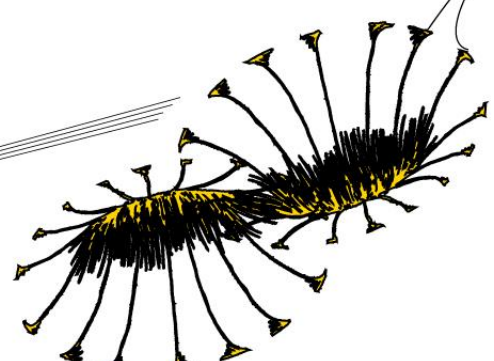





MASTER THESIS

## STRATEGIC BEHAVIOR BY STUDENTS

A Supplementary Explanation for Differences Between Marks on School Examinations and Central Examinations in Dutch Pre-University Education



**KEY WORDS:** CENTRAL EXAMINATIONS, SCHOOL EXAMINATIONS, STRATEGIC BEHAVIOR, PRE-UNIVERSITY EDUCATION

**Johan Leferink**  
April, 2015  
University of Twente  
Faculty of Behavioural, Management and Social sciences  
Educational Science and Technology

**Examination Committee**  
Dr. Hans J.W. Luyten  
Prof. Dr. Ir. Bernard P. Veldkamp



UNIVERSITY OF TWENTE.



## Table of contents.

Table of contents. ....	3
Summary. ....	4
Foreword. ....	5
Introduction and background.....	6
Possible causes of strategic behavior of students.....	7
Rules and regulations.....	10
Research questions and hypotheses.....	12
Research design.....	16
<i>Research method.</i> .....	16
<i>Respondents.</i> .....	16
<i>Instrumentation.</i> .....	16
<i>Data analyses.</i> .....	16
<i>Procedures.</i> .....	19
Results.....	20
Conclusions and discussion.....	32
Bibliography.....	35
Appendix A. ....	38

## **Summary.**

Discrepancies between students' marks on school examinations and central examinations in Dutch secondary educations can be attributed to strategic behavior of schools or can be attributed to strategic behavior of students. Strategic behavior of schools is researched in much literature. Strategic behavior of students as explanation has thus far hardly been considered as an alternative explanation. The difference between a school's average central examination mark for a subject, minus a school's average school examination mark for that subject, serves as an important indicator for assigning the quality of schools by the Dutch Inspectorate of Education.

A supplementary explanation that has not been considered in research so far is the possibility that discrepancies may also result from strategic behavior by students. Students with high grades on the school examinations may be less motivated to get high grades on the central examinations if the final grade is computed by averaging the mark on the school and central examinations.

A change in rules and regulations regarding graduating pre-university education provides an interesting opportunity to investigate if strategic behavior by students may account for discrepancies between marks on school examinations and central examinations. Since 2012, students need to get an average central examination mark of at least a 5.5. Therefore students with high marks on their school examinations can no longer afford to get to low marks on their central examinations and still need to perform on their central examinations. Before 2012 high grades on school examinations would largely compensate for low grades on central examinations (and vice versa). This research will compare differences between students' mark on their school examinations mark minus their central examinations mark in 2011 and 2012: before, and after this change in rules and regulations.

Data provided by the Dutch Inspectorate of Education were analyzed by both descriptive statistics and multi-level analyses. Multi-level analyses were used to discriminate in variation on school level and variation on student level. The data provided by the Dutch Inspectorate of Education consists of all students enrolled in pre-university education. Generalization of data to the whole population will therefore not be an issue. Effect sizes using Cohen's *d* were used to get information for interpreting the effect of the change in rules and regulations on discrepancies between marks on school examinations and central examinations.

The results of the analyses show a change in pattern between students' marks on school examinations and central examinations. This effect of a changing pattern is larger for students with higher SE marks. It seems likely that strategic behavior of students is a supplementary explanation for differences between students' marks on school examinations and students' marks on central examinations.

## **Foreword.**

This thesis is the result of my years studying at the University of Twente in Enschede. I started studying in 2006 intending to finish a bachelor and master in Biomedical Engineering. After only a few short months I found out that this study was not right for me. I decided to enroll in the bachelor Educational Design Management & Media in February 2007 and finished this bachelor in October 2012. I started the master Educational Science & Technology a month before finishing my bachelor. The attentive reader will have noticed a large discrepancy between the year starting my bachelor, 2007, and finishing my master, 2015. Although this time is long, I do not regret this time and I am proud that I will leave the University of Twente with a master's degree in Educational Science and Technology. There were many moments caused by health problems and other problems in the past nine years in which I thought I would never complete my bachelor or master study.

I would like to show my appreciation and gratitude for the involvement of my supervisors, starting with my first supervisor Dr. Hans J.W. Luyten. I admire the way he had patience with me and how all of his feedback was presented to me in a constructive manner. Second, I would like to thank Dr. B.A.N.M. Vreeburg of the Dutch Inspectorate of Education. He learned me a lot about statistics, was always available for answering my questions and had a lot of patience with me. The third supervisor I would like to thank is Prof. Dr. Ir. Bernard Veldkamp for receiving good feedback from him when my thesis was in the final stages.

Writing this thesis and completing my study in general would not have been possible without support of my parents and friends. I would like to thank everyone who supported me during my time being a student.

One chapter of my life ends. A chapter in which I developed myself as a person more than in every chapter before. I am looking forward to the next chapter of my life and hopefully this chapter will be as interesting and rewarding as my time of a student at the University of Twente.

Johan Leferink

Enschede, April 15, 2015.

## **Introduction and background.**

In the Netherlands there are two types of examinations in secondary educations. The first type of examination is the school examination. School examinations [SE] are the responsibility of the school and are an average of weighted marks in the last three years of pre-university education per subject depending on the program of assessment of the school (Programma van toetsing en Afsluiting). This mark per subject is an average of a student's performance over an extended time period. In contrast to the central examination, the school examination mark is based on numerous tests, assignment etc. but differs across schools. A school can decide on its own how to compose their SE marks. For VWO it can be composed of students' marks in their fourth, fifth or sixth year.

The central examination [CE] is the same per subject for each student in the Netherlands in a certain year, but is based on a student's performance at one single moment in time. It is made by Cito: a Dutch company making examinations. An important difference with school examinations is that the quality and difficulty of these examinations are guaranteed to be comparable with examinations of previous years (Kennisnet & CvTE, 2013).

Per subject, marks on both the school examination and the central examinations are averaged on an equal basis. Until 2011 a student graduated pre-university education when the average of his marks was above a 5.5. After 2011 a student is required to score at least an average CE mark of a 5.5.

The possibility that discrepancies between school examination marks and central examination marks could be attributed to strategic behavior of students, has thus far hardly been considered as an alternative explanation in literature. The Dutch Inspectorate of Education knows that students can perform strategically. This strategic behavior can be positive or negative. An example of negative strategic behavior is fraud by cheating on a test. An example of positive strategic behavior is a student studying longer for a test to maximize his chance of passing this test.

Strategic behavior of students as explanation for these differences is suggested, but not researched. A possible practical relevance of this study is a rule of thumb used by the Dutch Inspectorate of Education as one of many indicators about quality of schools. In the Netherlands there is a rule that the average mark difference for each school between school exams and central exams can be 0.5 point at most on a moving average of the last three years (Inspectie van het Onderwijs, 2014; Insectie van het Onderwijs 2007 in: Schildkamp, Rekers-mombarg, & Harms, 2012). This maximum of a 0.5 point difference serves as an indicator for the quality of education of a school and is considered a stable characteristic which is a reliable source of information about school revenues (Vreeburg, 2007). A small discrepancy between SE and CE marks is acceptable when it is caused by normal fluctuation and differences in mastery of subject matter by students. A discrepancy gets problematic when the direction of the difference gets the same for on average every student of a school. These students become systematically advantaged by their school relative to students of other schools on their changes of graduating VWO (Inspectie van het Onderwijs, 2008). Not meeting this indicator in combination with scoring insufficient on other indicators may change the opinion of the Dutch Inspectorate of Education on this school and can for example change the level of supervision of the school. This research might give more information about the validity of this indicator.

Differences and correlations between marks on school-based assessments and central exit examinations can differ between schools. Several studies have addressed this issue (De Lange & Dronkers, 2007; Himmler & Schwager, 2007; Reeves et al., 2001; Thomas et al., 1998; Wikström & Wikström, 2005; Willingham et al., 2002 in: Luyten & Dolkar, 2010).

Luyten & Dolkar (2010) summarize these as indicating a moderate degree of conformity within schools and nonconformity between schools on marks on school-based assessments and central exit examinations. The relative position of students within schools is largely the same on school-based assessments and central examinations, but the correlation of school averages on school-based assessments and central examinations is relatively weak. Students enrolled in the same school show a relatively strong correlation between school examination marks and central examination marks, whereas the correlation of school averages between central and school-based examinations tends to be lower. The central examinations in Dutch secondary education (and in most other countries) consists

of standardized tests given at one moment in time (two in case of a resit) and school examinations are a kind of resume of the school performance of a student over multiple years. Central examinations do have high stakes not only for students but also for the teachers and schools in its use for determining funding or school evaluations, strategic behavior to enhance output is predictable (Scheerens, Glas and Thomas, 2003 in: Luyten & Dolkar, 2010). In such cases, teachers may attempt to compensate for poor results on the central exam by overrating their students' performance in the school-based assessments. These overrating might be because of opportunistic behavior (Wößmann, 2005). A difference between different contents tested between an SE and CE can also result in differences between SE and CE marks (Vreeburg, 2007). An example is that speaking skills of languages are tested on the SE and reading comprehension is tested on the CE.

