

# Changes in Achievement in and Attitude toward Mathematics of the Finnish Children from Grade 0 to 9—A Longitudinal Study

Jari Metsämuuronen<sup>1,2</sup> & Laura Tuohilampi<sup>3</sup>

<sup>1</sup> Faculty of Behavioural Sciences, University of Helsinki, Finland

<sup>2</sup> Finnish Educational Evaluation Centre (FINEEC/KARVI), Finland

<sup>3</sup> Faculty of Behavioural Sciences, University of Helsinki, Finland

Correspondence: Jari Metsämuuronen, Finnish Education Evaluation Centre, PB 28 (Mannerheimintie 1A), 00101 Helsinki, Finland. Tel: 358-400-579-848. E-mail: jari.metsamuuronen@karvi.fi

Received: April 19, 2014

Accepted: September 14, 2014

Online Published: October 15, 2014

doi:10.5539/jedp.v4n2p145

URL: <http://dx.doi.org/10.5539/jedp.v4n2p145>

## Abstract

Recent years, Finland has been one of the countries of interest in education because of its success in international comparisons. Several attempts have been made to explain what could have been behind the positive results. However, some of the challenges of Finnish education, such as the productivity (achievement/costs) or its uniformity throughout the school years have not been emphasized. Further, it is under examined in Finland, as well as worldwide, the development of the performance and the attitude during the school years. Here, 3,502 stratified sampled Finnish students' achievement and attitude regarding mathematics were followed up from the beginning of the school (grade 0, age 7) to the end of the compulsory education (grade 9, age 16). The test scores from the different measurements were equated by using IRT modelling. The sharpest change in achievement happens during the lower grades and it evens out towards the upper grades. The achievement level of the student population entering the school is very heterogeneous. The actions during the first two years make the differences between the students disappear almost totally. The attitudes are declining during the years. During all the grades, boys feel themselves more self-efficacy in mathematics than the girls.

**Keywords:** longitudinal study, lower secondary school, mathematics teaching, primary school, student achievement, student attitudes

## 1. High Achievement and Low Attitudes in Finland?

### 1.1 High Achievement of the Students

In September 2012, an independent American research center, Pearson, ranked the Finnish Educational system as the top of the world (<http://thelearningcurve.pearson.com/index/index-ranking>). The result was based on the Global Index of Cognitive Skills and Educational Attainment which compares the performance of 39 countries and one region (Hong Kong) on two categories of education: Cognitive Skills and Educational Attainment. The Cognitive Skills were measured by the conjoint PISA, TIMSS and PIRLS scores in Reading, Maths and Science. The Educational Attainment was measured by literacy and graduation rates in the country.

This was not the first time the Finnish system was raised as an example for the others. Metsämuuronen, Kuosa & Laukkanen (2013) have noticed that, during the new millennium, the Finnish educational system has faced a new challenge: how to explain the glorious PISA results (OECD, 2001; 2003; 2007; 2010a; 2010b) produced with only a small variance between schools (Schleicher, 2006, 13), average national costs (OECD, 2005, 10-12) and, as regards the average duration of studies, relatively efficiently (e.g., SCP, 2004; Sutherland et al., 2007; Alfonso & St. Aybun, 2006; Clements, 2002). Explanations for this issue are searched for in many different ways. One possible approach is to focus on the basic **structures of the Finnish education system** in a European context (c.f. Aho, Pitkänen & Sahlberg, 2006; Laukkanen, 2008; 2013; Raivola, 2006; Sahlberg, 2006; Simola, 2005; Välijärvi, 2004; Välijärvi et al., 2007) and some specific features of it, such as in **teachers' quality** (see Niemi, 2010; 2011; Niemi & Jaku-Sihvonen, 2006; 2011; Sahlberg 2011a, 2011b; Schleicher, 2011). Another approach has been the **complex contextual factors** (see Lavonen & Laaksonen, 2009; Niemi, 2012; Reinikainen, 2012; Niemi, Toom, & Kallioniemi, 2012; Sulkunen et al., 2010). Still another way is to focus on the strengths of the

futures oriented **sustainable leadership** in the educational governance of Finland (i.e. Metsämuuronen, Kuosa & Laukkanen, 2013; Laukkanen, 2008; 2013; Sahlberg, 2007; Aho, Pitkänen & Sahlberg, 2006, pp. 126-133; Simola, 2005; Välijärvi, 2004; see also Hargreaves, 2006).

Naturally, many of the possibilities mentioned above may explain the results conjointly, or, maybe, the whole thing is a result of contingencies, unplanned decisions after each other as a Finnish educational sociologist Simola (2005) has argued. Whether the last is true or not, Metsämuuronen, Kuosa and Laukkanen (2013) suggested that it is hardly just a coincident in Finland that after 40 years of “common education for all”—after the first full generation of parents could help their children in their school assignments and give them home tuition — the results in student achievement are high. There may be some deep undercurrents also, which may explain the high results; working hard, for example—formerly known as “Lutheran working ethics” (Saarinen, 2005) or other words “In the sweat of thy face shalt thou eat bread” (1 Ms 3:16 according to King James Version)—is highly appreciated in Finland.

### 1.2 Low Attitudes of the Students

It is somewhat interesting that while the achievement level of mathematics is high, the attitudes toward mathematics and the general school satisfaction are quite low in Finland. Somewhat older comparisons in the international settings show that the Finnish pupils’ general school satisfaction was ranked either at the same level as that of other European pupils (Linnakylä, 1993; Arinen & Karjalainen, 2007, 63-65) or (at least partly) markedly lower (Linnakylä, 1993; Kannas, 1995). The most positive results regarding the attitude of pupils were noted in other Scandinavian countries and in the USA (Linnakylä, 1993). Also, research conducted by WHO (Kannas, 1995) showed that school negativity was clearly quite common in Finland. In a comparison of 20 countries, Finnish boys were ranked second to last. Leino (2003, 79) noted that pupils in Finland (as in other Nordic countries) are relatively humble when they describe their knowledge. This “humbleness” may also be reflected in attitude measurements.

It is quite clearly shown in several studies that the attitudes toward school have a tendency to decline during the years (see the discussion in Metsämuuronen, Svedlin & Ilic, 2012). Metsämuuronen studied the general attitudes toward the Mother tongue in Finnish language (2006a) and in Swedish language (2006b) and compared these (2006c) using a longitudinal design. The pupils were 7th and 9th graders. Among the Finnish-speaking pupils, the attitudes toward the subject declined by 8 percent units during three years. The data on Swedish-speaking pupils showed a significantly smaller reduction of 0.3 percent units. More recently, Metsämuuronen (2010) followed-up 4,545 pupils from grade 3 to 6 and noticed that the girls’ general attitudes toward Mathematics declined 13.6 percent units and boys’ attitudes 9.2 percent units. The study of Metsämuuronen, Svedlin and Ilic (2012), simultaneously with all grades of compulsory education in Finland, showed that the decline in attitudes is very intense after the two first grades but its evens out at grades 9 to 12. A deficiency of the study was the cohort design, that is, that the students were not followed up and hence it was not known how the individual attitudes changed over the years.

Derived from above, it is evident that at least the *end-product* of the Finnish system seems to be at quite a high level when it comes to achievement level though less high when it comes to the attitudes toward mathematics. To provide further clarifications, this study approaches the Finnish story by examining the contextual factors relating to students, their families, peer groups, teachers, schools, and demography. However, the contribution does not limit on explaining the success, as, there is much less evidence—if at all—on the *processes* or achievement level of the students during their path from zero level to the final grade. Thus, also the development of Finnish students’ achievement and attitude regarding mathematics is discussed. The development is especially interesting, as albeit the many possibilities that can be behind the good achievement, a recent result of Metsämuuronen (2013) notices that the *productivity* (achievement/costs) of Finnish education varies over time. On the basis of a longitudinal dataset of grade 7 and 9 students in Mother tongue it seems that the productivity in the Finnish lower secondary education is not very high. It was found that the productivity decreased by 7.5% during the lower secondary school years: i.e. costs increased by 7.5% more than the mean/median achievement level of the pupils. According to Metsämuuronen’s estimation, an apparent threshold value of pupils’ achievement levels in relation to school productivity was the 13–14 percent unit increase in achievement during the three years of lower secondary education. In other words, the schools appeared to be productive when the pupils increased their achievement levels by approximately +13 percent units of the maximum score, or more, within three years. This value equated to, approximately +5 percent units of the maximum score per year. In his study, the mean change over three years was +8 percent units. Hence, the lower secondary schools in Finland

(and within the subject of Mother tongue) were unproductive – they spent more money than they were capable of product achievement.

This article focuses on the path of the Finnish students on the basis of a large scale follow-up study by using equated test scores of 3,502 students. The article shows how the achievement level in Mathematics as well as the attitudes towards the subject changes from the beginning of the school within the nine years of compulsory education (Section 3.1), what kinds of profiles of change were found (Section 3.2), how the distributions of achievement in different grades deviate from each other (Section 3.3), how the boys and girls differ from each other (Section 3.4) and finally, which factors seem to explain the change in achievement (Section 3.5).

## 2. Methodological Solutions

In 2005, the Finnish National Board of Education (FNBE) assessed the learning outcomes in the Mathematics subject of pupils who had completed their second grade. The testing took place at the beginning of third grade. Largely, the same pupils were tested again in 2008 at the beginning of their sixth grade, and again in 2012 at the end of ninth grade. In what follows, the grades are called 3, 6, and 9 though actually they measure the learning outcomes at grades 2, 5, and 9.

### 2.1 Sample

Originally, 5,864 students participated for the grade 3 test in 2005. Of these, 4,679 students (80%) were able to be followed up when they started their 6<sup>th</sup> grade in 2008. When the students were at the end of their 9<sup>th</sup> grade, altogether 3,502 students of the original study group (60%) were able to be followed up. Of these, 1,800 were boys and 1,702 girls. All the students were selected by using the stratified sampling of the comprehensive schools, with a representation of different instruction languages (Finnish/Swedish), provinces and municipal groups (Cities/Population density areas/Rural areas).

Between the tests on grade 3 and 6, the number of low performing pupils dropped out from the data was higher than that of high performing pupils. Phone calls to the principals showed that most of the very low level students were taken out of the normal study groups to their own, supportive or intensive study groups. Hence, the dropping out was systematic. Nevertheless, there were still quite many low achieving students in the dataset in the 6<sup>th</sup> grade testing whose characteristics at 3<sup>rd</sup> grade did not differ from those who dropped out. Between grades 6 and 9, the students' drop-out was merely random. The reason was that usually the students change their school between the grade 6 and 7—from primary school to the lower secondary schools. All individual students were not followed but only those schools were selected where most students continued their studies.

