing Factors of the Reading Achievement between China and Finland

Journal of Electrical Systems

Abstract: - This paper is based on Big Data analysis in computer science to find the significant influencing factors in reading achievement in China and Finland and differentiate factors between the two countries so as to find the implications and suggestions for the government policy and then reach the education equality and educational core competition. Reading is the main subject in PISA 2018. In 2018, about 60,000 students from 79 participating nations and economies finished the evaluation, accounting for around 32 million 15-year-olds. In all, 12058 samples in China(B-S-J-Z) and 5496 samples in Finland participate in the PISA 2018. The Plausible Value in Cognitive Process Subscale of Reading entails locate information, understand, evaluate, and reflect. Reading achievement is related to student-level, school-family level and country level. We use student-level constructs to measure every construct including GRADE, MISS, POSS, READINT, CLSSIZE, SELF, SES, SEX, READACH and design a model of three-level HLM in Big Data of factors influencing reading achievement. The Big Data statistics shows the PISA outcome has unstable level in PISA in 2012, 2015, and 2018 and the outcome cannot be applied to other places in China because it just includes four cities, B-S-J-Z. Also, big data also reflects the higher teaching level and teaching staffs in China and Finland than OECD Average although we cannot get all the data of regions in two countries.

Keywords: Big Data, Influencing Factors, Reading Achievement, PISA, Education Equality, HLM

I. INTRODUCTION

Big data [1] provides the framework for education to rethink its business strategy and form the alliance of corporations, governments, and social entrepreneurs necessary to bring together the knowledge, resources, and inventiveness necessary to provide universal access to lifelong learning. To probe into the reading achievement based on Big Data in computer science have been our hotspot in PISA. Different governments have taken some step to promote the students' comprehension and proficiency in science, math, and reading, which can be measured in PISA using Big Data, students aged 15 who participate in the triennial survey in the modern societies around the world. In PISA 2018, reading was the primary subject evaluated. Reading achievement has been discussed in connection with family, ESCS, ICT resources, gender and reading attitude respectively. However, few studies have concerned the influencing factors of the reading achievement as a whole from the perspectives of Big Data. One way to explore such factors is to make a comparison in two countries with the statistics methods. For example, earlier studies in this aspect are carried out between Germany and Spain with PLSPATH and HLM (Hierarchical Linear Modelling) software[2]. Geske and Ozola discovered that the school environment had a bigger influence on males' literacy in reading with using SEM among five countries and more recently[3], Kılıc Depren showed that students' socioeconomic and cultural backgrounds and metacognition abilities are the factors with the ARF (Activity Region Finder) algorithm[4]. Besides, a variety of methods including Hierarchical Liner Methods, Item response models, applied multilevel analysis, ANOVA and multiple regression can be applied into the elements that affect a reader's achievement in reading. However, although the elements affecting a student's ability to read was demonstrated over fifteen years ago, little attention has been paid to the comparison of the reading achievement from student-, school-, country-level, not just single-level multivariate. The present paper uses Big Data analysis to present multilevel factors influencing the reading achievement in comparison of the differentiate features between Mainland China and Finland with the dataset PISA 2018 all the way to promote education reform in the so-called double reduction policy. HLM is a useful tool for examining educational big data. Data sets grouped together, in which observations are arranged into multilevel units utilizing HLM as a useful data mining tool in computer science, are frequently seen in educational research [5]. On the basis of multilevel factors in computer science, contrary to SEM, HLM are applied commonly from multilevel multivariate factors in this research because Hambleton and Kanjee have pointed out the usefulness of these techniques for examining descriptive models of the structure of cognitive abilities such as reading [6]. The combination of these two techniques in Big Data as a

[1] School of Education, City University of Macau, Taipa 999078, China
[2] Guangdong Polytechnic of Science and Technology, Zhuhai 519000, China
[3] Department of Marxism, Guangdong Polytechnic of Science and Technology, Zhuhai 519000, China
*Corresponding author: Yan Zhang

data mining tool can be a novel method in the multilevel analysis of multivariant in reading achievement and reflect their whole differentiate features between Mainland and Finland [7].