The validity and reliability of the pass or fail decision gets better by using both the central exams and the school exams results, because multiple samples are used. The average mark per subject on a school examinations is rather stable throughout the years, just like the level of central examinations is comparable between years (De Lange, M., & Dronkers, 2007; Kennisnet & CvTE, 2013; Vreeburg, Theunissen, & Coenen, 2010)

### **Possible causes of strategic behavior of students**

When making a decision, whether in general or about examinations, evaluating the outcomes of decision alternatives is a key step in any decision analysis (e.g. Clemen, 1996, in: Dekay & Patin, 2009). The standard procedure for making a choice between alternatives is weighing the relative desirability or utility of each alternative with the probability that these outcome will occur. The highest expected utility is selected. Intermediating factors, such as potential consequences are considered inputs to the evaluation of decision alternatives (Dekay & Patin, 2009). Besides the above consequentialist theory, people also often make non-consequentialist decisions that are affected by considerations other than consequences (Baron, 1994, in: Dekay & Patin, 2009). An example of such decision is that a person has a deadly disease and that there is an experimental vaccination with a high, almost certain, chance of death. In such cases, action is often chosen over inaction, even when death from the vaccination is considered worse than death from a disease. This behavior implies that evaluation of end results depends on the short term choices with their possible consequences (Dekay & Patin, 2009) and that students can make strategic decisions based on the amount of work or change that might lead to a certain outcome. This outcome can be a mark on a central examination, having a sufficient mark for a subject based on the average of a school examination and central examination mark or graduating for pre-university education in general. The above theory suggests that students can strategically study more for the exact same examinations if rules about graduating get harder. Interestingly, decision makers can influence people's decision making process towards certain outcomes by restructuring information, indicating that people have difficulty making clear distinctions between decisions and consequences. The above theory will now be illustrated with an example of a curriculum change in Germany where the education system changed to a system similar to the system used in the Netherlands.

In the last decade the German Educational System changed from a system with a curriculum-focused input regulation to a system focused on the attainment of learning outcomes. At first the curriculum per subject was content-driven. There was no compulsory testing. In the new situation there are end terms per subject and the curriculum is made to prepare students to this end terms. A risk of such a system is that only subject matter which is tested will be taught and no other important knowledge or skills (Jones, Jones & Hargrave, 2003 in: Jones, 2008). This risk is outweighed by numerous studies concluding that student achievement was higher by students having a central exit examination than by students not having a central exit examination (Jürges & Schneider, 2009; Oerke, Maag Merki, Holmeier, & Jäger, 2011; Schneider, 2003; Wößmann, 2005).

The key aim of this change was to improve student achievement by increasing the students' and their teachers' performance incentives by focusing on critical subject matter (Wößmann 2003, in: Oerke et al. 2011). Two relevant goals of central exit examinations for this paper are inducing teachers to set high standards and to motivate students to actually learn what they need to (Bishop, 2005). Teachers aligned their teaching to the specific needs of their national curriculum and teachers were forced to

find and solve weaknesses in their own curricula (Jones, 2008; Jürges, Schneider, Senkbeil, & Carstensen, 2012). In a study of Oerke et al. (2011) hypotheses about students' and teachers' motivation were tested. This study concludes that teachers' attributions to teaching and effort increased when their students' got higher marks. Probably because this teachers attribute parts of their students' success to themselves. An exception was the year in which their students had their central examination. A possible explanation is that the attribution to teaching was lower, because of stress about possible below par results of their students. Part of their study relates to changes in attribution to effort of students and hints towards strategic behavior of students. A significant interaction effect for effort relates to students who perceived themselves as successful. Students who put in more effort to pass their school exams and actually passed them felt more successful. A possible explanation of Oerke et al. (2011) is that students who score high on their school exam want to keep their scores high, preparing themselves well for their central examinations. This indicates that the introduction of central examinations in Germany seem to enhance advantageous attribution patterns, such as studying instead of doing nothing, in perceived 'successful' students (Oerke et al., 2011). Disadvantageous attributions of students perceiving themselves as less successful became less important. Oerke et al. (2011) do think that the reason for this is better teaching. This German case found clues for strategic behavior of students regarding their decision making about how much effort to put in for their central examination based on their perceived successfulness, which in turn is to a large extent based on marks on their school examinations.

The German case described and analyzed by Oerke et al., (2011) did already mention certain conditions effecting students attribution to effort. Several other factors found in literature will be described below.

#### ***School examination mark.***

The first factor to be discussed is a student's school examination mark per subject. Students with a high mark on their school examination wanted to keep their mark high and students that perceived themselves as successful expected themselves to be successful. Students can do this by using the opportunity to do a re-examination when he or she does not pass for a subject in one attempt. It could be possible that students try to pass an exam with the least amount of effort possible and take this chance because they know they have a second opportunity. Students could estimate this incorrectly and could fail by a small margin. In the second attempt a student could still pass because he has the opportunity to put more effort in a specific subject. With a little extra effort the student could pass using his second attempt (Kooreman, 2012). A side note is that Kooreman (2012) assumes that the total amount of effort needed and exerted by an individual student is the same for an individual student whether he needs a second opportunity or not. When rules about graduating change and a student need to put in more effort to pass a subject based on his school examination mark, a student may strategically decide to put in more effort for that subject.

#### ***Locus of control.***

The next possible reason for students to behave strategically is locus of control. Students perceiving themselves as more successful have more positive attributions to effort (Oerke et al., 2011). Can students' perceived successfulness be influenced to enhance a students' chance to put in more effort and therefore increase their chances of a pass? A students' personality trait which can influence the way students make decisions is locus of control. Locus of control refers to the extent in which students believe events are under personal control and this trait is relatively stable (Boon, Olffen, & Roijackers, 2004). Students perceiving themselves as highly successful have a high locus of control. Students perceiving themselves as less successful have a lower locus of control and attribute their fate more to factors like luck, chance, powerful individuals and institutions. These people feel their lives cannot entirely be controlled by their own actions (Lefcourt, 1982, in: Boon et al., 2004). Locus of control is additionally influenced by the believe students have in themselves. Students have a higher self-perceived intelligence and are more likely to demonstrate strategic actions when they are positive about themselves and do believe they are capable of performing academic tasks (Dermitzaki, Leondari, & Goudas, 2009; Dollinger & Clark, 2012). The main goal of the student is to graduate VWO. A student will try to optimize local goals to maximize the chance of reaching his main goal



(Chater & Oaksford, 2000). Examples of local goals are passing each different subject by having a high SE mark and a CE mark being high enough. Locus of control cannot be measured directly in the same way as the school examination mark can be measured. An assumption in this research is that the locus of control of the population does not differ in 2011 and 2012.

***Support.***

Students want that every piece of possible relevant information is used and considered when school leaders, teachers or parents make decisions or give feedback or advice to them. Students have the ‘human is better’ and ‘more is better’ attitude towards decision making processes. Students have the most positive attitude towards human decision making and for using all rather than some information (Eastwood, Snook, & Luther, 2012). Students have the least positive attitude towards decisions made by a decision maker who used a statistical formula which used little information. This despite past research showing that actuarial methods tend to outperform clinical methods across a range of domains (Kleinmuntz, 1990, in: Eastwood et al., 2012). This category, parental and school support, is mostly relevant for students who perceive themselves as students who have no good school examinations but expect to graduate by getting high enough marks on their central exams. Parents and schools can support students by influencing them, because intuition and deliberations are not stable and could change after consulting, or getting consulted by school leaders and/or parents (Laborde, Dosseville, & Scelles, 2010). Jürges & Schneider (2009) state that parental support for their children is the main determinant of individual educational success.

In this study support cannot be measured, just like locus of control. An assumption in this research is that support does not differ in 2011 and 2012.

***Interim summary.***

Table 1 shows the factors found which can lead to strategic behavior of students. Earlier is mentioned that action is chosen over inaction when both doing nothing and doing something will probably get the same result (Baron, 1994, in: Dekay & Patin, 2009). Besides the school examination mark, none of the factors found in table 1 can be measured directly, but they are assumed to be the same in 2011 and in 2012. It is expected that because of changes in rules and regulations between 2011 and 2012 regarding graduation student behavior changed as well. Especially the relation between SE marks and CE marks is expected to differ between 2011 and 2012.

Table 1

*Factors effecting student behavior regarding decisions about central examinations*

Factors			
Category	Factor	Source	Comment
Prior achievement	SE mark	(Oerke et al., 2011)	
	Locus of control	(Boon et al., 2004; Dermitzaki et al., 2009; Dollinger & Clark, 2012; Oerke et al., 2011)	Including perceived successfulness and academic self-belief
Expected output of effort	Prior effort used	(Kooreman, 2012)	
	Action over inaction	(Dekay & Patin, 2009)	When expected effort does not change expected outcome.
Support	Parental and school	(Eastwood et al., 2012; Laborde et al., 2010)	Affecting CE when SE is not that good
	Parental background	(Jürges & Schneider, 2009)	Affecting both SE and CE

## Rules and regulations

A student in Dutch pre-university education (VWO) before 2012 graduated when the average grade calculated over all subjects was a 5.5 or higher and additional conditions were met. The scale used is a scale from 1 till 10. Ten is the highest mark; a 5.5 or higher is regarded as sufficient. The grade per subject was calculated by averaging the mark on the school examination and the mark on the central exit examination. Both school and central examination marks are weighted even. The additional conditions are the following: one mark 5.0 can be compensated by other subjects when all other marks are a 6.0 or higher; one mark 4 can be compensated by other subjects when all other marks are a 6.0 or higher and the average mark on all marks is a 6.0 or higher. The second last condition is that two subjects being a mark 4 and a mark 5 or two marks 5.0 could be compensated when the average of all subjects is a 6.0 or higher. The last condition is that subjects marked with an insufficient, sufficient or good instead of the 10-point scale, needs to be graded at least sufficient (Rijksoverheid, 2013).