All in all, the final results are the most reliable when analyzing the change of the students with average or higher achievement level.

### 2.2 Test Instruments and Their Reliabilities

The mathematics tests comprise test items from three main areas arising from the Finnish National Core Curricula of grade 3 (FNBE 2004): (1) Numbers, Calculations, and Algebra, (2) Geometry, and (3) Data processing, Statistics, and Probability. Naturally, the items were very easy at the beginning of grade 3 and they gradually became more difficult. The equating and re-scaling of the test scores are handled at Section 2.4. The technical characteristics of the mathematics test are collected on Tables 1 and 2.

Table 1. Technical characteristics of test items

	Number of items			Maximum scores		
	Grade 3	Grade 6	Grade 9	Grade 3	Grade 6	Grade 9
Whole test	38	39	68 <sup>4</sup>	44	52	84 <sup>4</sup>
NCA <sup>1</sup>	22	21	36	24	28	40
GEO <sup>2</sup>	10	10	16	14	14	22
DSP <sup>3</sup>	6	8	7	6	10	9

1) Numbers, Calculations, and Algebra, 2) Geometry, 3) Data processing, Statistics, and Probability, 4) Includes also five items from the areas of Functions

Table 2. Reliabilities of the achievement tests

	reliability in the original datasets			reliability in the longitudinal datasets		
	Grade 3	Grade 6	Grade 9	Grade 3	Grade 6	Grade 9
	(n = 5 864)	(n = 5 560)	(n = 6 179)	(n = 3 502)	(n = 3 502)	(n = 3 502)
Whole test	0.86	0.85	0.94 <sup>4</sup>	0.89	0.86	0.94 <sup>4</sup>
NCA <sup>1</sup>	0.81	0.78	0.88	0.86	0.78	0.87
GEO <sup>2</sup>	0.67	0.66	0.83	0.70	0.73	0.82
DSP <sup>3</sup>	0.55	0.47	0.61	0.53	0.46	0.61

1) Numbers, Calculations, and Algebra, 2) Geometry, 3) Data processing, Statistics, and Probability, 4) Includes also five items from the areas of Functions

As a whole, the tests were accurate enough for reliable inferences in all years ( $\alpha = 0.86\text{--}0.94$ ). Also the sub-scores of Numbers, Calculations, and Algebra ( $\alpha = 0.78\text{--}0.87$ ) and Geometry ( $\alpha = 0.70\text{--}0.82$ ) were accurate enough. However, the sub-score of Data processing, Statistics, and Probability ( $\alpha = 0.46\text{--}0.64$ ) stayed too low for accurate inferences. The reason for the last was that there were two sets of linking items between grades 3 and 6 which were kept in the grade 6 test even though it was noticed that the items were too easy for the 6<sup>th</sup> graders. In four items out of 8, more than 90% of the 6<sup>th</sup> graders gave a correct answer and hence, the item discrimination of these items stayed low. This lowered the reliability remarkably. In what follows in Section 3, mainly the total score is reported.

The attitude scale used in the different datasets is a modified version of Fennema-Sherman Mathematics Attitude Scales (Fennema & Sherman, 1976; Metsämuuronen 2012a). A shortened version of the original test with nine dimensions is used in several international comparisons, like in Trends in International Mathematics and Science Study 2007 (TIMSS, Mullis, Martin, & Foy, *et al.*, 2008) or 2011 (see <http://timssandpirls.bc.edu/data-release-2011/pdf/Overview-TIMSS-and-PIRLS-2011-Achievement.pdf>) and its predecessors 1995, 1999, and 2003 as well as in Programme for International Student Assessment (PISA). The shortened structure of the Fennema-Sherman scales includes three dimensions with four items in each and two negative items in each of the first two dimensions. The names of the factors can be “Liking Math”, “Self-efficacy in Math”, and “Experiencing utility in Math” (compare naming in, e.g., Kadujevich, 2006; 2008).

In the Finnish national achievement testing, a modified Fennema-Sherman test, with the same dimensions as in the international settings, has been used in numerous assessment questionnaires in several subjects (e.g., in Mathematics, Mother tongue, Science, Languages, Arts, and Physical education tests) in different grades (grades 4, 6, 7, and 9). The original Fennema-Sherman test has been amended by using the following principles: 1) to include less negative items (just one for each dimension), 2) to include simpler wordings, 3) to focus—not in “mathematics” but – more concrete “mathematics lessons” and “mathematic as a school subject”, 4) to omit the third dimension, utility when assessing the pupils of low grades because it includes the irrelevant questions of their work life or further studies, and 5) to use the 5-point Likert scale (-2, -1, 0, +1, +2) instead of the 4-point Likert scale without value 0 (scale is actually -2, -1, +1, +2) as used in the TIMSS and PISA studies. Though the dimensions are the same, the item-wise changes are so radical that the Finnish test is no more Fennema-Sherman test but rather “loosely based on Fennema-Sherman test” as described by Metsämuuronen (2009, p. 20). Further, Metsämuuronen (2012a; 2012b) argues, on the basis of extensive study with TIMSS dataset, that the Finnish version would be more suitable in the international settings because it is evident that the original test is not optimal with the lowest achieving students (Metsämuuronen 2012a) and in many Asian countries (Metsämuuronen, 2012b).

Four things are noteworthy of the attitude scales. (1) At the third grade, the dimension of feeling Utility in Mathematics was not used. Hence only two dimensions were used. All three dimensions were used at the grades 6 and 9. (2) At the original standard version in Finland, there are 5 items on each dimension. All these were used in the testing at the grades 6 and 9. At the third grade, only four items per dimension were used. (3) The wordings of the items were changed slightly to fit the third graders level (see the comparison at Table 3). All in all, the changes in wordings are small. Because of these reasons, all the attitude scores were changed into percentages of maximum score. Hence, as the most positive case, the student would get 100 which is strictly

100% of the maximum score. As the most negative case, the students would get zero which corresponds with 0% of the maximum score.

The classical item discriminations are high (Table 3). The reliabilities of the attitude scores are high enough for accurate inferences ( $\alpha = 0.79\text{--}0.92$ ) (Table 4). Mainly in what follows in Section 3, the attitude as a whole is reported.

Table 3. Item discrimination of the parallel items in the longitudinal dataset (N = 3,502)

Items used at the grade 3	Item Discrimination (Corrected item-total correlation)			Items used at the grades 6 and 9
	G3	G6	G9	
1) Mathematics is easy	0.65	0.66	0.79	1) Mathematics is an easy subject
2) I like Mathematics lessons	0.75	0.71	0.68	2) I like Mathematics lessons
3) I'm good in Mathematics tasks	0.58	0.64	0.76	3) I think I'm good in Mathematics
4) I can manage even the difficult tasks in Mathematics	0.54	0.52	0.68	4) I can manage even the difficult tasks in Mathematics
5) <sup>1</sup> Mathematics is boring	0.63	0.61	0.65	5) <sup>1</sup> Mathematics is a boring subject
6) <sup>1</sup> Many things in Mathematics lessons are difficult	0.36	0.47	0.59	6) <sup>1</sup> Many things in Mathematics are difficult
7) I like to learn Mathematics	0.61	0.73	0.80	7) I like to study Mathematics
8) Mathematics is one of my favorite subjects	0.69	0.73	0.77	8) Mathematics is one of my favorite subjects

<sup>1</sup> Items are inversed before summing up

Table 4. Reliabilities of the scores of Attitude scales in the follow-up datasets (N = 3,502)

	$\alpha$ reliability		
	Grade 3	Grade 6	Grade 9
Attitude as a whole (Liking Math + Self-Efficacy)	0.86	0.88	0.92
Liking Math	0.88	0.89	0.90
Self-Efficacy	0.79	0.82	0.88

### 2.3 Change Score Reliability

Usually the reliability of the change score is measured by using Intra-Class Correlation (ICC, Shrout & Fleiss, 1979). In the case of achievement scores, the *Two-way random effects ANOVA*, Model 2 in Shrout & Fleiss (1979,) is assumed. Then it is interpreted that the measurement points (that is, “Judges”) are independent and random—many other kinds of tests could have been done. The attitude tests are interpreted as *One-way random effects ANOVA*, Model 1 in Shrout and Fleiss (1979), because each student has assessed him-/herself three times. The values of ICCs are collected in Tables 5 and 6.

Table 5. Change score stability and reliability in the follow-up dataset (N = 3,502)

	Intra-class correlation (ICC)	95% Confidence Interval		$\alpha$ reliability
		lower	higher	
Achievement as a whole	0.61	0.59	0.63	0.83
Attitudes as a whole <sup>1</sup>	0.35	0.33	0.37	0.62

1) Only the components of Self-efficacy and Liking the subject

Table 6. Change score stability and reliability between the grades (N = 3,502)

	Intra-class correlation (ICC)	$\alpha$ reliability
Achievement as a whole 3 – 6		
6 – 9	0.75	0.86
3 – 9	0.56	0.72
Attitudes as a whole <sup>1</sup> 3 – 6	0.30	0.54
6 – 9	0.40	0.62
3 – 9	0.05	0.37

1) Only the components of Self-efficacy and Liking the subject

If the results would have been that the stability was  $ICC = 1$ , it would have meant that the achievement and attitudes did not change at all during the years. On the basis of ICC, it seems then obvious that the attitudes are changing more ( $ICC = 0.05-0.40$ ) than the achievement ( $ICC = 0.56-0.75$ ). It is noteworthy that the attitudes change radically between grades 3 and 9 ( $ICC = 0.05$ ), whereas, at the same time frame, the achievement changes only mildly ( $ICC = 0.56$ ) though clearly. It is also noteworthy that the achievement seems to change more between grades 3 and 6 ( $ICC = 0.58$ ) than between 6 and 9 ( $ICC = 0.75$ ) even though there were four years between the latter frame and three years between the former. This tells that the value added given by the schools might be higher at the lower grades than at the higher grades.

#### 2.4 Equating of the Test Scores

Before it is meaningful to compare the scores of the tests of different grades, the scores should be calibrated into the same scale, that is, the scores should be equated. The final tests were constructed so that a certain amount of identical items, representing different content areas, linked the tests to each other. Thus, it was possible to equate the test scores with Item Response Theory (IRT) modelling (Rasch, 1960; Lord & Novick, 1968; Hambleton, 1993; of equating, see Béguin, 2000) and to acquire the comparable latent ability of each student over the different versions. IRT modelling is the very tool for equate test scores in the well-known international comparisons of PISA and TIMSS studies, too. The estimation was done by using one parametric logistic model (that is, Rasch model) with OPLM software (Verhelst, Glas, & Verstralen, 1995). This means that only the difficulty parameter of the items was calibrated into the same metric.