A. Research Aims and Objectives

Reading literacy is the ability of pupils to comprehend, apply, assess, consider, and interact with texts in order to fulfill their own objectives, advance their understanding and capacity, and take part in society [8]. Finland and China are strong performers in PISA including reading achievement and their advantages in success of it have differ in various factors. To find the difference and learn from each other has a great significance on education in China although the population and GDP of Finland is less than China. The aim of this present paper is to find the significant influencing factors in reading achievement in China and Finland and differentiate factors between the two countries so as to find the implications and suggestions for the government policy and then reach the education equality and educational core competition. Reading achievement is indispensable ability for daily life, which concerns literacy, acquisition of new knowledge, execution of critical thinking, communication and the world view. To find the problem in education, we try to explore the significant factors of reading achievement in China, which is carried out by the assessment PISA 2018. The OECD conducts a triennial survey of students aged 15 called the International Student Assessment Program (PISA). It is aimed at assessing reading, mathematics, science and problem solving and then evaluates how well students can apply what they have learned and draw conclusions from it in new situations, both within and outside of the classroom [8]. It focuses on the all-round development for future life and lifelong education and education has complex association with ESCS, ICT, gender and reading ability and so on. To be more specific, we have three objectives: (1) to find the influencing factors in reading achievement between Mainland China and Finland. (2) to find the differentiate factors and the reason of this phenomenon in reading achievement between Mainland China and Finland. (3) to find the targeted suggestions in multilevel factors for the improvement of reading achievement and then the education reform in China.

B. Research Focus

To promote the education reform and development and find the influencing factors in reading between Mainland China and Finland, we should make a comparisons between them in multivariant of multilevel factors, including ESCS, gender difference, ICT, school climate, cultural difference, we seek to solve these following questions:

(1) Is there a distinct difference in influencing factors of reading achievement between Mainland and Finland?

(2) What is the most significant difference in variables of reading achievement?

(3) Is there a gender gap in reading achievement in China and Finland?

(4) Which level (family, school, country and student) may be the factors determining the students' accomplishments in reading?

II. METHODOLOGY

PISA is a triennial worldwide study conducted by the OECD on pupils aged 15 years. Evaluations aim to determine not only whether students approaching the conclusion of their required education can duplicate what they have learned, but also how well they can apply what they have learned in new contexts of both inside and outside of the classroom [9]. The PISA started with the year 2000, which focus on reading, mathematics, science, reading and problem solving. Triennial PISA has its main subject in assessment, to be more accurate. In 2000 and 2009, the most popular subjects were reading and mathematics; in 2003 and 2012, science was the most popular subject; in 2006 and 2015, science was the most popular subject; and in 2009 and 2018 there was the most reading. The primary emphasis of the PISA 2018 study was reading, with modest assessments in mathematics, science, and global competency, and young people's financial literacy optionally although reading is the main subject in PISA 2018. In 2018, about 60,000 students from 79 participating nations and economies finished the evaluation, accounting for around 32 million 15-year-olds. Different from other PISA assessments, PISA 2018 is carried out as a two-hour test via computer. Reading literacy in PISA 2018, rather than reading, identified four processes, specifically finding information, comprehending, assessing, and reflecting, as well as reading proficiently. PISA 2018 divide the student proficiency into three levels of proficiency: comparatively high, moderate, and poor in accordance to Item I-II, III-V, VI-VII based on Item response theory models. The reading literacy is divided into eight levels correspond to below 1c, 1c,1b, 1a, 2, 3, 4, 5 and 6 with increasing more difficult tasks.

A. PISA 2018 Dataset in Mainland and Finland

The proportion of 15-year-olds is Over 96 percentage in Finland and 81 percentage in Mainland China who covered by the PISA sample, and it may be on the account of remote location, level of education, Lack of competency in the exam language, physical or intellectual handicap, and accessibility of the school, or insufficient exam materials available in the instruction's language. Lacking economic source , dropout and exclusion may have impact on the low percentage in Mainland China. The mean score in reading achievement of Mainland China (B-S-J-Z) and Finland are statistically significantly above the the OECD average.The current investigation involved sampling from the original dataset in Mainland and Finland and analyzing the influencing factors of reading achievement in three levels. three levels were selected from the factors such as bullying, disciplinary climate, The following factors have been covered in previous studies: student truancy and tardiness; A student's sense of belonging at school, their level of self-efficacy and fear of failing, their development mindset, their instructors' excitement, their support and teaching techniques, their conduct and the way they teach, their competition and collaboration, and their sense of belonging at school are all factors. 12058 samples in China (B-S-J-Z) and 5496 samples in Finland participate in the PISA 2018. And the dataset is originated from https://www.oecd.org/pisa/data/2018database/. The Plausible Value in Cognitive Process Subscale of Reading entails locate information, understand, evaluate and reflect. Below Level 2（1a,1b,1c）according to the PISA 2018 defines as "low proficiency level [9]. 5.1% of Mainland and 13.5 of Finland students are at a lower proficiency level while the OECD is 24% in the PISA 2018 reading achievement. Above level 4 is defined as the upper proficiency level[10]. The distinction is seen in Figure 1 The Percentage of reading achievement level by country. To make sure the differentiate features of reading achievement in Mainland and Finland, we use the HLM (Hierarchical linear modelling) to analyze the factors in country level, student level and school-family levels. Referred to as Figure 2, the conceptual framework of elements influencing reading performance describes the components that affect reading achievement.