A change in regulations became active starting at school year 2011-2012. From this year on the previous rule were still active, but one additional rule became active. Not only the average grades per subject needed to be a 5.5 or higher, but also the average grade of all central exit examination marks needed to be a 5.5 or higher (Rijksoverheid, 2013). Table 2 below show

Table 2

### *Rules and regulations for school examinations*

Rule number	Rule	Active when rule # is met	Active when rule # is not met	Year(s) active
1	Average grade over all subject is at least 5.5	-		All Years
2	Maximum of one 5 and all other marks are at least 6.0	1		All years
3	Maximum of one 4 and all other marks are at least 6.0 and all marks average is at least 6.0	1	2	All years
4	Maximum of one 4 and one 5 or 2 times a 5 and all marks average at least 6.0	1	2, 3	All years
5	Average grade of central exit examinations marks is at least 5.5	1, 2, 3		2011-2012

A student graduates for his VWO when the conditions in the table above are met. When conditions are not met there still is the possibility of graduation through a reexamination. It is possible for a student to take an extra examination in a subject. The highest grade on the central exit examination is the valid mark. When the reexamination mark is lower than the first examination mark, the first mark is used by calculating the average grade per subject. The pass or fail for VWO is calculated in the same way as if no reexamination was done in one or more subjects (Rijksoverheid, 2013).

The extra requirement since 2012 of at least an average mark on the central examination of a 5.5 means that students with high scores on the school examinations could no longer afford low scores on the central examinations. Therefore it is expected not only that the differences between school and central examinations are smaller in 2011-2012 than the year(s) before, but also that the changes are most marked for student with high scores on the school exams. Nothing has changed for student with low scores on the school exams. They still need to compensate their low scores with high scores on the central examination. Until 2010-2011 students with high scores on the school exams could afford low scores on the central examinations, but since 2011-2012 they are required to get an average of at least 5.5 on the central examinations. It is therefore expected that in 2011-2012 especially the students with

high scores on the school examinations will get higher scores on the central examinations than students with similar scores on the school examinations in 2010-2011.

## **Research questions and hypotheses.**

The main goal of this research project is to detect if strategic behavior by students can account for any discrepancies in school examination and central examination grades. Changes in rules and regulations between 2011 and 2012 about graduating for pre-university education are exploited to find evidence of strategic behavior by students. It is assumed that students with the same school examination mark would, on average, have the same marks on their central examination in 2011 and 2012 if the rules for graduating had remained unchanged.

This research tries to find out whether discrepancies between results of school exams and central exams can be attributed to strategic behavior of students instead of schools. No literature so far appears to have addressed this possibility.

The main research question is therefore the following:

- Can discrepancies between results on school exams and central exams be explained by strategic behavior of students?

To answer this question it is important that at least the following questions will be answered as well:

1. Is there a difference in level and variation between school examination marks and central examination marks per subject after the rules and regulation change?
2. Is there a difference between school examination marks and central examination marks per subject caused by variation in school level and by variation on student level after the rules and regulations change?

### ***Hypotheses about differences between SE and CE marks in 2011 and 2012.***

It is expected that changes in rules and regulations (see table 1) will cause a different relation between marks on school exams and marks on central exams per subject. An example is that students with good marks on their school exams think they will graduate anyway and do not mind getting low grades on their central exams (2011,  $R_1$ ). With the new rules in 2012 ( $R_2$ ), they still need to get good results. Therefore it is expected that the differences between school examination marks and central examination marks will be smaller.

Especially students having high grades on their SE are expected to score higher under  $R_2$  than under rules  $R_1$  because, under  $R_2$ , they cannot compensate low CE marks with high SE marks anymore like it was possible under  $R_1$ . Therefore students will be categorized in four groups:

- SE mark below 5.5;
- SE mark higher than or equal to 5.5 and lower than 6.5;
- SE mark higher than or equal to 6.5 and 7.5 and
- SE mark equal to or higher than 7.5.

Each of the four groups will be analyzed two times:

- The difference between CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ );
- The difference between the SE minus CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ );

The four groups analyzed two times gives the eight hypotheses below. Hypotheses one till four do form the first set of hypotheses: differences per subject, per category, on CE marks between  $R_1$  and  $R_2$ .

1. SE mark lower than 5.5 and differences between CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )
  - $H_0: CE_{SE<5.5} R_2 = CE_{SE<5.5} R_1$
  - $H_1: CE_{SE<5.5} R_2 > CE_{SE<5.5} R_1$
2. SE mark higher than 5.5 and lower than 6.5 and differences between CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )
  - $H_0: CE_{SE5.5 \leq x < 6.5} R_2 = CE_{SE5.5 \leq x < 6.5} R_1$
  - $H_1: CE_{SE5.5 \leq x < 6.5} R_2 > CE_{SE5.5 \leq x < 6.5} R_1$
3. SE mark higher than 6.5 and 7.5, differences between CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )
  - $H_0: CE_{SE6.5 \leq x < 7.5} R_2 = CE_{SE6.5 \leq x < 7.5} R_1$
  - $H_1: CE_{SE6.5 \leq x < 7.5} R_2 > CE_{SE6.5 \leq x < 7.5} R_1$
4. SE mark equal to or higher than 7.5, differences between CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )
  - $H_0: CE_{SE \geq 7.5} R_2 = CE_{SE \geq 7.5} R_1$
  - $H_1: CE_{SE \geq 7.5} R_2 > CE_{SE \geq 7.5} R_1$

Hypotheses five till eight do form the second set of hypotheses: differences per subject, per category, on SE-CE marks between  $R_1$  and  $R_2$

5. SE mark lower than 5.5 and differences between SE minus CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )
  - $H_0: SE-CE_{SE<5.5} R_2 = SE-CE_{SE<5.5} R_1$
  - $H_1: SE-CE_{SE<5.5} R_2 < SE-CE_{SE<5.5} R_1$
6. SE mark higher than 5.5 and lower than 6.5 and differences between SE minus CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )
  - $H_0: SE-CE_{SE5.5 \leq x < 6.5} R_2 = SE-CE_{SE5.5 \leq x < 6.5} R_1$
  - $H_1: SE-CE_{SE5.5 \leq x < 6.5} R_2 < SE-CE_{SE5.5 \leq x < 6.5} R_1$
7. SE mark higher than 6.5 and 7.5, differences between SE minus CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )
  - $H_0: SE-CE_{SE6.5 \leq x < 7.5} R_2 = SE-CE_{SE6.5 \leq x < 7.5} R_1$
  - $H_1: SE-CE_{SE6.5 \leq x < 7.5} R_2 < SE-CE_{SE6.5 \leq x < 7.5} R_1$
8. SE mark equal to or higher than 7.5, differences between SE minus CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )
  - $H_0: SE-CE_{SE \geq 7.5} R_2 = SE-CE_{SE \geq 7.5} R_1$
  - $H_1: SE-CE_{SE \geq 7.5} R_2 < SE-CE_{SE \geq 7.5} R_1$

Effect sizes larger than 0.0 support hypotheses one till four. Effect sizes lower than 0.0 support hypotheses five till eight. A more negative effect size means the effect is larger.

Two extra hypotheses will be answered about the differences between the four groups based on their SE marks. These two hypotheses do form the third set of hypotheses.

9. Groups with a higher SE mark will have a higher effect size ( $d$ ) than groups with a lower SE mark based on the difference in CE marks in 2011 ( $R_1$ ) CE marks in 2012 ( $R_2$ )
  - $H_0: d CE_{SE<5.5} = d CE_{SE5.5 \leq x < 6.5} = d CE_{SE6.5 \leq x < 7.5} = d CE_{SE \geq 7.5}$
  - $H_1: d CE_{SE<5.5} < d CE_{SE5.5 \leq x < 6.5} < d CE_{SE6.5 \leq x < 7.5} < d CE_{SE \geq 7.5}$
10. Groups with a higher SE mark will have a higher effect size ( $d$ ) than groups with a lower SE mark based on the difference in SE minus CE marks in 2011 ( $R_1$ ) and SE minus CE marks in 2012 ( $R_2$ )
  - $H_0: d SE-CE_{SE<5.5} = d SE-CE_{SE5.5 \leq x < 6.5} = d SE-CE_{SE6.5 \leq x < 7.5} = d SE-CE_{SE \geq 7.5}$
  - $H_1: d SE-CE_{SE<5.5} < d SE-CE_{SE5.5 \leq x < 6.5} < d SE-CE_{SE6.5 \leq x < 7.5} < d SE-CE_{SE \geq 7.5}$

***Hypothesis about the amount of difference between SE and CE marks which can be attributed to students.***

Strategic behavior by some schools is the most common explanation for differences between school examination results and central examinations. The explanation is that school examinations in some schools are relatively easy compared to central examinations, whereas the school examinations are more difficult in other schools.

The variation in differences between students within schools should be relatively small if the main cause for the differences in results on school examinations and central examinations would be the standards that schools set. All students would profit (or suffer) to the same extent if their school sets low (or high) standards for the school examinations. This hypothesis can be tested by means of multilevel analysis. The variation in differences can be decomposed in a between schools component and a within schools component. These analyses can be conducted for each examination subject separately and also for the differences in grade list averages. In that case the analyses relate to the difference between a student's average score across all school exams and the average across all central exams. A side note is that a lower percentage of variation at school level does not necessarily mean that a school showed less strategic behavior. According to the College voor Toetsen en Examens [CvE], the level of difficulty of the central examination was the same in 2011 and 2012 (Inspectie van het Onderwijs, 2014). Therefore, the variable year will be account for changes in rules and regulations between 2011 and 2012.