Equating the test scores with IRT modelling was administered with the following principles and practices. The scores are transformed into the same scale on the basis of characteristics of IRT models that the latent ability level of a learner ( $\theta$ ) and difficulty level of an item ( $\beta$ ) are identical when certain preconditions are met (see Wright, 1968). The latent ability level for each pupil can be determined in the same metric for every test as far as there are the linking items connecting the versions. A brief technical description of the equation process is as follows (see more exhaustively in Béguin, 2000, 17–36):

- 1) Define the structure of the test so that the linking items are connecting the tests into each other. Because the values of difficulty parameter of the linking items are exactly the same in each version the difficulty levels of all other items are calibrated into the same scale as the linking items are.
- 2) Use *Conditional Maximum Likelihood* (CML) procedure to estimate the difficulty level ( $\beta$  parameter) for each item.

- 3) Use *Marginal Maximum Likelihood* (MML) procedure to estimate the distribution of each student's latent ability ( $\theta$  parameter) in each version.
- 4) Estimate the  $\theta$  parameter of the scores of each version using means and deviations of distributions of  $\beta$  and  $\theta$ . This results in a unique latent value, however measured in a common scale, for each observed value of the scores in all versions.

In the case, the equating of the scores could have been done in several ways depending on, first, which dataset is selected as the reference dataset and, second, how the difficulty parameters are fixed. In all cases, the differences between the years stay the same. The difference of the output comes from the decision of which will be the "average" person in the dataset. Now, the equating was done so that the original test of grade 9 is kept as the reference dataset and all the other datasets are calibrated into that scale. The average student in grade 9 is the reference point: (s)he will get the value 0 in the original metrics. When the students' achievement levels are higher than this "average" student, their value will be greater than 0 and, parallel, if their levels are lower than the "average", their value will be less than zero. Technically speaking, the item difficulty parameters were first estimated freely in the dataset including all grade 9 students. Then the item parameters were fixed and the 3<sup>rd</sup> and 6<sup>th</sup> graders datasets were taken into the estimation. The original metrics was further transformed into the PISA and TIMSS scale by using 10xT transformation. Hence, the original 0 is transformed into 500 and the Standard Deviation of the total distribution is 100.

### 2.5 Zero Level Estimation

The follow-up study started when the pupils were finished their second grade. In order to get a full picture of the change in the achievement, a simulation data was prepared to model the achievement in the zero class—at the beginning of the school. Three experts of the Mathematics in Early Childhood Education were asked their subjective opinion on how many percent of the pupils at the beginning of the first grade would be able to give a correct answer in each item of the third graders' test. The experts did not know the average percentage of correct answers in the tasks. No consensus was sought. However, two experts tend to give a conjoint answer to most items. In many cases, the mission was easy for the experts: except some random cases, none of the pupils would be able to solve the task, for example, related with multiplication. In some cases, the opinions of the experts vary mildly. The average percentage given by the experts was calculated for the final data processing.

The zero level dataset was constructed from the dataset of grade 3 on the basis of the following principles: (1) A pupil at the zero level does not give more correct answers than (s)he gave at grade 3. (2) If the pupil was not able to solve the task at grade 3, (s)he would not be able to solve it at the zero level. (3) The lower level students at grade 3 would not have managed the tasks at the zero level better than the higher level students. The last point means that when the correct answers were reduced, they were systematically taken from the lower level pupils. Hence, for example, if the experts indicated that not more than 5% of the zero level students would have been able to give a correct answer in a specific item, 95% of the lowest level students in the grade 3 dataset were given an incorrect answer and if any of the pupils in the remaining 5% made a mistake at the test of grade 3, it was not corrected to be a correct one.

### 2.6 Methods Used in the Analysis

Basic methods are used when describing and analyzing the differences between the years. The Repeated measures procedure is used to test the main differences between the years. The t-test is used when comparing the boys and girls and the related measure for effect size is Cohen's d (Cohen, 1988). The p-values are not adjusted. Proportions are tested in a traditional way (z-test) and the related measure for the effect size is Cohen's h.

Because the material is clustered – as the dataset sampled from schools usually is—the modeling of the predicted variables for the change is done by using Multilevel modeling known also as Hierarchical linear modeling (Goldstein 1986; Bryk & Raudenbush 1987; Raudenbush & Bryk 2002). Before the multilevel analysis, however, the basic principles of Analysis of covariance (ANCOVA) were applied. Namely, the grade 9 results were first explained by the grade 3 result and the remaining variance (residual) is then explained by other factors. Hence, the grade 3 result is used as a covariate in all the models. Multilevel modeling was done by using SPSS Linear Mixed Model procedure with Restricted Maximum Likelihood (REML) estimation.

Because of a large number of background variables, the data mining tool of SPSS software, Decision Tree Analysis (DTA) with CHAID algorithm (Kass, 1980) was used in the preliminary phase to identify such variables which may be valuable to take into closer consideration. Statistical analysis was done in SPSS20 environment.

All in all, the dataset is large and accurate enough to allow the credible inferences of the changes of achievement of students. The results are the most accurate when it comes to the measurement points of grades 3, 6, and 9 and the zero line estimation with simulated dataset give valuable information about the remaining part of the school years from zero to the end of grade 2.

### 3. Results

#### 3.1 Change in Achievement and Attitudes—A General View

The average achievement level of the pupils participating in the longitudinal study was 83 units at the zero level, 375 units at the beginning of third grade, 463 units at the beginning of sixth grade, and 502 units at the end of ninth grade. The detailed figures are condensed in Table 7 and illustrated in Figure 1. In all, during the school years learning increases by an average of 419 units, the sharpest change being observed during the lower grades and evening out towards the upper grades. Based on simulation of the achievement at the zero level, the greatest increase in learning seems to take place during the first two grades (Figure 1). Since most of the mathematical operations and their interconnections are learnt, in practice, at school, it can be concluded that school produces significant added value for the pupils.

Table 7. Descriptive statistics of the achievement scores

	Grade	mean (all)	min.	max.	Standard Deviation	Boys' means	Girls' means	Sig.	Cohen's d
Achievement as whole	0	83	-120	418	102,11	84	82	ns	0.03
	3	375	68	703	77,34	377	373	ns	0.05
	6	463	289	703	45,56	464	461	0.054	0.07
	9	502	324	746	50,68	505	498	< 0.001	0.14
Numbers, Calculations, and Algebra	0	-23	-60	431	88,17	-24	-23	ns	-0.01
	3	370	0	666	90,30	374	367	0.018	0.08
	6	457	228	673	52,56	460	454	< 0.001	0.12
	9	502	256	710	53,00	505	498	< 0.001	0.15
Geometry	0	235	38	443	95,08	236	234	ns	0.03
	3	409	38	575	97,27	408	411	ns	-0.02
	6	472	294	579	57,87	471	474	ns	-0.05
	9	502	331	664	63,23	503	501	ns	0.04
Data processing, Statistics, and Probability	0	40	34	340	30,12	40	40	ns	0.00
	3	344	34	481	101,15	343	345	ns	-0.01
	6	468	206	573	59,26	470	467	ns	0.06
	9	501	184	647	69,28	507	495	< 0.001	0.18



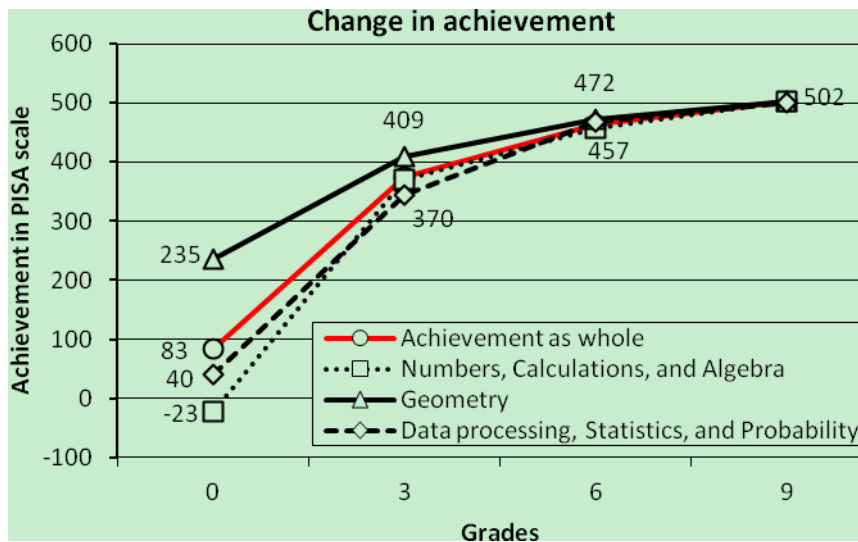


Figure 1. Change in achievement during the school years

The statistical analysis of the total scores shows the obvious fact that the differences between the means of the years are statistical significant (Tables 8 and 9). All the indicators used in the standard Repeated measures procedure for the effect, Pillai's trace, Wilks Lambda, Hotellings Trace, and Roys Largest Root indicate the difference in the means of different years ( $p < 0.001$ ) and between boys and girls ( $p = 0.004$ ) (Table 8). However, Mauchly's W ( $W = 0.243, p < 0.001$ ) indicates the lack of sphericity in the data and hence, the p-values are corrected by using Greenhouse-Geisser correction factor (GG). After correction, the main result remains: the differences in means are statistically significant ( $p < 0.001$ ). After using GG, however, there seems to be no difference between sexes ( $p = 0.187$ ) (Table 9). The deeper analysis of differences between the boys and girls (Table 10) shows, nevertheless, that there actually is difference at grade 9 ( $p < 0.001$ ). The effect size is, though, very low ( $d = 0.18$ ); in a graph, one would barely see any difference.