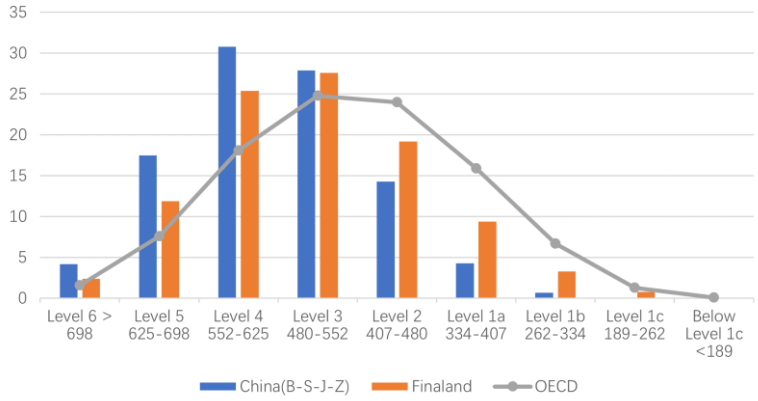

Figure 1 The Percentage of reading achievement level by country

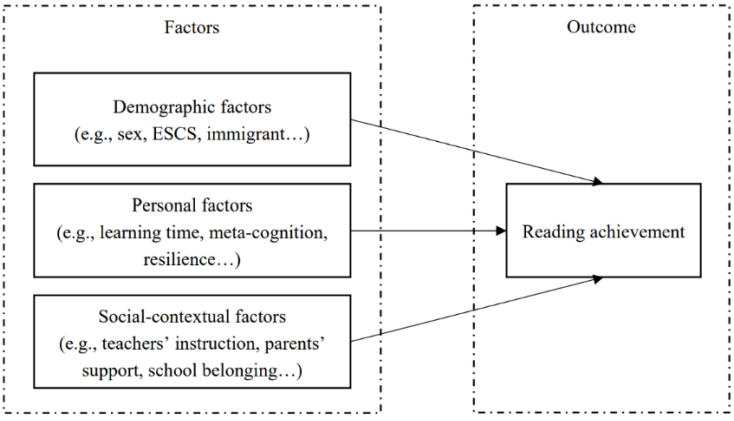

Figure 2 Conceptual framework of factors influencing reading achievement

B. Instruments and Procedures

In the 1990s and early 2000s, the social sciences began talking about HLM as a "new" methodology for handling nested data. Hierarchical Linear Modeling (HLM), different simple linear model, examines changes in the outcome variables when the predictor variables are situated at various levels of the hierarchical structure using a sophisticated version of ordinary least squares (OLS) regression [11]. HLM or multilevel models (sociology, political science) is a technique widespread in education which allows for the common use of HLM in multilevel data analysis. Popularized by education research, which produced textbooks that derived HLM as a generalization of regression models. In this study, we use HLM to find out the influencing factors in Mainland and Finland in level 1 student level, level 2 school-family level, level 3 country level. The HLM can predict the reading achievement of student level in country level. For example, disadvantaged countries (level-3 variable) will have low income in family and poor software and hardware of manufacture in school, which lead to a relatively weak reading achievement compared to advantaged countries. School-family level, student level and country level are combined to influence reading achievement. And we choose the variables significantly in statistics from the PISA 2018 and divide them into three levels. The variables of factors in Mainland and Finland are seen in Table 1. Such factors as the index of teacher enthusiasm, Adaptive education, teacher-directed instruction, and teachers encouraging students' reading engagement, enjoyment of reading, teacher behavior, student cooperation, student competition, sense of belonging, parental involvement, parents' perceived school quality, positive feelings are positively or negatively linked to reading achievement. Besides, home possessions, ICT resources, ESCS, disciplinary climate are also important factors in reading achievement. However, such factors may be different in Mainland China and Finland.