A relatively large percentage of variation at the school level would support the idea that differences in standard setting between schools is the main cause for differences between school and central examinations. From school effectiveness research we know (Scheerens & Bosker, 1997) that on average about 10-15% of the variance in student achievement scores relates to difference between school means. The remaining 85% relates to differences between students within schools. If the school level variance in differences clearly exceeds the "standard amount" of 15%, this can be considered as support for the idea that school standards account to some extent for the differences between school examinations and central examinations. The school variance in differences can also be compared to the school variance in central examination scores.

In the case of strategic behavior by schools, a substantial amount of school level variance in differences between school examination marks and central examination marks is expected. In the case of opportunistic student behavior mostly variation in central examination and school examination differences at the student level is expected.

Intra class variation per subject (both on school examination and central examination) serves as a benchmark. Central examination marks have previously been found to vary more strongly than school examination marks. A side mark is that central examination marks probably also show more school level variance by having a higher intra class variation. Reliability of tests scores is probably lower for central examination marks, because a central examination mark is based on a single test versus a school examination mark being an average over many tests in case of school exams. The same analyses will be done for the school examination marks, for the central examination marks and for the school examination marks minus the central examination marks. These analyses are: per subject per year (2011 and 2012), and multilevel analyses (zero-model with intra class correlations). In total four multilevel analyses per subject will be done. The amounts of variation attributed to the school will be compared between 2011 and 2012. Hypothesis eleven and twelve will be answered after performing these analyses.

It is possible that the proportion of variation attributed to the school is lower, but the total amount of strategic behavior of the school was higher. Hypotheses eleven and twelve are stated as below and do form the final fourth set of hypotheses:

11.  $H_0: ICC\_2011\_CE_{school}R_2 = ICC\_2012\_CE_{school}R_1$   
 $H_1: ICC\_2011\_CE_{school}R_2 > ICC\_2012\_CE_{school}R_1$

12.  $H_0: ICC\_2011\_SE-CE_{school}R_2 = ICC\_2012\_SE-CE_{school}R_1$   
 $H_1: ICC\_2011\_SE-CE_{school}R_2 > ICC\_2012\_SE-CE_{school}R_1$

## **Research design.**

### ***Research method.***

Evidence for a possible supplementary explanation for discrepancies between students' marks on school and central examinations will be sought by statistically analyzing students' data. IBM SPSS Statistics 22 [SPSS] is a statistical software package. SPSS will be used to do all the analyses. The planned analyses are described in the section 'data analysis', which is found below.

### ***Respondents.***

The respondents of this study are all students that were enrolled in pre-university secondary education in the Netherlands and took their central examinations in the year 2011 or in the year 2012.

The researcher signed an agreement in which they guaranteed that it will not be possible to retrace results back to individual students or schools. The dataset provided by the Dutch Inspectorate of Education does contain literally all students who participated in central examinations in a certain year. The total amount of respondents in 2011 are 36760 and in 2012 36794. In total there are 73509 respondents.

The selection of subjects to be analyzed is done by conveniently select every subject having both SE and CE examinations.

### ***Instrumentation.***

The datasets are supplied by the Dutch Inspectorate of Education, but originate from DUO. DUO is the Dutch Education Executive Agency. The examination marks are delivered by the school to DUO and are the official marks on which students may enter higher education after graduating their secondary education. Therefore the quality of the data is high.

Marks on school examinations and central examinations of fourteen different subjects will be analyzed, as well as the average school examination and central examination mark per student. These subjects including the amount of students per subjects are found in appendix A. These subjects are selected because they had both school examinations and central examinations. Using fourteen different subjects means that lots of instruments are used: per year fourteen central exit examinations made by Cito and fourteen school examination marks per subject per year. These school examination marks differ per school and are composed of multiple tests per subjects.

### ***Data analyses.***

All data is provided by the Dutch Inspectorate of Education. This data contains information about every student in secondary education in the Netherlands who participated in the Dutch central examinations in a particular year. This information includes school examination and central examination marks, as well as students' background information.

Having literally all the data of the students enrolled in the selected subjects in secondary education renders concerns about statistical significance due to sampling irrelevant: all data will be used and analyzed. Having all the data means that statistical significance tests are redundant because there is no need to generalize conclusions to the whole population. Still, information is needed about observed differences between 2012 and 2011. Describing differences between groups by effect sizes are suitable in such situation (Neill, 2008). An effect size of at least .2 will be considered small, an effect score of at least .5 will be considered moderate and an effect size of at least .8 will be considered large (Cohen, 1988). Differences between SE marks in 2011 and 2012 are not analyzed. This study distributes student in four groups based on their SE mark. This arbitrary decision makes it not relevant to compare these two years.



Two different sorts of analyses will be used to answer the main research question: descriptive statistics and multilevel analyses. Both will now be introduced, starting with the descriptive statistics.

*Descriptive statistics.*

Descriptive statistics will be used to answer the first ten hypotheses. These are the first, second and third set of hypotheses. First it will be used to show differences in level and variation between school examinations marks and central examination marks. Averages and standard deviations will be given for each subject the students had examinations for.

For each subject, descriptive statistics will be presented four times: the average central examination mark in 2011, the difference between school examination mark and central examination mark in 2011 and the previous two analyses again for the year 2012.

These analyses will show differences in level and variation between school examinations and central examinations. These differences between 2010-2011 and 2011-2012 are relevant because of changes in regulations. Poor results on central examination can no longer be compensated by good results on school examinations. This change in regulations can be used to find different patterns in behavior of students. The first set of hypothesis will be answered after these analyses.

Descriptive statistics will also be used to show if and, in case of yes, what differences there are between school examination marks and central examination marks per subject, per year. This will be done by calculating the average central examination mark per year. The hypothesis is that in subjects having new rules, a sufficient mark on central examinations is needed; the mark on central examinations for each subject will be higher. It is expected not only that the differences between school and central examinations are smaller in 2012 than in 2011, but also that the changes are most marked for student with high scores on the school exams. Nothing has changed for student with low scores on the school exams. They still need to compensate their low scores with high scores on the central examination. Until 2011 students with high scores on the school exams could afford low scores on the central examinations, but since 2012 they are required to get an average of at least 5.5 on the central examinations. It is therefore expected that in 2012 especially the students with high scores on the school examinations got higher scores on the central examinations than students with similar scores on the school examinations in 2011.

Effect sizes will be given every time differences between marks on school examinations and central examinations will be analyzed. Microsoft Excel 2013 [Excel] will be used to calculate effect sizes because in SPSS it is not possible to calculate effect sizes. The following formula will be used to calculate the effect sizes:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Where  $s$ , the pooled standard deviation, is calculated using the following formula:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

The standard error of  $d$  gives information about the confidence interval of  $d$ . A relative low standard error of  $d$  in relation to a relatively high value of  $d$  means that the effect size is more informative than a standard error of  $d$  of approximately the same size of  $d$ , or even smaller than  $d$ . The size of  $d$  should be at least two times the standard error of  $d$  to be 95% certain that the effect is bigger than 0 (Cooper, Hedges, & Valentine, 2009). The standard error of  $d$  is calculated with the following formula (Cooper & Hedges, 1994):

$$SE d = \sqrt{\left(\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)}\right) \left(\frac{n_1 + n_2}{n_1 + n_2 - 2}\right)}$$

Effect sizes will be calculated for hypotheses one till ten. These categories are based on different marks: insufficient ( $x < 5.5$ ), sufficient ( $5.5 \leq x < 6.5$ ), satisfying ( $6.5 \leq x < 7.5$ ) and good ( $x \geq 7.5$ ).

### *Multilevel modeling.*

Multilevel modeling is used to answer the fourth set of hypotheses. The percentage of variance in student achievement which can be attributed to the school will be calculated for two years: 2011 and 2012. The data is corrected for the student's school. The influence of the change in rules and regulations can be seen by comparing the zero-model of 2011 with the zero-model of 2012. The word 'brin' in the formula below refers to the unique id of a school.

Both zero-models will be used two times. The CE marks of students are compared per subject and the SE-CE marks of students will be calculated per subject.

Multilevel analyse 1: zero model CE 2011

$$CE_{\text{mark}} = SE_{\text{mark},0} + \text{brin}_{\text{aj}} + e_{ij}$$

Multilevel analyse 2: zero model CE 2012

$$CE_{\text{mark}} = SE_{\text{mark},0} + \text{brin}_{\text{aj}} + e_{ij}$$

Multilevel analyse 3: zero model SE-CE 2011

$$SE\text{-}CE_{\text{mark}} = SE_{\text{mark},0} + \text{brin}_{\text{aj}} + e_{ij}$$

Multilevel analyse 4: zero model SE-CE 2012

$$SE\text{-}CE_{\text{mark}} = SE_{\text{mark},0} + \text{brin}_{\text{aj}} + e_{ij}$$

### ***Procedures.***

The datasets available consist of two datasets: datasets of the years 2011 and 2012. These datasets will be joined to form a new dataset containing both the years 2011 and 2012. Records of students who were not enrolled in pre-university secondary education will be removed from the newly created dataset. Non-relevant data will be removed too. An example of non-relevant data is students' postal codes, ethnicity and school's denomination and the area's urbanization.