Table 8. Basic statistics of repeated measures

Multivariate Tests <sup>a</sup>							
Effect		Value	F	Hypothesis	dfError	df	Sig.
Achievement	Pillai's Trace	,970	38023,073 <sup>b</sup>	3,000	3498,000		,000
	Wilks' Lambda	,030	38023,073 <sup>b</sup>	3,000	3498,000		,000
	Hotelling's Trace	32,610	38023,073 <sup>b</sup>	3,000	3498,000		,000
	Roy's Largest Root	32,610	38023,073 <sup>b</sup>	3,000	3498,000		,000
Achievement * sex	Pillai's Trace	,004	4,376 <sup>b</sup>	3,000	3498,000		,004
	Wilks' Lambda	,996	4,376 <sup>b</sup>	3,000	3498,000		,004
	Hotelling's Trace	,004	4,376 <sup>b</sup>	3,000	3498,000		,004
	Roy's Largest Root	,004	4,376 <sup>b</sup>	3,000	3498,000		,004

a. Design: Intercept + sex

Within Subjects Design: Achievement

b. Exact statistic

Table 9. Greenhouse-Geisser corrected tests of within-subject effects

Tests of Within-Subjects Effects						
Measure: MEASURE_1						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Achievement	GG	375972856,277	1,583	237483858,640	58023,667	,000
Achievement * sex	GG	11080,179	1,583	6998,813	1,710	,187
Error(Achievement)	GG	22678763,174	5541,029	4092,879		

Table 10. Linear contrast of sexes between different grades

Tests of Within-Subjects Contrasts						
Measure: MEASURE_1						
Source	Achievement	Type III Sum of Squares	df	Mean Square	F	Sig.
	G0 vs. G3	977,307	1	977,307	,354	,552
Achievement * sex	G3 vs. G6	364,290	1	364,290	,108	,742
	G6 vs. G9	14664,030	1	14664,030	12,765	,000

One thing may be worth emphasizing in Table 7 and Figure 1: It seems that the tasks related to geometry are much easier to the zero level pupils (mean score 235) than, for example, Numbers, Calculations, and Algebra (mean score -23). The reason for this is that even though the academic studies are not given before the first grade in Finland—and hence the pupils usually do not calculate before coming to school – they are, though, taught the basic shapes which are elementary in learning geometry. Their ability of recognizing such basic shapes as the triangle, circle, and square were asked in the test at the grade 3.

Another note, related to the previous, is that the real value added for the children seems to come to the area of Numbers, Calculations, and Algebra, at least in the Finnish system. Compared with some other educational systems, where the children start their academic education very early, in the Finnish system, the parents are not encouraged to teach mathematical operations before starting the school (that is, age 7). The idea is to teach all the children conjointly the elementary mathematical operations in the school in such a way which allows later the proper understanding about multiplication, division, and proportions.

One may also infer from Table 7 that the differences between boys and girls are very mild in the early school years. From the sixth grade on, the achievement level of girls seems to get somewhat lower than that of boys. At the ninth grade, the differences are statistically significant (all  $p < 0.001$ ) in all areas except Geometry though the effect sizes remain small (Cohen's  $d$  ranges  $d = 0.14-0.18$ ). These differences are discussed in detail in Section 3.3.

One may ask whether the change of this magnitude in achievement is high or low. According to Metsämuuronen's (2013) longitudinal study of productivity in the Finnish schools on the basis of change in achievement in Mother tongue at the grades 7 to 9, an approximate increase of +5 percent units (of the maximum score) in achievement per year would show productivity—the schools would produce more achievement than they spend money. When changing the equated scores to percentages of maximum score instead of the PISA- or TIMSS scale, the scores are as follows: 51.4% at grade nine, 35.7% at grade six, 16.1% at grade three, and 0.1% at grade zero. Hence, the difference between grades 0 and 3 (that is, till the end of grade 2) is 16 percent units which equal 8 percent units per year. The difference between grades 3 and 6 is 19.5 percent units which equal 6.5 percent units per year. The difference between the beginning of grade 6 and end of grade 9 is 15.8 percent units which equal 3.9 percent units per year. Hence, roughly estimated by using Metsämuuronen's (2013) 5 percent units as a measurement stick, it seems that the education in Finland is *very* productive at the first two years, *quite* productive between grade 3 and 6 and *unproductive* at grades 6 to 9. It is, though, good to keep in

mind that the number of allocated teaching hours in Mathematics in Finland is higher in the lower grades than at the higher grades and hence, more rigorous study is needed to confirm the rough figures presented here.

The attitude, as whole, declines from 70% of the maximum score to 52% from the beginning of grade 3 till the end of grade 9. On the basis of rough logistic modeling on the basis of the averages in grades 3 to 9, the attitude level at the zero grade may be as high as 83% of the maximum (Figure 2). Hence the decline may be somewhat 30 percent units during all the school years. The decline seems to be the greatest at the area of Liking Mathematics: somewhat 50 percent units. The Finnish children seem to be quite aware of their level in mathematics—the items on the score of Self-efficacy quite strictly reflect the reality the children face in the classroom where they are, at quite early age, compared with the peer pupils. This may explain why the decline in this area is milder than in the other areas.

The result of declining attitudes was expected because of the previous results (Metsämuuronen, 2006a; 2006b; 2006c; 2010; Metsämuuronen, Svedlin & Ilic, 2012). Compared with the results on Metsämuuronen, Svedlin and Ilic (2012), the decline in the data, however, seems to be quite mild.

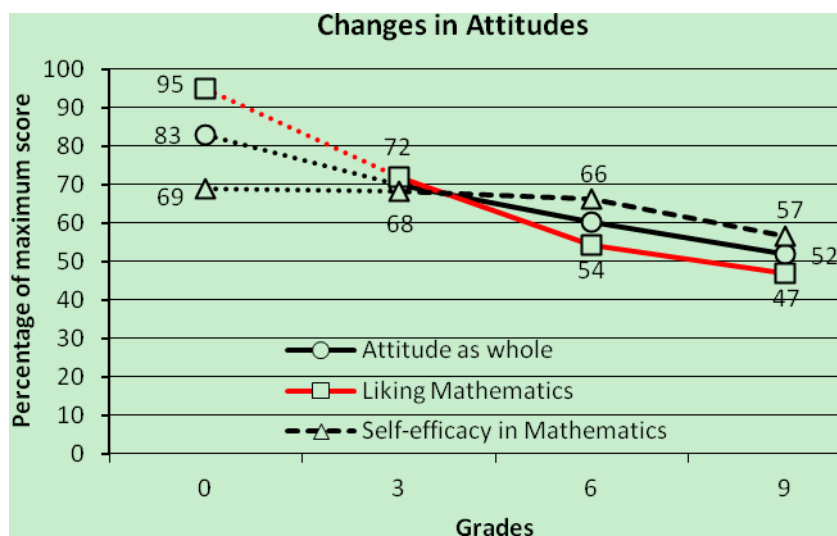


Figure 2. Change in Attitudes toward Mathematics during the school years

### 3.2 Profiles of the Change

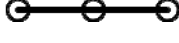
On the basis of three last measurement points, one finds nine different profiles of change (Table 11). The profiles were formed by using the knowledge of whether the achievement or attitude between two measurement points increase, decline, or remain the same. The increasing, declining, and remaining were defined by the effect size as an indicator; when the change showed lower than the boundary for “low” effect size (Cohen’s  $d \leq 0.25$ ), the change was not great enough to be said “change”. When, on the other hand, the change exceeded this boundary—in either direction—the change was called “change”. On Table 11, this “change” is indicated either +1 (increasing), 0 (remaining), and -1 (declining).

Table 11 tells that the main profiles in the Finnish dataset are “Increasing continuously” (57% of the students) and “Increasing evening out” (27%). Of the group of “Increasing continuously”, a specific subgroup was extracted, “increasing strongly”. For this group, the effect size was set to indicate high effect size in both measurement points, that is, Cohen’s  $d$  should exceed  $d \geq 0.80$ . Altogether 3.7% of the students belonged to this group. When it comes to the attitudes, most frequent profiles are those where, at least some kind of, a decline trend is seen. Most cases fell into category “Declining continuously” (19.5%).

One note may be worth making of Table 11. The test scores are equated so that the scores are at the same scale and hence they are comparable. If the achievement level was not increasing between three measurement points, that is, the student was remaining at the same level from grade 3 to grade 9, it actually means that the student

was regressing. Namely, it is more or less a normal phenomenon to see a natural development when the children grow up. The students with no change did not show that kind of development. In these rare cases (1.9%) the students either (1) were not willing to show their best performance, (2) were extremely good at the first measurement(s) but their enthusiasm or skills, or both, declined during the years, or (3) they just did not develop the normal way during the years.

Table 11. Profiles of changes and their frequencies in the longitudinal dataset (n = 3502)

Profile	Description	Change 3-6 <sup>1</sup>	Change 6-9 <sup>1</sup>	Frequency (%)	
				Achievement as whole	Attitudes as whole
	Increasing continuously	+1	+1	56,6	2,8
	Increasing later on	0	+1	8,2	3,9
	Recovering later on	-1	+1	2,4	15,6
	Increasing evening out	+1	0	27,4	4,8
	not increasing = regressing	0	0	1,9	8,2
	Declining evening out	-1	0	0,4	16,4
	Regressing later on	+1	-1	2,9	12,2
	Declining later on	0	-1	0,2	16,7
	Declining continuously	-1	-1	0	19,5

<sup>1</sup> 0 = No change, +1 = Change in a positive direction = increasing, -1 = Change in a negative direction = decreasing

### 3.3 Distributions of Achievement

The distributions of the values of latent ability (Theta) of the student populations in different years tell an interesting story of the educational system in Finland (Figure 3). On the basis of the distribution of the zero line, the student population entering the school is very heterogeneous. On the basis of the simulation dataset, there seems to be four populations: (1) the students who are totally ignorant of any mathematical concepts (or who still at the third grade cannot read the test papers) and hence, who would have gained zero score, (2) the students who may have some ideas of geometrical shapes but hence would gain only one point in the test, (3) the majority of the students who have gained some points from the geometric items but none from the other areas, and (4) few students who would have gained quite well even in the grade 3 test.

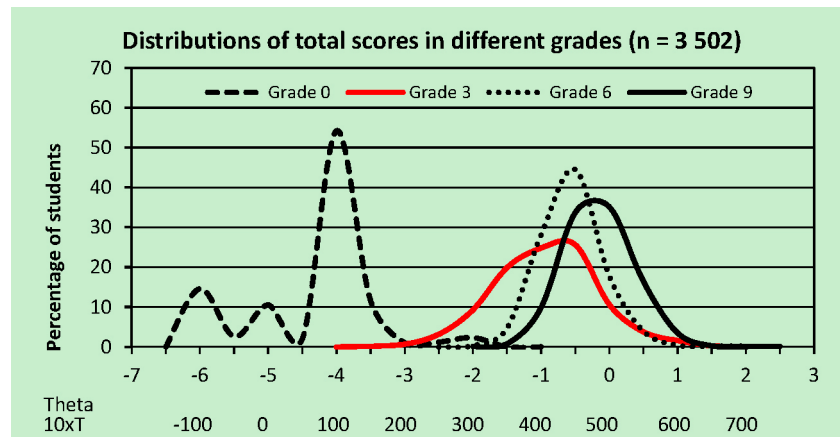


Figure 3. Distributions of total score

One interesting finding can be made on the basis of comparing the distributions of grades zero and three. The distribution of grade 3 is obviously not Normal. However, compared with the zero level distribution, within two years, the wide differences between the students have almost disappeared. The classroom actions during the first two years—mainly with the specialized teacher for the early childhood education in Finland—seem to bring a great value added especially for those students with low performance at the beginning. More specific test for the beginners at the zero level would reveal a more nuanced picture of the first graders' real achievement level.