C. Data Analysis

Table 1 Descriptive statistics of the variables in Finland

Variable(Finland)	Obs	Mean	Std. Dev.
METASPAM	5,169	0.19	1.00879
JOYREADP	0		
SCREADCOMP	5,396	0.09	1.000802
METASUM	5,240	0.022067	1.006544
PISADIFF	5,387	-0.25698	0.966521
UNDREM	5,247	-0.09985	1.027155
SCREADDIFF	5,407	-0.10555	1.048627
MASTGOAL	5,265	-0.11835	0.921665
EUDMO	5,298	0.058156	0.942835
COMPETE	5,355	-0.02694	0.974617
WORKMAST	5,269	-0.31748	0.944679
RESILIENCE	5,269	-0.03392	0.952264
ATTLNACT	5,402	0.019175	0.95
SWBP	5,293	-0.11926	0.93
LMINS	4,829	150.8397	45.37
REPEAT	5,526	0.031849	0.18
HEDRES	5,566	-0.3	0.91
ESCS	5,557	0.3	0.79
ICTRES	5,573	0.15	0.73
IMMIG	5,524	1.09	0.38
EMOSUPS	5,143	-0.049	0.99
DISCLIMA	5,535	-0.11	0.95
BEINGBULLIED	5,017	-0.03	0.96
STIMREAD	5,419	-0.2	0.95
PERFEED	5,433	-0.16	0.93
PERCOMP	5,057	0.1	0.86
BELONG	5,264	0.01	0.99
TEACHINT	5,434	-0.15	0.93
PERCOOP	4,926	0.08	0.90
ADAPTIVITY	5,437	0.06	0.96
DIRINS	5,488	-0.11	0.97
TEACHSUP	5,525	0.21	0.91

Hierarchical Linear model is used in this study to explore the reading achievement from the student, family and school, country level[2]. We construct the hierarchical level in the figure 3. It tests the links between the three levels and reading achievement. The missing value can be deleted from the original dataset. In the second step, we process the data via Stata software to form the descriptive statistics of reading achievement in Mainland China and Finland. In the third step, we construct the three level from the factors of descriptive statistics. In the fourth step, every model in country level, family-school level, student level can be estimated to decide the important model in reading achievement between Mainland China and Finland in Table 1. Country variable might affect the family and school level [12], and all the way to affect the student level, including motivation, self-efficacy, resilient and learning time. All of them can be reflected on the students' reading achievement in a given countries. In the Level 1 student level , we use the i to represent the student, the outcome variable Yij, which refers to the student i in school j so that the model can be

$$Yijj=\beta 0jk+\beta 1jkXjk+eijk$$

In the level 2 school level, the model can be

$$\beta 0jk=\gamma 00k+\gamma\ 01kw1jk+\mu 0jk$$

$$\beta 1jk=\gamma 10k+\gamma\ 11kw1jk+\mu 1jk$$

here γ00k and γ10k is stable element, μ0jk and μ1jk is random element. In level 3 country level, the model can be

$$\gamma\ 00k=\Pi 000+\Pi 001Zook + e00k$$

$$\gamma\ 01k=\Pi 010+\Pi 011Z01k+e10k$$

$$\gamma\ 10k=\Pi 100+\Pi 101Z10k+e10k$$

$$\gamma\ 11k=\Pi 110+\Pi 111Z11k+e11k$$

In this study, the reading achievement can be estimated as the model:

$$Yijk=Y00K+a\ \beta\ ijk + b\ \gamma\ JK+cejk+d\ \mu\ jk$$

Here, the outcome variable Yijk of student i in school j in country k has Y00k , the mean reading achievement country. β ijk represents student level, γ jk represents family level, ejk represents school level, μ jk represents country level. a, b, c, d to assess the average impacts of these factors on reading achievement, need to be estimated [12-13]. The expected outcome will be clarified into the Chapter Result.