Fourteen of the subjects students had examinations marks for, are used. The other subjects had too little respondents. The fourteen subjects are, in alphabetical order, the following: biology, chemistry, Dutch language, economics, French language, geography, history, management and organization, mathematics A, mathematics B, Mathematics C and physics.

Data which can be traced back to individual schools or students will not be published, or will only be published anonymized only with approval of the Dutch Inspectorate of Education.

## Results.

The results of the different analyses described in the method section will be presented here. The hypotheses answered using descriptive statistics will be answered first. The hypotheses answered using multilevel modeling will be answered the first and second set of hypotheses.

### *First set of hypotheses (one till four).*

Hypothesis one is tested using descriptive statistics and calculated by using effect sizes. Hypothesis one is that the differences per subject between school examination marks and central examination marks will be smaller in 2012 than in 2011 for each of the categories: insufficient ( $x < 5.5$ ), sufficient ( $5.5 \leq x < 6.5$ ), satisfying ( $6.5 \leq x < 7.5$ ) and good ( $x \geq 7.5$ ). The results of hypotheses one till eight will be given below. A negative effect size (d) means that the CE mark is higher in 2012 than in 2011 for hypotheses one till four. A negative effect size (d) means that the difference between SE-CE marks 2012 is larger than in 2011 for hypotheses five till eight.

The results of the analyses on CE marks and SE-CE marks are not necessarily the same. The distribution of SE marks inside a category does not have influence on the analyses of the CE mark, but does have an influence on the SE-CE analyses.

### *Hypothesis 1.*

1. SE mark lower than 5.5 and differences between CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )

$$H_0: CE_{SE < 5.5} R_2 = CE_{SE < 5.5} R_1$$

$$H_1: CE_{SE < 5.5} R_2 > CE_{SE < 5.5} R_1$$

Table 3

### *Effect sizes for CE marks of students scoring lower than a 5.5 on their SE*

Subject	2011			2012			$\Delta m$	d	SE d
	n	m	SD	n	m	SD			
Biology	662	5.15	0.97	647	5.06	0.95	-0.09	-0.09	0.06
Chemistry	2134	5.23	1.05	2081	5.41	1.07	0.18	0.17	0.03
Dutch	1100	5.61	0.99	1157	5.61	0.98	0.00	0.00	0.04
Economics	2086	5.10	1.02	2049	5.19	1.09	0.09	0.09	0.03
English	2016	4.89	0.92	2132	4.96	1.02	0.07	0.07	0.03
French	1299	5.31	1.02	1322	5.62	1.03	0.31	0.30	0.04
Geography	246	4.94	0.90	344	4.94	0.94	0.00	0.00	0.08
German	1848	5.38	0.97	1824	5.52	0.97	0.14	0.14	0.03
History	613	4.92	1.10	740	5.18	1.10	0.26	0.24	0.05
M & O	755	5.02	1.08	747	4.90	1.04	-0.12	-0.11	0.05
Mathematics A	2369	5.16	1.11	2093	5.25	1.19	0.09	0.08	0.03
Mathematics B	3198	4.84	1.29	2836	4.99	1.22	0.15	0.12	0.03
Mathematics C	428	5.23	1.25	366	5.41	1.20	0.18	0.15	0.07
Physics	1612	5.11	0.93	1632	5.06	1.00	-0.05	-0.05	0.04
All subjects	479	4.95	0.73	529	5.01	0.73	0.06	0.08	0.06

Table 3 shows that  $H_0$  cannot be discarded for five of fourteen subjects. These are the following subjects: biology, Dutch language, geography, physics and management and organization.  $H_0$  can be discarded for the other subjects. All effect sizes can be considered (very) small using the classification of Cohen (1988) because most are less than 0.2. The two exceptions are French language and history. The other d's are still relevant and do give information even though they are not relevant according to

Cohen. The large  $n$  and having data on the entire population makes every difference between 2011 and 2012 relevant. The value of  $d$  is less than two times the standard error of  $d$ . This means that it is less than 95% certain that there is an effect, a difference, between 2011 and 2012.

*Hypothesis 2.*

2. SE mark higher than 5.5 and lower than 6.5 and differences between CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )

$$H_0: CE_{SE5.5 \leq x < 6.5} R_2 = CE_{SE5.5 \leq x < 6.5} R_1$$

$$H_1: CE_{SE5.5 \leq x < 6.5} R_2 > CE_{SE5.5 \leq x < 6.5} R_1$$

Table 4

*Effect sizes for CE marks of students scoring higher than or equal to a 5.5 and lower than a 6.5 on their SE*

Subject	2011			2012			$\Delta m$	$d$	SE $d$
	$n$	$m$	SD	$n$	$m$	SD			
Biology	7426	5.90	0.90	7494	5.82	0.89	-0.08	-0.09	0.02
Chemistry	7447	5.99	0.89	7496	6.21	0.89	0.22	0.25	0.02
Dutch	10883	5.90	0.89	11616	5.93	0.92	0.03	0.03	0.01
Economics	7914	5.86	0.95	7990	6.01	0.95	0.15	0.16	0.02
English	10697	5.53	0.93	11747	5.64	1.02	0.11	0.11	0.01
French	5528	5.75	1.01	5836	5.98	1.04	0.23	0.22	0.02
Geography	4297	5.69	0.84	4494	5.80	0.81	0.11	0.13	0.02
German	7243	5.64	0.96	7286	5.72	0.97	0.08	0.08	0.02
History	6256	5.67	0.91	6659	5.99	0.88	0.32	0.36	0.02
M & O	3937	5.81	0.99	4003	5.67	0.95	-0.14	-0.14	0.02
Mathematics A	6109	5.93	0.99	6316	6.07	0.97	0.14	0.14	0.02
Mathematics B	5905	5.87	1.21	5826	6.01	1.13	0.14	0.12	0.02
Mathematics C	940	6.02	1.11	865	6.18	1.06	0.16	0.15	0.05
Physics	6993	5.90	0.86	7083	5.88	0.91	-0.02	-0.02	0.02
All subjects	15803	5.79	0.55	16400	5.90	0.56	0.11	0.20	0.01

Table 4 shows that  $H_0$  can be discarded for eleven of fourteen subjects.  $H_0$  is true for the following subjects: biology, physics and management and organization.  $H_0$  can be discarded for the other subjects. Three effect sizes are bigger than 0.2: French language, history and chemistry. The other  $d$ 's are still relevant and do give information even though they are very small according to Cohen. The large  $n$  and having data on the entire population makes every difference between 2011 and 2012 relevant.

*Hypothesis 3.*

3. SE mark higher than 6.5 and 7.5, differences between CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )

$$H_0: CE_{SE6.5 \leq x < 7.5} R_2 = CE_{SE6.5 \leq x < 7.5} R_1$$

$$H_1: CE_{SE6.5 \leq x < 7.5} R_2 > CE_{SE6.5 \leq x < 7.5} R_1$$

Table 5

*Effect sizes for CE marks of students scoring higher than or equal to a 6.5 and lower than a 7.5 on their SE*

Subject	2011			2012			$\Delta m$	$d$	SE $d$
	$n$	$m$	$SD$	$n$	$m$	$SD$			
Biology	7874	6.69	0.94	7972	6.64	0.92	-0.05	-0.05	0.02
Chemistry	6668	6.68	0.89	6677	6.87	0.83	0.19	0.22	0.02
Dutch	18637	6.25	0.88	18333	6.37	0.93	0.12	0.13	0.01
Economics	6905	6.60	1.01	6630	6.73	0.91	0.13	0.14	0.02
English	15352	6.25	0.90	14679	6.46	0.96	0.21	0.23	0.01
French	5814	6.34	1.02	5710	6.62	1.14	0.28	0.26	0.02
Geography	5116	6.30	0.84	5252	6.37	0.78	0.07	0.09	0.02
German	6976	6.09	0.98	6850	6.25	1.04	0.16	0.16	0.02
History	8183	6.28	0.86	8065	6.61	0.83	0.33	0.39	0.02
M & O	3974	6.47	1.01	4143	6.40	0.92	-0.07	-0.07	0.02
Mathematics A	5700	6.56	0.97	5960	6.70	0.90	0.14	0.15	0.02
Mathematics B	4824	6.79	1.18	4840	6.97	1.13	0.18	0.16	0.02
Mathematics C	733	6.68	1.04	698	6.75	1.04	0.07	0.07	0.05
Physics	6554	6.60	0.91	6478	6.67	0.93	0.07	0.08	0.02
All subjects	16639	6.47	0.58	16193	6.63	0.56	0.16	0.28	0.01

Table 5 shows that  $H_0$  can be discarded for eleven of fourteen subjects. The effect sizes of biology and management and organization are too small and the standard error of is bigger than half of the effect size for Mathematics C.  $H_0$  cannot be discarded for the other subjects. Two effect sizes are bigger than 0.2: French language and history. The other  $d$ 's are still relevant and do give information even though they are very small according to Cohen. The large  $n$  and having data on the entire population makes every difference between 2011 and 2012 relevant.