At the grade 3, the best of pupils seems to be almost at the level of average pupils at grade 6 but there still is this other population with remarkably lower results and hence, the distribution is quite wide. At grade 6, the distribution is nicely normal which may tell something of the (late) latent developmental phase of the children: the student population has not deviated too widely to those who are not very willing to attend school and those who take it seriously. This kind of division of students may explain the wide shape of the grade 9 students' quasi-normal distribution.

### 3.4 Longitudinal Effect to Girls' Dropping out from the Best Sequence of Students

Somewhat sad and alarming a result, from the viewpoint of the Finnish educational system, is that the girls are dropping out from the best performing students during the higher grades (Figure 4). Especially, at the ninth grade the number of the girls is, in the midst of the students in the highest quintile, statistically lower than that of boys ( $z = -4.42, p < 0.001$ ) and even less within the students in the highest decile ( $z = -4.69, p < 0.001$ ). The effect sizes are, though, moderate (for the highest quintile Cohen's  $h = 0.34$  and for the highest decile  $h = 0.51$ ). Though there might be some good reasons for this phenomenon—for example, that the girls are optimizing their high marks at the school leaving certificate by concentrating on languages—it, in any case, reduces girls' opportunities to get a study place in the areas where the extremely good mathematical skills are needed.

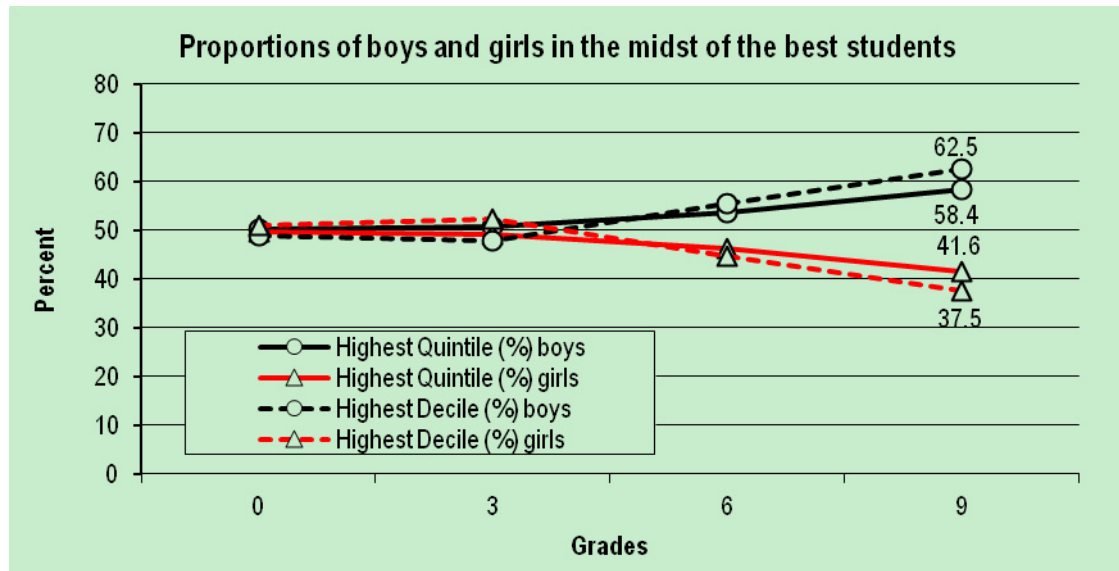
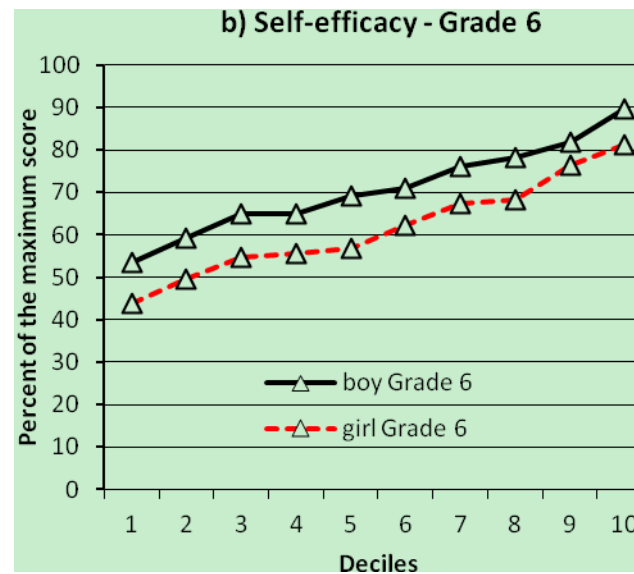
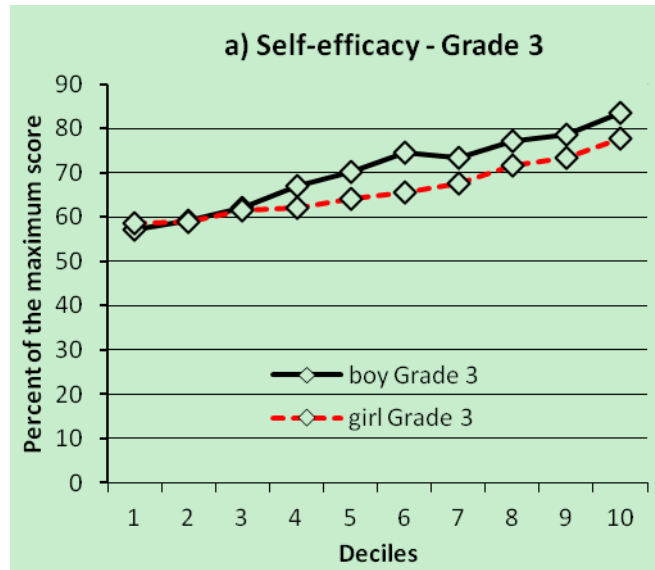


Figure 4. Proportions of boys and girls in the group of best students

A possibly related phenomenon is seen in the area of attitudes toward mathematics. During all the grades and within almost all the deciles, boys feel statistically significantly more self-efficacy in mathematics than the girls (Figures 4a-4c.). The phenomenon is obvious though the effect sizes at grade 3 in all deciles remain lower than  $d = 0.47$ , at grade 6 lower than  $d = 0.68$ , and at grade 9 lower than  $d = 0.70$ . Hence, during the years, girls seem to feel themselves less capable than boys in mathematics even though their actual capacity is equal with the boys. Williams and Williams (2010) observed, on the basis of 2003 PISA results (OECD, 2004), that the difference between boys' and girls' self-efficacy in mathematics was one of the widest in Finland. Tuohilampi and Hannula (2013) compared this with the results of PISA 2009 (OECD 2010a) and observed that the difference was reduced to be below the average. They remind us though, on the basis of Else-Quest, Hyde and Linn (2010), that the better the equality between the genders is in the society the less there are differences between the achievements but the more there seems to be differences in attitudes toward mathematics.

Though the causal connection of achievement and attitude is not clear (see Leder, 2006 and compare Pajares & Miller, 1994 and Mägi et al., 2010), Tuohilampi and Hannula (2013) showed that the achievement predicts better the attitude than other way round.



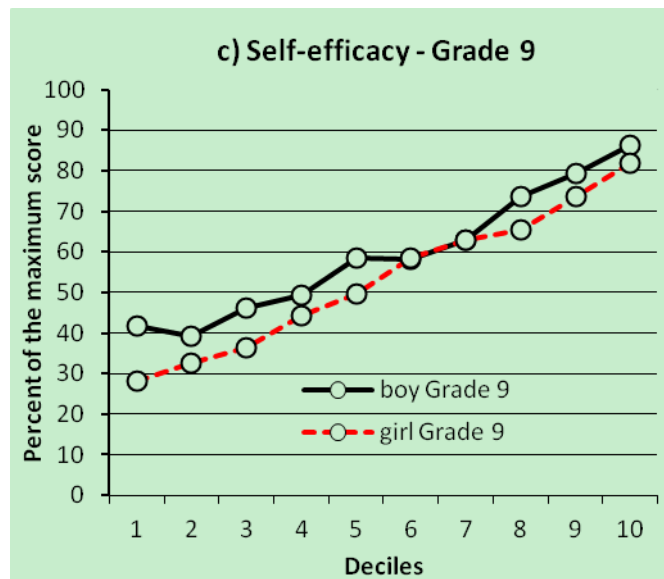


Figure 5a-5c. Boys' and girl's self-efficacy in mathematics in different deciles and grades

### 3.5 Selected Variables Explaining the Change—A Classical Approach

The dataset of three different grades combined with background questionnaires for students, teachers, and head teachers, as well as demographic information related with the school, provided us with a rich—maybe too rich—dataset to be analyzed effectively. On the other hand, it gives us possibilities to find some relevant reasons why some students gained more than the others. Because of the huge dataset (over 900 variables), it is possible to reveal just some connections in the dataset here. To simplify the analysis, only the residual of the ANCOVA, after explaining the grade 9 results with grade 3 results, is used as a dependent variable. The correlation between these two measurement points is  $r = 0.612$  and hence  $R^2 = 0.374$  (Figure 5). Practically speaking, there are students whose achievement increased more than what was expected on the basis of their achievement at grade 3 and, on the other hand, there are students whose achievement level increased radically less than what was expected on the basis of their grade 3 results. The latter analysis tries to explain which factors may explain this discrepancy.