III. RESULTS AND DISCUSSION

A. *Results*

This chapter explain the differences in students' reading achievement of student, family, school, country variables [14]. The variables of China and Finland are seen in the Table 2. It is clear that the impact of every prediction in Mainland China and Finland separately in the perspectives of model1, model 2, model 3, model 4. And in the figure 3 the mean reading achievement in Mainland China and Finland are all above the OECD average. Finland has the relatively stable level from PISA 2012 to PISA 2019 while China(B-S-J-Z) has fluctuated in these years. As It can be seen in the figure 1, Low achievers in reading achievement of Below Level 2 has slightly increased in Finland while top performance in reading keep a rather stable level in Finland.

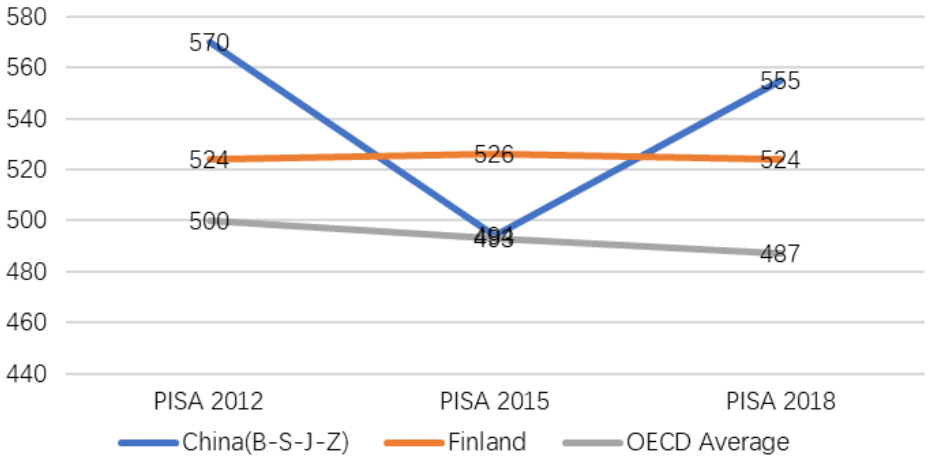

Figure 3 Change in mean reading achievement between Mainland China and Finland

In this study, three-level analyses via HLM software can draw the conclusions of difference between China and Finland in reading achievement. Reading achievement is related to student-level, school-family level and country level. However, every variable or factor may influence reading achievement, the significant statistics in reading achievement have differed among them. These factors can make us to adopt corresponding teaching strategies and the way to keep a good and balanced family and school relationship. In addition, the level is the most important and significant constructs, which might be valuable to educational institutions. The government can adopt the policy to promote the equality of education from macrolevel and microlevel, avoid the involution of education, such as the double reduction policy, TPACK teaching model, favor rural education, moderately prosperous society, strengthen the quality of teachers.

Table 2 Descriptive statistics of the variables in Mainland China and Finland

Variable(China）	**Obs**	**Mean**	**Std. Dev.**
METASPAM	11,842	0.09	0.96
JOYREADP	0	/	/
SCREADCOMP	11,907	0.08	0.87
METASUM	11,834	-0.12	0.97
PISADIFF	11,951	0.03	0.92
UNDREM	11,767	0.2	0.99
SCREADDIFF	11,932	0.12	0.96
MASTGOAL	11,935	0.06	0.91
EUDMO	11,957	0.089	0.92
COMPETE	11,954	0.42	0.82
WORKMAST	11,943	0.29	0.89
RESILIENCE	11,982	-0.07	0.96
ATTLNACT	11,968	0.16	0.93
SWBP	11,896	0.1	0.89
LMINS	11,905	266.2	116.8
REPEAT	11,990	0.06	0.24
HEDRES	11,988	0.27	0.99
ESCS	11,990	-0.36	1.087
ICTRES	11,988	-0.4	0.95
IMMIG	11,952	1.003	0.076
EMOSUPS	11,952	0.0057	0.93
DISCLIMA	11,987	0.82	1.03
BEINGBULLIED	11,899	-0.23	0.88
STIMREAD	11,986	0.63	1.03
PERFEED	11,968	0.35	1.04
PERCOMP	11,919	0.16	0.95
BELONG	11,977	-0.15	0.91
TEACHINT	11,990	0.38	0.97
PERCOOP	11,916	0.23	1.01
ADAPTIVITY	11,966	0.43	1.04
DIRINS	11,990	0.51	1.02
TEACHSUP	11,988	0.42	0.89

B. Discussion

According to the descriptive statistics between Finland and China, they keep the same level in the METASPAM, SCREASCOMP, METASUM, PISADIFF, UNDREM, SCREADIFF, MASTGOAL, EUDMO, COMPETE, WORKMAST, RESILIENCE, ATTLNACT, SWBP etc. And China has relatively higher gap than Finland in ESCS and HEDRES. Chinses students usually spend more time in studying than Finland, and their LMINS are 2400 and 840, respectively.