*Hypothesis 4.*

4. SE mark equal to or higher than 7.5, differences between CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )

$$H_0: CE_{SE \geq 7.5} R_2 = CE_{SE \geq 7.5} R_1$$

$$H_1: CE_{SE \geq 7.5} R_2 > CE_{SE \geq 7.5} R_1$$

Table 6

*Effect sizes for CE marks of students scoring higher than or equal to 7.5 on their SE*

Subject	2011			2012			$\Delta m$	$d$	SE $d$
	$n$	$m$	SD	$n$	$m$	SD			
Biology	2491	7.81	0.92	2507	7.81	0.85	0.00	0.00	0.03
Chemistry	3664	7.77	0.87	3690	7.85	0.76	0.08	0.10	0.02
Dutch	6099	6.80	0.85	5597	7.06	0.92	0.26	0.29	0.02
Economics	2971	7.77	1.07	2831	7.70	0.91	-0.07	-0.07	0.03
English	8659	7.15	0.81	8135	7.43	0.85	0.28	0.34	0.02
French	3256	7.49	0.99	2909	7.90	1.18	0.41	0.38	0.03
Geography	1209	7.20	0.80	1187	7.16	0.75	-0.04	-0.05	0.04
German	2886	7.00	1.02	2587	7.30	1.16	0.30	0.28	0.03
History	3084	7.14	0.84	2800	7.44	0.80	0.30	0.37	0.03
M & O	1902	7.51	1.00	1835	7.39	0.93	-0.12	-0.12	0.03
Mathematics A	2720	7.45	0.94	2961	7.48	0.83	0.03	0.03	0.03
Mathematics B	3576	8.22	1.17	3742	8.39	1.11	0.17	0.15	0.02
Mathematics C	266	7.33	1.02	226	7.50	1.02	0.17	0.17	0.09
Physics	3503	7.80	0.94	3480	7.87	0.92	0.07	0.08	0.02
All subjects	3821	7.62	0.57	3627	7.77	0.56	0.15	0.27	0.02

Table 6 shows that  $H_0$  can be discarded for nine of fourteen subjects.  $H_0$  cannot be discarded for the following subjects because of the effect size is too small: economics, geography and management and organization. The standard errors of  $d$  are higher than half of  $d$  for the following subjects: Mathematics A and mathematics C. It is interesting that the differences between CE marks between 2011 and 2012 are largest for each of the four languages, with 2012 being higher. Five effect sizes are bigger than 0.2: English language, Dutch language, German language, French language and history. The other  $d$ 's are still relevant and do give information even though they are very small according to Cohen. The large  $n$  and having data on the entire population makes every difference between 2011 and 2012 relevant.

**Second set of hypotheses (five till eight).**

*Hypothesis 5.*

5. SE mark lower than 5.5 and differences between SE minus CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )

$$H_0: SE-CE_{SE \geq 5.5} R_2 = SE-CE_{SE < 5.5} R_1$$

$$H_1: SE-CE_{SE < 5.5} R_2 < SE-CE_{SE < 5.5} R_1$$

Table 7

*Effect sizes for SE-CE marks of students scoring higher than or equal to a 5.5 and lower than a 6.5 on their SE*

Subject	2011			2012			$\Delta m$	$d$	SE $d$
	$n$	$m$	SD	$n$	$m$	SD			
Biology	662	0.02	0.95	647	0.14	0.92	0.12	0.13	0.06
Chemistry	2134	-0.21	0.98	2081	-0.39	0.99	-0.18	-0.18	0.03
Dutch	1100	-0.48	1.04	1157	-0.51	1.02	-0.03	-0.03	0.04
Economics	2086	-0.07	0.98	2049	-0.17	1.03	-0.10	-0.10	0.03
English	2016	0.20	0.92	2132	0.13	1.03	-0.07	-0.07	0.03
French	1299	-0.27	1.04	1322	-0.59	1.06	-0.32	-0.30	0.04
Geography	246	0.24	0.88	344	0.27	0.89	0.03	0.03	0.08
German	1848	-0.34	1.01	1824	-0.47	1.01	-0.13	-0.13	0.03
History	613	0.21	1.07	740	-0.05	1.04	-0.26	-0.25	0.05
M & O	755	0.03	1.02	747	0.16	0.99	0.13	0.13	0.05
Mathematics A	2369	-0.30	1.05	2093	-0.33	1.12	-0.03	-0.03	0.03
Mathematics B	3198	0.02	1.21	2836	-0.11	1.14	-0.13	-0.11	0.03
Mathematics C	428	-0.37	1.19	366	-0.55	1.14	-0.18	-0.15	0.07
Physics	1612	-0.08	0.87	1632	-0.01	0.95	0.07	0.08	0.04
All subjects	497	0.35	0.67	529	0.29	0.69	-0.06	-0.09	0.06

Table 7 shows that  $H_0$  can be discarded for eight subjects except the following: biology, geography, physics and management and organization because of a high  $d$  and Dutch Language and Mathematics A because of a standard error being too big.  $H_0$  cannot be confirmed for these four subjects. Two effect sizes are they are bigger than 0.2: French language and history. The other  $d$ 's are still relevant and do give information even though they are very small according to Cohen. The large  $n$  and having data on the entire population makes every difference between 2011 and 2012 relevant. The value of  $d$  is less than two times the standard error of  $d$ . This means that it is less than 95% certain that there is an effect, a difference, between 2011 and 2012.



*Hypothesis 6.*

6. SE mark higher than 5.5 and lower than 6.5 and differences between SE minus CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )

$$H_0: SE-CE_{SE5.5 \leq x < 6.5} R_2 = SE-CE_{SE5.5 \leq x < 6.5} R_1$$

$$H_1: SE-CE_{SE5.5 \leq x < 6.5} R_2 < SE-CE_{SE5.5 \leq x < 6.5} R_1$$

Table 8

*Effect sizes for SE-CE marks of students scoring higher than or equal to a 5.5 and lower than a 6.5 on their SE*

<u>Subject</u>	<u>2011</u>			<u>2012</u>			<u><math>\Delta m</math></u>	<u><math>d</math></u>	<u>SE <math>d</math></u>
	<u><math>n</math></u>	<u><math>m</math></u>	<u>SD</u>	<u><math>n</math></u>	<u><math>m</math></u>	<u>SD</u>			
Biology	7426	0.15	0.88	7494	0.23	0.86	0.08	0.09	0.02
Chemistry	6668	0.21	0.86	6677	0.02	0.80	-0.19	-0.23	0.02
Dutch	10883	0.17	0.90	11616	0.15	0.92	-0.02	-0.02	0.01
Economics	7914	0.13	0.93	7990	-0.02	0.92	-0.15	-0.16	0.02
English	10697	0.51	0.92	11747	0.39	0.99	-0.12	-0.13	0.01
French	5528	0.26	1.01	5836	0.03	1.04	-0.23	-0.22	0.02
Geography	4297	0.39	0.82	4494	0.28	0.79	-0.11	-0.14	0.02
German	7243	0.37	0.97	7286	0.28	0.99	-0.09	-0.09	0.02
History	6256	0.39	0.90	6659	0.06	0.86	-0.33	-0.38	0.02
M & O	3974	0.43	0.98	4143	0.50	0.89	0.07	0.07	0.02
Mathematics A	6109	0,06	0,98	6316	-0,07	0,96	-0,13	-0,13	0,02
Mathematics B	5905	0,10	1,18	5826	-0,04	1,09	-0,12	-0,14	0,02
Mathematics C	940	-0,04	1,09	865	-0,20	1,05	-0,15	-0,04	0,05
Physics	6554	0.29	0.88	6478	0.23	0.89	-0.06	-0.07	0.02
All subjects	15803	0.35	0.52	16400	0.24	0.51	-0.11	-0.21	0.01

Table 8 shows that  $H_0$  can be discarded for eleven out of fourteen subjects.  $H_0$  cannot be discarded for the following subjects because of a high value of  $d$ : biology and management and organization and for mathematics C because of the standard error being too high.. Three effect sizes are bigger than 0.2: French language, history and chemistry. The other  $d$ 's are still relevant and do give information even though they are very small according to Cohen. The large  $n$  and having data on the entire population makes every difference between 2011 and 2012 relevant.

*Hypothesis 7.*

7. SE mark higher than 6.5 and 7.5, differences between SE minus CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )

$$H_0: SE-CE_{SE6.5 \leq x < 7.5} R_2 = SE-CE_{SE6.5 \leq x < 7.5} R_1$$

$$H_1: SE-CE_{SE6.5 \leq x < 7.5} R_2 < SE-CE_{SE6.5 \leq x < 7.5} R_1$$

Table 9

*Effect sizes for SE-CE marks of students scoring higher than or equal to a 6.5 and lower than a 7.5 on their SE*

<u>Subject</u>	<u>2011</u>			<u>2012</u>			<u><math>\Delta m</math></u>	<u>d</u>	<u>SE d</u>
	<u>n</u>	<u>m</u>	<u>SD</u>	<u>n</u>	<u>m</u>	<u>SD</u>			
Biology	7874	0.19	0.89	7972	0.24	0.86	0.05	0.06	0.02
Chemistry	6668	0.21	0.86	6677	0.02	0.80	-0.19	-0.23	0.02
Dutch	18637	0.67	0.88	18333	0.53	0.91	-0.14	-0.16	0.01
Economics	6905	0.28	0.97	6630	0.15	0.88	-0.13	-0.14	0.02
English	15352	0.67	0.86	14679	0.46	0.92	-0.21	-0.24	0.01
French	5814	0.56	0.99	5710	0.28	1.10	-0.28	-0.27	0.02
Geography	5116	0.57	0.81	5252	0.50	0.75	-0.07	-0.09	0.02
German	6976	0.80	0.96	6850	0.64	1.01	-0.16	-0.16	0.02
History	8183	0.63	0.84	8065	0.28	0.81	-0.35	-0.42	0.02
M & O	3974	0.43	0.98	4143	0.50	0.89	0.07	0.07	0.02
Mathematics A	5700	0.34	0.95	5960	0.20	0.89	-0.14	-0.15	0.02
Mathematics B	4824	0.10	1.15	4840	-0.07	1.09	-0.17	-0.15	0.02
Mathematics C	733	0.18	1.03	698	0.14	1.03	-0.04	-0.04	0.05
Physics	6554	0.29	0.88	6478	0.23	0.89	-0.06	-0.07	0.02
All subjects	16639	0.42	0.51	16193	0.27	0.48	-0.15	-0.30	0.01

Table 9 shows that  $H_0$  can be discarded for eleven out of fourteen subjects.  $H_0$  cannot be discarded for the following subjects because of a high d: biology and management and organization and for mathematics C because of the standard error being too high. Four effect sizes are bigger than 0.2: English language, French language, history and chemistry. The other d's are still relevant and do give information even though they are very small according to Cohen. The large  $n$  and having data on the entire population makes every difference between 2011 and 2012 relevant.