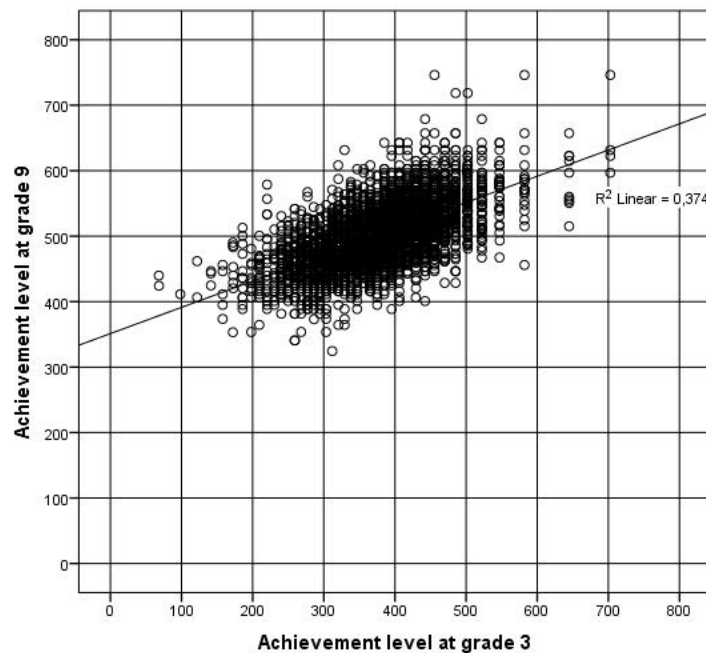


Figure 6. Basic design of ANCOVA

The dataset consists of more than 900 variables related to the results. To do any meaningful analysis in the dataset, a sketchy modeling of this complex phenomenon of learning is in use in FNBE (Metsämuuronen, 2009). The variables are divided into eight dimensions; the student factors, family factors, peer group factors, teacher factors, school factors divided into managerial factors and physical factors, economic factors and demographic factors. Of these, the economic factors are not handled here. This rough systemic model was used to organize the variables. The data mining tool of SPSS software, Decision Tree Analysis, DTA, was used in the preliminary phase to identify such variables which may be valuable to take into closer consideration. Altogether 26 variables were found to be the most prominent factors explaining the results; they all explain the residual remarkably when using them as individual fixed factors. These factors are collected in Table 12 in a rough way—more detailed information is shown on Tables 13 and 14.

After finding these interesting variables from each of the eight (or seven) dimensions of background factors, the multilevel modeling was administered to analyze them all in once. The educational provider was selected as the basis of the upper level hierarchy. The school level was not meaningful because the students had changed their schools in the midst of the process. The instruction language in the school explained the change in achievement remarkably; in the Swedish-speaking schools, the change was much greater than in the Finnish-speaking schools. Also, the explaining factors were somewhat radically different than those in the Finnish-speaking schools. Out of 26 interesting variables nine showed the independent main effect within the Finnish-speaking schools (Table 13) and ten within the Swedish-speaking schools (Table 14).

Table 12. Most prominent individual variables explaining the change in the dataset

Variable <sup>1</sup>	Difference <sup>2</sup>	Degree of Determination $\eta^2$	Effect size Cohen's f
<b>Student factors</b>			
School contentment at the grades 6 and 9	+39.1	3.0 %	0.18
Activity in home works during the grades 7–9	+38.2	2.0 %	0.13
Achievement level in Math at the grade 3 assessed by the teacher	+25.6	3.7 %	0.20
Readiness to learn at the grade 3 (a diagnostic test within the grade 3 test)	+22.3	2.0 %	0.14
Absence from school during the grade 9	+21.8	0.4 %	0.06
Achievement level in Mother tongue at the grade 3 assessed by the teacher	+13.3	1.0 %	0.10
Sex	+5.4	0.5 %	0.07
<b>Home factors</b>			
Home language	+31.5	1.8 %	0.14
Appreciation of education in school	+23.0	3.7 %	0.20
Parents' education by passing in the Matriculation examination	+20.8	4.5 %	0.21
<b>Peer group factors</b>			
Bullying in the school	+18.6	0.3 %	0.05
Peacefulness in the classroom	+13.0	1.2 %	0.11
<b>Teacher- and teaching factors</b>			
Given supporting teaching at the grades 6 and 9	+55.2	12.2 %	0.37
Given supporting- and special education at the grades 6 and 9 (together)	+52.2	13.6 %	0.40
Given special education at the grades 6 and 9	+44.9	9.2 %	0.32
Length of education in Mother tongue of the teacher at the grades 1 and 2	+23.3	2.4 %	0.16
Classroom activities: each solves tasks suitable for their achievement level	+22.4	3.1 %	0.18
Classroom activities: Students help each other	+21.3	1.9 %	0.14
Classroom activities: Applying the mathematics in the everyday life situations	+19.8	2.2 %	0.15
Classroom activities: common teaching for all by the teacher	+17.8	2.7 %	0.17
<b>School-related- and demographic factors</b>			
Size of the teaching group at the grade 9	+36.1	1.3 %	0.11
Type of municipality (in the Swedish-speaking schools)	+18.2	4.3 %	0.21
Instruction language (Finnish/Swedish)	+14.7	1.6 %	0.13
Province (in the Swedish-speaking schools)	+11.6	2.3 %	0.15
Average level of achievement in the school at the beginning of lower secondary education	+11.7	1.1 %	0.11
Way of forming the study group (Fixed / Flexible)	+8.5	0.4 %	0.06

1) The variables are ordered on the basis of the Difference.

2) Difference from the expected value between the lowest and highest group after explaining the grade 9 results with grade 3 result (PISA scale).

Table 13. Statistically significant variables in multilevel modeling explaining the change in the Finnish-speaking schools

Variable as a fixed factor <sup>1</sup>	df <sub>1</sub>	df <sub>2</sub>	F	p
Constant				
Bullying in the school (0 = bullied frequently during G9, 1 = bullied every week at G9 or every now and then at G6 and G9, 2 = not bullied at G9 or G6) <sup>2</sup>	2	1699	45.13	<0.001
Average level of achievement in the school at the beginning of lower secondary education (Proportion of low achieving students at G6: 0 = < 10%, 1 = 10–18.2%, 2 = >18.2%)	2	1699	32.16	<0.001
Parents' education assess by passing in the Matriculation examination (MA) (0 = none passed MA, 1 = either passed MA, 2 = Both passed MA)	2	1699	10.36	<0.001
Size of the teaching group at grade 9 (0 = < 11, 1 = 11–15, 2 = 16–20, 3 = 21–25, 4 = > 25)	4	1699	7.61	<0.001
Classroom activities: Students help each other (0 = not at all, ..., 4 almost always)	4	1699	4.89	0.001
Way of forming the study group (0 = flexible, 1 = fixed)	1	1699	7.83	0.005
Peacefulness in the classroom (0 = never, 1 = in some classes, 2 = usually or always)	2	1699	5.34	0.005
Activity in home works during the grades 7–9 (0 = never, ..., 4 almost always)	4	1699	2.84	0.023
School contentment at grades 6 and 9 (0 = very poor, ..., 3 = very well)	4	1699	2.76	0.027

- 1) The variables are organized on the basis of their statistical significance; at the top, there are the variables with the lowest p-value.
- 2) The categories are formed on the basis of DTA

When the students' starting level and the homogenizing effect of the educational provider were taken into account, out of two Finnish-speaking pupils at the same achievement level at grade 3, the one will get statistically higher results at the grade 9

- who *was not bullied at the higher grades*,
- who *went to a lower secondary school where the number of low achieving students is below 10%*,
- whose *both parents have passed the Matriculation examination* (that is, they are more educated),
- who *studied in a somewhat large study group* (which means that (s)he did not need a teaching in a small study group),
- in whom *classes the students often helped each other*,
- who *studied in a study group formed flexible*,
- whose *class does not encounter problems in peacefulness in the classroom*,
- who *did diligently his/her home works*, and
- who *did not have problems in school contentment at grade 9*.

Table 14. Statistically significant variables in multilevel modeling explaining the change in the Swedish-speaking schools

Variable as a fixed factor <sup>1</sup>	df <sub>1</sub>	df <sub>2</sub>	F	Sig.
Constant	1	170	6.387	0.012
Province (0 = Province of South Finland 1 = Province of Western Finland) <sup>2</sup>	1	170	22.194	< 0.001
Way of forming the study group (0 = flexible, 1 = fixed)	1	170	17.876	< 0.001
Average level of achievement in the school at the beginning of lower secondary education (Proportion of low achieving students at G6: 0 = < 10%, 1 = 10–18.2%, 2 = >18.2%)	2	170	9.874	< 0.001
Type of the municipality (0= city, 1 = population center 2 = rural)	2	170	8.597	< 0.001
Size of the teaching group at the grade 9 (0 = < 11, 1 = 11–15, 2 = 16–20, 3 = 21–25, 4 = > 25)	4	170	6.796	< 0.001
Classroom activities: every one solves tasks suitable for their achievement level (0 = never, ..., 4 almost always)	4	170	4.743	0.001
Bullying in the school (0 = bullied frequently during G9, 1 = bullied every week at G9 or every now and then at G6 and G9, 2 = not bullied at G9 or G6)	1	170	10.041	0.002
Absence during the grade 9 (0 = 0–5 days, 1 = 6–10 days, 2 = 11–20 days, 3 = over 20 days)	3	170	3.799	0.011
Parents' education assessed by passing in the Matriculation examination (MA) (0 = none passed MA, 1 = either passed MA, 2 = Both passed MA)	2	170	4.558	0.012
Activity in home works during the grades 7–9 (0 = never do, ..., 4 almost always do)	4	170	3.062	0.018

- 1) The variables are organized on the basis of their statistical significance; at the top, there are the variables with the lowest p-value.
- 2) The categories are formed on the basis of DTA

When the students' starting level and the homogenizing effect of the educational provider are taken into account, out of two Swedish-speaking pupils at the same achievement level at grade 3, the one will get statistically higher results at grade 9

- who studied at the Province of South Finland,
- who studied in the flexible formed study groups in the city or population center,
- who studied in a somewhat large study group (which means that (s)he did not need a teaching in a small study group),
- in whose classes every one solves tasks suitable for their achievement level,
- who was not bullied at the higher grades,
- who was diligently at school and made the home works, and
- whose both parents have passed the Matriculation examination.

#### 4. Discussion and Questions Raising from the Results

On the basis on seven years' follow up of 3,502 students – from the end of grade 2 to the end of grade 9 – it seems evident that the sharpest change in achievement happens during the lower grades and it evens out towards the upper grades. Based on simulations, the greatest increase in learning seems to take place during the first two grades. Since, in practice, much of the mathematical operations and their interconnections are learnt at school, it can be concluded that school produces significant added value for the pupils. The achievement level of the student population entering the school is very heterogeneous. During the first two years, the actions seem, however, to cause the differences between the students almost totally disappear. Compared with Metsämuuronen's (2013) figures for productivity in the Finnish lower secondary schools, it seems that the education in Finland is *very* productive at first two grades, *quite* productive between grade 3 and 6 and *unproductive* between the grades 6 to 9.