Reading achievement is related to student-level, school-family level and country level. We use student-level constructs to measure every construct including GRADE, MISS, POSS, READINT, CLSSIZE, SELF, SES, SEX, READACH and design a model of three-level HLM of factors influencing reading achievement.

IV. CONCLUSION

In this study, three-level analyses via HLM software can draw the conclusions of difference between China and Finland in reading achievement. Reading achievement is related to student-level, school-family level and

country level. However, every variable or factor may influence reading achievement, the significant statistics in reading achievement have differed among them. These factors can make us to adopt corresponding teaching strategies and the way to keep a good and balanced family and school relationship.

The PISA outcome have unstable level in PISA 2012, 2015 and 2018 and the outcome cannot be applied to other places in China because it just includes four cities, B-S-J-Z. The big data with data mining using HLM reflects the higher teaching level and teaching staffs in China and Finland than OECD Average although we cannot get all the data of regions in two countries.

REFERENCES

[1] Che D., Safran M., & Peng Z. (2013). From Big Data to Big Data Mining: Challenges, Issues, and Opportunities. In Hong B., Meng X., Chen L., Winiwarter W., & Song W. (Eds.), Database Systems for Advanced Applications (Vol. 7827, pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40270-8_1

[2] Kotte, D., Lietz, P., & Lopez, M. M. (2005). Factors Influencing Reading Achievement in Germany and Spain: Evidence from PISA 2000. International Education Journal, 6(1), 113–124.

[3] Geske, A., & Ozola, A. (2009). Different Influence of Contextual Educational Factors on Boys' and Girls' Reading Achievement. Online Submission, 6(4), 38-44.

[4] Kılıç Depren, S., & Depren, Ö. (2022). Cross-cultural comparisons of the factors influencing the high reading achievement in Turkey and China: Evidence from PISA 2018. The Asia-Pacific Education Researcher, 31(4), 427-437.

[5] Weng W., & Luo W. (2023). A Comparative Analysis of Data Mining Methods and Hierarchical Linear Modeling Using PISA 2018 Data (SSRN Scholarly Paper 4501442). https://papers.ssrn.com/abstract=4501442

[6] Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. European Journal of Psychological Assessment, 11(3), 147-157.

[7] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. IEEE transactions on knowledge and data engineering,26(1), 97-107.

[8] OECD. (2019). PISA 2018 Results (Volume I): What Students Know and Can Do. OECD. https://doi.org/10.1787/5f07c754-en

[9] OECD. (2013). PISA 2012 Results: Ready to Learn (Volume III): Students' Engagement, Drive and Self-Beliefs. OECD. https://doi.org/10.1787/9789264201170-en

[10] Kilic Depren, S. (2018). PREDICTION OF STUDENTS' SCIENCE ACHIEVEMENT: AN APPLICATION OF MULTIVARIATE ADAPTIVE REGRESSION SPLINES AND REGRESSION TREES. Journal of Baltic Science Education, 17(5), 887–903. https://doi.org/10.33225/jbse/18.17.887

[11] Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. Tutorials in quantitative methods for psychology, 8(1), 52-69.

[12] Chiu, M. M., & McBride-Chang, C. (n.d.). Family and Reading in 41 Countries: Differences Across Cultures and Students. 31.

[13] Ning, B., Van Damme, J., Gielen, S., Vanlaar, G., & Van den Noortgate, W. (2016). What Makes the Difference in Reading Achievement? Comparisons Between Finland and Shanghai. Scandinavian Journal of Educational Research, 60(5), 515–537. https://doi.org/10.1080/00313831.2015.1062413

[14] Porras, A. M. L., García, M. D. M. R., Muñoz, D. M., & Rodríguez, B. C. (2018). Identifying the Factors Influencing the Scientific Competence in Andalusia: A Multilevel Study of the PISA 2012 Results. The Eurasia Proceedings of Educational and Social Sciences, 9, 200–208.