*Hypothesis 8.*

8. SE mark equal to or higher than 7.5, differences between SE minus CE marks per subject in 2011 ( $R_1$ ) and 2012 ( $R_2$ )

$$H_0: SE-CE_{SE \geq 7.5} R_2 = SE-CE_{SE \geq 7.5} R_1$$

$$H_1: SE-CE_{SE \geq 7.5} R_2 < SE-CE_{SE \geq 7.5} R_1$$

Table 10

*Effect sizes for SE-CE marks of students scoring higher than or equal to 7.5 on their SE*

Subject	2011			2012			$\Delta m$	$d$	SD $d$
	$n$	$m$	SD	$n$	$m$	SD			
Biology	2491	0.08	0.84	2507	0.09	0.75	0.01	0.01	0.03
Chemistry	3664	0.29	0.76	3690	0.20	0.65	-0.09	-0.13	0.02
Dutch	6099	1.05	0.83	5597	0.79	0.89	-0.26	-0.30	0.02
Economics	2971	0.23	0.95	2831	0.29	0.82	0.06	0.07	0.03
English	8659	0.85	0.75	8135	0.56	0.80	-0.29	-0.37	0.02
French	3256	0.58	0.87	2909	0.12	1.04	-0.46	-0.48	0.03
Geography	1209	0.64	0.75	1187	0.67	0.71	0.03	0.04	0.04
German	2886	1.00	0.93	2587	0.68	0.99	-0.32	-0.33	0.03
History	3084	0.79	0.78	2800	0.48	0.75	-0.31	-0.40	0.03
M & O	1902	0.48	0.89	1835	0.60	0.84	-0.12	-0.14	0.03
Mathematics A	2720	0.55	0.87	2961	0.52	0.79	-0.03	-0.04	0.03
Mathematics B	3576	0.00	0.99	3742	-0.18	0.94	-0.18	-0.19	0.02
Mathematics C	266	0.62	0.94	226	0.43	0.98	-0.19	-0.20	0.09
Physics	3503	0.27	0.80	3480	0.19	0.79	-0.08	-0.10	0.02
All subjects	3821	0.30	0.42	3627	0.14	0.43	-0.16	-0.38	0.02

Table 10 shows that  $H_0$  can be discarded for ten of fourteen subjects.  $H_0$  cannot be discarded for the following subjects because of a high  $d$ : biology, economics and geography and the standard error of mathematics A is too high. It is interesting that the differences between CE marks between 2011 and 2012 are the largest for each of the four languages, with 2012 being higher. Five effect sizes are bigger than 0.2: English language, Dutch language, German language, French language and history. The other  $d$ 's are still relevant and do give information even though they are very small according to Cohen. The large  $n$  and having data on the entire population makes every difference between 2011 and 2012 relevant.

***Interim summary about the first two sets of hypotheses.***

$H_0$  can be discarded for most subjects. Table 11 shows the effect sizes when the average mark of all students is analyzed for two sets of hypotheses: differences in CE marks between 2011 and 2012 and differences between SE-CE marks between 2011 and 2012. Both sets of hypotheses are tested for the four categories.

Table 11 shows that the effect size is larger for students having a high SE mark compared with students having a lower SE mark. The effect sizes are larger for the analyses of the SE-CE than for the analyses of the CE mark. It should be noted that the value of  $d$  is less than two times the standard error of  $d f$  for both the SE and SE-CE analyses for the categories with SE marks lower than a 5.5. This means that it is less than 95% certain that there is an effect, a difference, between 2011 and 2012 for these categories.

Table 11

*A summary of the effect sizes about the differences between 2011 and 2012 on CE and SE-CE for all subjects*

	<u>CE</u>				<u>SE-CE</u>			
	<u>&lt;5.5</u>	<u>&lt;6.5</u>	<u>&lt;7.5</u>	<u>=&gt;7.5</u>	<u>&lt;5.5</u>	<u>&lt;6.5</u>	<u>&lt;7.5</u>	<u>=&gt;7.5</u>
All subjects	0.08	0.20	0.28	0.27	-0.09	-0.21	-0.30	-0.38

***Third set of hypotheses (nine and ten).***

*Hypothesis 9 and 10.*

This third set of hypotheses is about differences between groups. All effect sizes of the previous eight hypotheses are shown in table 12. This table will be used to answer hypotheses nine and ten. Hypotheses 9 and 10 are the following:

1. Groups with a higher SE mark will have a higher effect size (d) than groups with a lower SE mark based on the difference in CE marks in 2011 ( $R_1$ ) CE marks in 2012 ( $R_2$ )

$$H_{01} d CE_{SE<5.5} = d CE_{SE5.5 \leq x < 6.5} = d CE_{SE6.5 \leq x < 7.5} = d CE_{SE \geq 7.5}$$

$$H_{11} d CE_{SE<5.5} < d CE_{SE5.5 \leq x < 6.5} < d CE_{SE6.5 \leq x < 7.5} < d CE_{SE \geq 7.5}$$

2. Groups with a higher SE mark will have a lower effect size (d) than groups with a lower SE mark based on the difference in SE minus CE marks in 2011 ( $R_1$ ) and SE minus CE marks in 2012 ( $R_2$ )

$$H_{01} d SE-CE_{SE<5.5} = d SE-CE_{SE5.5 \leq x < 6.5} = d SE-CE_{SE6.5 \leq x < 7.5} = d SE-CE_{SE \geq 7.5}$$

$$H_{11} d SE-CE_{SE<5.5} > d SE-CE_{SE5.5 \leq x < 6.5} > d SE-CE_{SE6.5 \leq x < 7.5} > d SE-CE_{SE \geq 7.5}$$

Table 12

*A summary of the effect sizes about the differences between 2011 and 2012 on CE and SE-CE*

<u>Subject</u>	<u>CE</u>				<u>SE-CE</u>			
	<u>&lt;5.5</u>	<u>&lt;6.5</u>	<u>&lt;7.5</u>	<u>=&gt;7.5</u>	<u>&lt;5.5</u>	<u>&lt;6.5</u>	<u>&lt;7.5</u>	<u>=&gt;7.5</u>
Biology	-0.09	-0.09	-0.05	0.00	0.13	0.09	0.06	0.01
Chemistry	0.17	0.25	0.22	0.10	-0.18	-0.24	-0.23	-0.13
Dutch Language	0.00	0.03	0.13	0.29	-0.03	-0.02	-0.16	-0.30
Economics	0.09	0.16	0.14	-0.07	-0.10	-0.16	-0.14	0.07
English Language	0.07	0.11	0.23	0.34	-0.07	-0.13	-0.24	-0.37
French language	0.30	0.22	0.26	0.38	-0.30	-0.22	-0.27	-0.48
Geography	0.00	0.13	0.09	-0.05	0.03	-0.14	-0.09	0.04
German Language	0.14	0.08	0.16	0.28	-0.13	-0.09	-0.16	-0.33
History	0.25	0.36	0.39	0.37	-0.25	-0.38	-0.42	-0.40
M & O	-0.11	-0.14	-0.07	-0.12	0.13	0.14	0.07	0.14
Mathematics A	0.08	0.14	0.15	0.03	-0.03	-0.13	-0.15	-0.04
Mathematics B	0.12	0.12	0.16	0.15	-0.11	-0.12	-0.15	-0.19
Mathematics C	0.15	0.15	0.07	0.02	-0.15	-0.15	-0.04	-0.20
Physics	-0.05	-0.02	0.08	0.08	0.08	0.02	-0.07	-0.10
All subjects	0.08	0.20	0.28	0.27	-0.09	-0.21	-0.30	-0.38

*Hypothesis 9.*

$H_0$  is can be discarded for four subjects: English language, Dutch language, biology and German language.  $H_0$  is nearly discarded for four subjects: French Language, history and physics, mathematics A. Only one of four categories differs from  $H_1$  for these four subjects.

$H_0$  is not discarded for the remaining six subjects: mathematics B, mathematics C, economics, geography, chemistry and management and organization.

The effect sizes of a students' average CE marks of all the students' subjects is displayed in the last row.  $H_0$  can almost be discarded: the effect size of students with a higher SE mark than a 7.5 is just a bit too low. However, the general line expectation can be observed: the effect size of students having a high SE is higher than the effect size of students higher a lower SE.

*Hypothesis 10.*

$H_0$  is discarded for five subjects: English language, Dutch language, biology, Mathematics B and physics.  $H_0$  is nearly discarded for six subjects: mathematics A, German language, French language, history, chemistry and management and organization. Only one of four categories differs from  $H_1$  for these subjects.  $H_1$  is rejected for three subjects: mathematics C, economy and geography. The effect sizes of a students' average SE-CE marks of all the students' subjects is displayed in the last row.  $H_1$  is approved here.

***Fourth set of hypotheses (eleven and twelve).***

The following two hypotheses make up the final set of hypotheses and are tested using multilevel analyses. The percentages of variation which can be attributed to the school are computed for differences between two years: 2011 and 2012. Hypothesis 11 looks into differences in CE marks between these years and hypothesis 12 looks into differences between SE-CE marks between these years.