An obvious question to ask is why the increase in the development of achievement slows down during the upper grades? Is it a law of nature that the adolescents just are not that interested in mathematics and hence the motivation slows down? At least students are more interested in learning at the beginning of school years than later on (Tuohilampi & Hannula, 2013), so it might be possible that the greatest development happens during these early school years. Could the interest be lower at the lower secondary school years because of less ambitious curriculum at the higher grades, causing the deteriorating development of achievement? Most probably this is not true. Maybe, contrarily, the *wide* and *ambitious* curriculum at the higher grades blocks the deep learning of the mathematical content, leading to more negative attitude and decreasing outcome results. In any case, it may be valuable to compare the possible differences of the curves in the high- and low performing countries. Finland can be taken as one of those high performing countries as discussed above. What kind of profiles may be found in some of the lowest level countries involved PISA or TIMSS studies? The comparisons may be easy to do especially in the countries with exhaustive examination culture.

A known phenomenon in the Finnish educational system is that the attitudes are declining during the years. In the dataset, the decline was the greatest at the area of Liking Mathematics (somewhat 50 percent units) and milder in the area of Self-efficacy in Mathematics (17 percent units). During all the grades and within almost all the deciles, boys feel statistically significantly more self-efficacy in mathematics than the girls. The reasons for this phenomenon are not discussed here but it may have an effect on girls to drop out from the highest segment of students at the highest grades seen clearly in the follow-up dataset.

First question to ask is the same as above: Is it normal to see the declining trend in attitudes toward mathematics while the pupils are growing older? The development can be seen at least to some extent natural, as the attitude is constructed based on social responses, including negative ones. After an almost omnipotent view of the self in the childhood that covers also an extreme interest of surrounding (Harter, 1999), a certain number of negative, significant responses contribute to a more realistic view. However, here the data shows that in some cases the trend could be positive. Is it realistic to think that the change in attitudes should, in a large scale, show an increasing trend? What about the girls: should we be worried of the girls' low self-efficacy in mathematics? If we should, what could or should be done? If the achievement explains better the attitudes than other way round, as Tuohilampi and Hannula (2013) found out, maybe the girls receive too negative feedback of their real achievement level, and more realistic, i.e. more positive feedback would be needed. On the other hand, girls tend to rate themselves in a more modest way than boys (Syzmanowicz & Furham, 2011), which may explain some of the difference. It is even possible that the girls, being more matured than boys, understand that one gets good marks much easier in some other area, such as in languages, and put their effort on that.

On the whole, learning increased more when the pupils were diligent in doing their homework at the upper grades; were not bullied at school; did not need to participate in special needs education or have remedial teaching needs; when their parents are more educated; and when education and mathematics as a subject are valued at home. These are not surprising factors. In any case, they make sense in increasing the achievement level of the students.

One may ask how local these kinds of explaining factors are. The factors above may be universal ones but, for example, the facts that the language group (Finnish/Swedish) or Province explain the results, may be very Finnish reason for differences. Are these factors important to rise as basic results? Would it be more important or interesting to go deeper in the broader dimensions with more specified theories from psychology, sociology, pedagogy, leadership and management, economy and so on? What should we do with these kinds of factors explaining the average increase in achievement? Should we change something in our practices?

The last question to ask is what can be learnt from the results? First thing to learn is that the most valuable work for the school children is done during the very first years—in Finland, within two first years, seems so. It is known that the variance between schools is very low in Finland (Schleicher 2006, 13). This means that where ever in the country the parents put their children at the school, it is expected that the end product will be the same. This homogenizing process seems to happen within the lower grades. Second, it seems evident that all the students have their own path to go—nine different profiles were identified and many more could be found. The real question is how to help the young ones at the beginning of school in an optimal way to find their own path to productive citizens for the society, bearing in mind that the future success is not always dependent of school performance. It is worth remembering that, in Finland and in many other countries too, the first grader pupil will spend at school somewhat 13 years after which (s)he enters the work life or further studies. What kinds of tools we are able to give those children to play with in the future world? This responsibility is given to schools and teachers specifically.

## References

- Aho, E., Pitkänen, K., & Sahlberg, P. (2006). *Policy Development and Reform Principles of Basic and Secondary Education in Finland since 1968*. May 2006. Washington, D.C., U.S.A.: The World bank.
- Alfonso, A., & St. Aybun, M. (2006). Cross-Country Efficiency of Secondary Education Provision: A Semi-Parametric Analysis with non-Discretionary Inputs. *Econometric Modelling* 23(3), 476-493. <http://dx.doi.org/10.1016/j.econmod.2006.02.003>
- Arinen, P., & Karjalainen, T. (2007). *PISA 2006 ensituloksia. 15-vuotiaiden koululaisten luonnontieteiden, matematiikan ja lukemisen osaamisesta*. [in Finnish] Opetusministeriö. Retrieved from <http://www.minedu.fi/export/sites/default/OPM/Julkaisut/2007/liitteet/opm38.pdf?lang=fi>. [in Finnish]
- Béguin, A. (2000). *Robustness of Equating High-Stake Tests*. Enschede: Febodruk B.V.
- Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 104, 147-158. <http://dx.doi.org/10.1037/0033-2909.101.1.147>
- Clements, B. (2002). How Efficient Is Education Spending in Europe? *European Review of Economics and Finance* 1(1), 3-27.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, 136(1), 103-127. <http://dx.doi.org/10.1037/a0018053>
- Fennema, E., & Sherman, J. A. (1976). Fennema-Sherman mathematics attitudes scales. *JSAS Catalog of Selected Documents in Psychology*, 6, 31.
- FNBE. (2004). *National Core Curriculum for Basic Education 2004*. Finnish National Board of Education. Retrieved from [http://www.oph.fi/english/publications/2009/national\\_core\\_curricula\\_for\\_basic\\_education](http://www.oph.fi/english/publications/2009/national_core_curricula_for_basic_education).
- Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 73, 223-232. <http://dx.doi.org/10.1093/biomet/73.1.43>
- Hambleton, R. K. (1993). Principles and selected Applications of Item Response Theory. In R. N. Linn (Ed.), *Educational Measurement* (3rd ed.). American Council of Education. Series of Higher Education. Oryx Press.
- Hargreaves, A. (2006). *Sustainable Leadership & Development in Education: Creating the Future, Conserving the Past*. Paper presented to the EU Presidency Conference on “Lifelong Learning: Equity and Efficiency”. Retrieved from [http://www.minedu.fi/export/sites/default/OPM/Tapahtumakalenteri/2006/09/eu\\_28\\_2909/Andy\\_Hargreaves\\_2.pdf](http://www.minedu.fi/export/sites/default/OPM/Tapahtumakalenteri/2006/09/eu_28_2909/Andy_Hargreaves_2.pdf)
- Harter, S. (1999). *The construction of the self. A developmental perspective*. New York, NY: The Guildford Press.
- Kadijevich, D. (2006). Developing trustworthy TIMSS background measures: A case study on mathematics attitude. *The Teaching of Mathematics*, 9(2), 41-51. Retrieved from <http://elib.mi.sanu.ac.yu/journals/tm/17/tm924.pdf>
- Kadijevich, D. (2008). TIMSS 2003: Relating Dimensions of Mathematics Attitude to Mathematics Achievement. *ZbornikInstitutaZapedagoskaIstrazivanja*, 40(2), 327-346. <http://dx.doi.org/10.2298/ZIPI0802327K>

- Kannas, L. (Ed.). (1995). *Koululaisten kokema terveys, hyvinvointi ja kouluviihtyvyys*. Helsinki: Opetushallitus. [In Finnish]
- Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2), 119-127. <http://dx.doi.org/10.2307/2986296>
- Laukkanen, R. (2008). Finnish Strategy for High-Level Education for All. In N.C. Soguel, & P. Jaccard (Eds.), *Governance and Performance of Educational Systems* (pp. 305-324). Springer.
- Laukkanen, R. (2013). Finland's experiences of compulsory education development. *ArtsEduca*, 5, 140-166.
- Lavonen, J., & Laaksonen, S. (2009). Context of Teaching and Learning School Science in Finland: Reflections on PISA 2006 Results. *Journal of Research in Science Teaching*, 46(8), 922-944. <http://dx.doi.org/10.1002/tea.20339>
- Leder, G. C. (2006). Affect and mathematics learning: Concluding comments. In J. Maasz, & W. Schloeglmann (Eds.), *New mathematics education and practice* (pp. 257-261). Sense Publishers: Netherlands.
- Leino, K. (2003). Computer Usage and Reading Literacy. In S. Lie, P. Linnakylä, & A. Rue (Eds.), *Northern Lights on PISA* (pp. 71-81). Department of Teacher Education and School Development, University of Oslo, Norway.
- Linnakylä, P. (1993). Miten oppilaat viihtyvät peruskoulun yläasteella? Kouluelämän laadun kansallinen ja kansainvälinen arviointi. In V. Brunell, & P. Kupari (Eds.), *Peruskoulu oppimisympäristönä—peruskoulun arviointi 90-tutkimuksen tuloksia*. Jyväskylä: Kasvatustieteiden tutkimuslaitos. [in Finnish]
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of Mental test Scores*. Reading, Mass.: Addison-Wesley Publishing Company.
- Metsämuuronen, J. (2006a). *Äidinkieli ja kirjallisuus -oppiaineen oppimistulosten ja asenteiden muuttuminen perusopetuksen ylempien luokkien aikana. Oppimistulosten arviointi 3/2006*. Opetushallitus. Helsinki: Yliopistopaino. [in Finnish]
- Metsämuuronen, J. (2006b). *Förändringar i kunskapsnivån i ämnet modersmål och litteratur under de högre årskurserna i den grundläggande utbildningen. Utvärdering av inlärningsresultat 4/2006*. Utbildningsstyrelsen. Helsinki: Yliopistopaino. [in Swedish]
- Metsämuuronen, J. (2006c). Oppimistulosten ja asenteiden muuttuminen perusopetuksen ylempien luokkien aikana. Äidinkieli ja kirjallisuus ja modersmål och litteratur -oppiaineiden näkökulma. Tekninen raportti. Oppimistulosten arviointi 5/2006. Opetushallitus. Helsinki: Yliopistopaino. [in Finnish]
- Metsämuuronen, J. (2009). Methods Assisting Assessment; Methodological solutions for the National Assessments and Follow-Ups in the Finnish National Board of Education. Oppimistulosten arviointi 1/2009. Opetushallitus. Helsinki: Yliopistopaino. [In Finnish.]
- Metsämuuronen, J. (2010). Osaamisen ja asenteiden muutos perusopetuksen 3.-5. luokilla. In E. K. Niemi, & J. Metsämuuronen (toim.), *Miten matematiikan taidot kehittyvät? Matematiikan oppimistulokset peruskoulun viidennen vuosiluokan jälkeen vuonna 2008* (pp. 93-136). Opetushallitus. [In Finnish]
- Metsämuuronen, J. (2012a). Challenges of the Fennema-Sherman Test in the International Comparisons. *International Journal of Psychological Studies*, 4(3), 1-22. <http://dx.doi.org/10.5539/ijps.v4n3p1>
- Metsämuuronen, J. (2012b). Comparison of Mental Structures of Eighth-Graders in Different Countries on the basis of Fennema-Sherman test. *International Journal of Psychological Studies*, 4(4), 1-17. <http://dx.doi.org/10.5539/ijps.v4n4p1>
- Metsämuuronen J (2013). Total factor productivity in lower secondary education—A Finnish perspective. *International Journal of Educational Research*. (Accepted)
- Metsämuuronen, J., Svedlin, R., & Ilic, J. (2012). Change in Pupils' and Students' Attitudes toward School as a Function of Age—A Finnish Perspective. *Journal of Educational and Developmental Psychology*, 2(2). <http://dx.doi.org/10.5539/jedp.v2n2p134>
- Metsämuuronen, J., Kuosa, T., & Laukkanen, R. (2013). Sustainable leadership and future-oriented decision making in the educational governance—A Finnish case. *International Journal of Educational Management*, 27(4). <http://dx.doi.org/10.1108/09513541311316331>
- Mullis, I. V. S., Martin, M. O., Foy, P., Olson, J. F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J. (2008).

- TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grade*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from <http://pirls.bc.edu/timss2007/mathreport.html>
- Mägi, K., Lerkkanen, M.-K., Poikkeus, A.-M., Rasku-Puttonen, H., & Kikas, E. (2010). Relations between achievement goal orientations and math achievement in primary grades: A follow-up study. *Scandinavian Journal of Educational Research*, 54(3), 295-312. <http://dx.doi.org/10.1080/00313831003764545>
- Niemi, H. (2010). Teachers as high level professionals—What does it mean in teacher education? Perspectives from the Finnish teacher education. In K. G. Karras, & C. C. Wolhuter (Eds.), *International Handbook of Teacher Education: Issues and Challenges* (Vol. I & II, pp. 237-254). Athens Greece: Atrapos.
- Niemi, H. (2011). Educating student teachers to become high quality professionals—A Finnish case. *Center for Educational Policy Studies Journal*, 1(1), 43-66.
- Niemi, H. (2012). The societal factors contributing to education and schooling in Finland. In H. Niemi, A. Toom, & A. Kallioniemi (Eds.), *The Miracle of Education: The Principles and Practices of Teaching and Learning in Finnish Schools* (pp. 19-38). Rotterdam: Sense Publishers. [http://dx.doi.org/10.1007/978-94-6091-811-7\\_2](http://dx.doi.org/10.1007/978-94-6091-811-7_2)
- Niemi, H., & Jakku-Sihvonen, R. (2006). Research-based teacher education in Finland. In R. Jakku-Sihvonen, & H. Niemi (Eds.), *Research-Based Teacher Education in Finland—Reflections by Finnish Teacher Educators* (pp. 31-51). Turku: Finnish Educational Research Association
- Niemi, H., & Jakku-Sihvonen, R. (2011). Teacher education in Finland. In M. Valenčič Zuljan, & J. Vogrinc (Eds.), *European Dimensions of Teacher Education: Similarities and Differences* (pp. 33-51). Slovenia: University of Ljubljana & The National School of Leadership in Education.
- Niemi, H., Toom, A., & Kallioniemi, A. (Eds.) (2012). *The Miracle of Education: The Principles and Practices of Teaching and Learning in Finnish Schools*. Rotterdam: Sense Publishers. <http://dx.doi.org/10.1007/978-94-6091-811-7>
- OECD. (2001). *Knowledge and Skills for Life*. First results from PISA 2000. Paris: OECD. <http://dx.doi.org/10.1787/9789264195905-en>
- OECD. (2003). *Education at a Glance*. OECD Indicators 2003. Paris: OECD.
- OECD. (2004). *Learning for tomorrow's world. First results from PISA 2003*. Paris: OECD Publishing. <http://dx.doi.org/10.1787/9789264006416-en>
- OECD. (2005). *Equity in Education*. Thematic review. Finland, Country note. Retrieved from <http://www.oecd.org/dataoecd/49/40/36376641.pdf>
- OECD. (2007). PISA 2006 results. Retrieved from [http://www.pisa.oecd.org/document/2/0,3343,en\\_32252351\\_32236191\\_39718850\\_1\\_1\\_1\\_1,00.html#ES](http://www.pisa.oecd.org/document/2/0,3343,en_32252351_32236191_39718850_1_1_1_1,00.html#ES)
- OECD. (2010a). *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science* (Vol. I). Paris: OECD. <http://dx.doi.org/10.1787/9789264091450-en>
- OECD. (2010b). *Finland: Slow and Steady Reform for Consistently High Results*. Retrieved from <http://www.oecd.org/dataoecd/34/44/46581035.pdf>
- Pajares, F., & Miller, M. 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. <http://dx.doi.org/10.1037/0022-0663.86.2.193>
- Raivola, R. (2006). How far can we learn anything practical from the study of foreign systems of education? Finland and the PISA model. Athens: *Comparative and International Educational Review*, 6, 11-23.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Pædagogiske Institut. Studies in Mathematic Psychology I. Copenhagen: Nielsen & Lydiche.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Application and Data Analysis Methods* (2nd ed.). (Advanced Quantitative Techniques in the Social Sciences Series.) Thousands Oaks: Sage Publications.
- Reinikainen P. (2012). Amazing PISA results in Finnish comprehensive schools. In H. Niemi, A. Toom, & A. Kallioniemi (Eds.), *The Miracle of Education: The Principles and Practices of Teaching and Learning in*



- Finnish Schools* (pp. 3-18). Rotterdam: Sense Publishers. [http://dx.doi.org/10.1007/978-94-6091-811-7\\_1](http://dx.doi.org/10.1007/978-94-6091-811-7_1)
- Saarinen, R. (2005). Otsa hiessä: luterilaisuuden vaikutus suomalaiseen ajatteluun. *Niin & näin*, 12(3), 79-84. [In Finnish]
- Sahlberg, P. (2006). Education Reform for Raising Economic Competitiveness. *Journal of Educational Change*, 7(4), 259-287. <http://dx.doi.org/10.1007/s10833-005-4884-6>
- Sahlberg, P. (2007). Education policies for raising student learning: The Finnish approach. *Journal of Education Policy*, 22(2), 147-171. <http://dx.doi.org/10.1080/02680930601158919>
- Sahlberg, P. (2011a). The Professional Educator: Lessons from Finland. *American Educator*, 35(2), 34-38.
- Sahlberg, P. (2011b). Lessons from Finland: Where the Country's Education System Rose to the Top in Just a Couple Decades. *Education Digest*, 77(3), 18-24.
- Schleicher, A. (2006). *The economics of knowledge: Why education is key for Europe's success*. Policy Brief. Brussels: Lisbon Council. Retrieved from <http://www.oecd.org/dataoecd/43/11/36278531.pdf>
- Schleicher, A. (2011). Is the Sky the Limit to Education Improvement? *Phi Delta Kappan*, 93(2), 58-63. <http://dx.doi.org/10.1177/0031721711109300213>
- SCP, Social and Cultural Planning Office. (2004). *Public sector Performance*. SCP-publication 2004/8. The Hague.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86(2), 420-428. <http://dx.doi.org/10.1037/0033-2909.86.2.420>
- Simola, H. (2005). The Finnish Miracle of PISA: Historical and Sociological Remarks on Teaching and Teacher Education. *Comparative Education*, 41(4), 455-470. <http://dx.doi.org/10.1080/03050060500317810>
- Sulkunen, S., Välijärvi, J., Arffman, I., Harju-Luukkainen, H., Kupari, P., & Nissinen, K. (2010). *PISA 2009 Ensituloksia*. [Initial Finnish Results of PISA2009, in Finnish].
- Sutherland, D., Price, R., Joumard, I., & Nicq, C. (2007). *Performance Indicators for Public Spending Efficiency in Primary and Secondary Education*. OECD Economics Department Working Papers 546.
- Syzmanowicz, A., & Furnham, A. (2011). Gender differences in self-estimates of general, mathematical, spatial and verbal intelligence: Four meta analyses. *Learning and Individual Differences*, 21(5), 493-504. <http://dx.doi.org/10.1016/j.lindif.2011.07.001>
- Tuohilampi, L., & Hannula, M. S. (2013). Matematiikkaan liittyvien asenteiden kehitys sekä asenteiden ja osaamisen välinen vuorovaikutus 3., 6. ja 9. luokalla. In J Metsämuuronen (Ed.), *Perusopetuksen matematiikan oppimistulosten pitkäjäsenarviointi vuosina 2005–2012*. Koulutuksen seurantaraportit 2013:4. Opetushallitus. Tampere: Juvenes Print-Suomen Yliopistopaino Oy. [In Finnish]
- Verhelst, N. G., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *One-Parameter Logistic Model OPLM*. Arnhem: Cito.
- Williams, T., & Williams, K. (2010). Self-efficacy and performance in mathematics: Reciprocal determinism in 33 nations. *Journal of Educational Psychology*, 102(2), 453-466. <http://dx.doi.org/10.1037/a0017271>
- Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference of Testing Problems*. Princeton, NJ: Educational Testing Service.
- Välijärvi, J. (2004). The System and How Does it Work—Some Curricular and Pedagogical Characteristics of the Finnish Comprehensive Schools. *Educational Journal*, 31(2) & 32(1), 31-55. Retrieved from [http://hkier.fed.cuhk.edu.hk/journal/wp-content/uploads/2009/10/ej\\_v31n2-v32n1\\_31-55.pdf](http://hkier.fed.cuhk.edu.hk/journal/wp-content/uploads/2009/10/ej_v31n2-v32n1_31-55.pdf)
- Välijärvi, J., Kupari, P., Linnakylä, P., Reinikainen, P., Sulkunen, S., Törnroos, J., & Arffman, I. (2007). *The Finnish success in PISA—And some reasons behind it. PISA 2003*. Jyväskylä: University of Jyväskylä.

### Copyrights

Copyright for this article is retained by the author, with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).