*Hypothesis 11.*

The percentage of variation in students' CE marks attributed to the school should be smaller in 2012 than in 2011.

1.  $H_0: ICC\ CE_{school}R_2 = ICC\ CE_{school}R_1$   
 $H_1: ICC\ CE_{school}R_2 < ICC\ CE_{school}R_1$

The results of the analyses using the multilevel model analyzing differences between CE marks can be seen in table 13.

Table 13

*Variances of residuals and intercepts of the multilevel model of CE marks corrected for school influence.*

Subject	Zero model 2011				Zero model 2012				Difference
	residual	intercept	Percentages	school	residual	intercept	Percentages	school	
Biology	1.209	0.099	92.45%	7.55%	1.211	0.085	93.47%	6.53%	-1.02%
Chemistry	1.277	0.123	91.24%	8.76%	1.175	0.101	92.08%	7.92%	-0.85%
Dutch	0.821	0.059	93.29%	6.71%	0.929	0.075	92.55%	7.45%	0.73%
Economics	1.435	0.150	90.55%	9.45%	1.274	0.125	91.07%	8.93%	-0.53%
English	1.159	0.068	94.46%	5.54%	1.380	0.082	94.42%	5.58%	0.03%
French	1.320	0.196	87.04%	12.96%	1.553	0.206	88.30%	11.70%	-1.25%
Geography	0.833	0.123	87.18%	12.82%	0.730	0.122	85.70%	14.30%	1.48%
German	1.082	0.133	89.07%	10.93%	1.223	0.140	89.72%	10.28%	-0.66%
History	0.969	0.127	88.39%	11.61%	0.939	0.101	90.26%	9.74%	-1.86%
M & O	1.348	0.164	89.12%	10.88%	1.272	0.109	92.13%	7.87%	-3.01%
Mathematics A	1.326	0.136	90.70%	9.30%	1.239	0.115	91.54%	8.46%	-0.84%
Mathematics B	2.430	0.302	88.94%	11.06%	2.373	0.215	91.69%	8.31%	-2.75%
Mathematics C	1.513	0.099	93.87%	6.13%	1.352	0.156	89.66%	10.34%	4.21%
Physics	1.319	0.124	91.38%	8.62%	1.451	0.094	93.94%	6.06%	-2.56%
All subjects	0.610	0.048	92.70%	7.30%	0.628	0.041	93.86%	6.14%	-1.16%

The zero models of each year are compared. The school is a random intercept in both models.  $H_0$  cannot be discarded for four of the twelve subjects: English language, Dutch language, geography and mathematics C. The percentage of variation between 2011 and 2012 on students' marks which can be attributed to the school is bigger in 2012 than in 2011 for these subjects. The other eight subjects behave like expected, showing a smaller percentage of variation in students' marks in 2012 in relation to 2011. The last row in table 13 shows the analysis done on all subjects.  $H_0$  can be discarded for this analysis.

About ten till fifteen percent of the variance in student achievement scores relates to difference between school means according to Scheerens & Bosker (1997). The data of this analyses does not comply with the earlier research. Only six out of fourteen subjects do score higher than ten percent variation attributed to school level in 2011 and only four subjects in 2012.

#### *Hypothesis 12.*

The percentage of variation in students' SE-CE marks attributed to the school should be smaller in 2012 than in 2011.

1.  $H_0: ICC\ SE-CE_{school}R_2 = ICC\ SE-CE_{school}R_1$   
 $H_1: ICC\ SE-CE_{school}R_2 < ICC\ SE-CE_{school}R_1$

The results of the analyses using the multilevel model analyzing differences between CE marks can be seen in table 14.

Table 14

*Variances of residuals and intercepts of the multilevel model of SE-CE marks corrected for school influence.*

Subject	Zero model 2011			Zero model 2012			Difference		
	residual	intercept	Percentages	residual	intercept	Percentages			
		school	Residual	school		Residual	school		
Biology	0.673	0.118	85.07%	14.93%	0.641	0.089	87.76%	12.24%	-2.70%
Chemistry	0.639	0.139	82.16%	17.84%	0.605	0.119	83.57%	16.43%	-1.41%
Dutch	0.826	0.084	90.76%	9.24%	0.838	0.087	90.57%	9.43%	0.18%
Economics	0.784	0.157	83.31%	16.69%	0.728	0.127	85.13%	14.87%	-1.82%
English	0.690	0.081	89.55%	10.45%	0.796	0.079	91.02%	8.98%	-1.47%
French	0.853	0.182	82.38%	17.62%	1.015	0.193	84.04%	15.96%	-1.66%
Geography	0.536	0.140	79.24%	20.76%	0.473	0.134	77.89%	22.11%	1.35%
German	0.964	0.128	88.32%	11.68%	1.000	0.135	88.12%	11.88%	0.20%
History	0.618	0.155	79.99%	20.01%	0.590	0.138	81.03%	18.97%	-1.03%
M & O	0.763	0.211	78.32%	21.68%	0.714	0.140	83.63%	16.37%	-5.31%
Mathematics A	0.855	0.156	84.58%	15.42%	0.814	0.131	86.11%	13.89%	-1.52%
Mathematics B	1.018	0.299	77.29%	22.71%	0.939	0.218	81.15%	18.85%	-3.86%
Mathematics C	1.088	0.143	88.38%	11.62%	1.018	0.177	85.21%	14.79%	3.16%
Physics	0.611	0.147	80.61%	19.39%	0.668	0.109	85.99%	14.01%	-5.38%
All subjects	0.226	0.037	85.96%	14.04%	0.219	0.026	89.31%	10.69%	-3.35%

The zero models of each year are compared. The school is a random intercept in both models.  $H_0$  cannot be discarded for four of the twelve subjects: Dutch language, geography, German language and mathematics C. The percentage of variation between 2011 and 212 on students' marks which can be attributed to the school is bigger in 2012 than in 2011 for these subjects. The other eight subjects behave like expected, showing a smaller percentage of variation in students' marks in 2012 in relation to 2011. The last row in table 14 shows the analysis done on all subjects.  $H_0$  can be confirmed for this analysis.

About ten till fifteen percent of the variance in student achievement scores relates to difference between school means according to Scheerens & Bosker (1997). The data of these analyses do partly comply with this research for individual subjects in 2012. The percentage of variation attributed to the school for individual subjects is a lot higher in 2011. The analysis done on all subjects shows that 14.04 percent of variation between SE-SE marks in 2011 were attributed to the school. This complies with Scheerens & Bosker (1997). It is very interesting that the percentage of variation attributes to the school for SE-CE marks on all subjects is lower in 2012 than in 2011. This discards  $H_0$ .

The difference between the results in table 13 and the results in table 14 are interesting. The analyses in table 14 do take into account the average SE marks of schools. The amount of strategic behavior attributed to the school is higher in table 14 than in table 13. This higher percentage indicates that schools demonstrate strategic behavior by making their school examinations easier than the central examinations.

## **Conclusions and discussion.**

Discrepancies between students' marks on school examinations and central examinations in Dutch secondary educations are often attributed to strategic behavior of schools or are attributed to strategic behavior of students. Strategic behavior of schools is researched in much literature. Strategic behavior of students as explanation has thus far hardly been considered as an alternative explanation. Literature did give clues about the possibility of strategic behavior of students influencing their central examination marks. Factors effecting students behavior regarding decisions about central examinations are the following: the first category is prior achievement with as factors students' SE mark per subject and the students' locus of control and the students' prior effort. The second category is the students' expected output of effort and the preference for action over inaction. The third and final category is the support a student receives. The factors of this category are parental support, school support and parental background.

A big limitation of this study is the amount of factors which can be measured. Only one of the factors mentioned above can be measured directly: the students' school examination marks. A change in rules and regulations regarding graduating VWO made it possible to do research on the possible supplementary explanation of strategic behavior of students, being responsible for Discrepancies between students' marks on school examinations and central examinations in Dutch secondary educations. The factors found in literature mentioned above are used to predict how the introduction of new rules and regulations would change the relation between school examination marks and central examination marks between the 2011 and 2012. The expected patterns are explained and expressed as hypotheses and are tested.

### ***Sub questions.***

The first set of hypotheses, hypotheses one till four, looked into changing data patterns in CE marks in 2011 and CE marks in 2012. The second set of hypotheses, hypotheses five till eight, looked into changing data patterns in SE-CE marks and SE-CE marks in 2011 and 2012. Students were grouped into four categories: SE mark below 5.5; SE mark higher than or equal to 5.5 and lower than 6.5; SE mark higher than or equal to 6.5 and 7.5 and SE mark equal to or higher than 7.5. The expectation is that students with the same SE marks will score higher on their CE in 2012 than in 2011. Hypotheses nine and ten are about differences in between the four groups of students categorized by their SE marks. Hypotheses one till ten will together answer sub question one:

Is there a difference in level and variation between school examination marks and central examination marks per subject after the rules and regulation change?

The results of the analyses done for hypotheses one till eight show a pattern in line with the hypothesis: CE marks of students with the same SE marks were higher in 2012 than they were 2011. The patterns are most clear in the analyses done on SE-CE marks. The distribution of SE marks is taken into account in these analyses. Ten out of fourteen subjects do show the expected pattern in the group of students having a subject SE mark lower than a 5.5. Eleven out of fourteen subjects to show the expected pattern in the group of students having a subjects SE mark equal to or higher than a 5.5 and lower than a 6.5. The three subjects not showing this pattern are the same subjects in the previous group: biology, physics and management and organization. Twelve out of fourteen subjects to show the expected pattern in the group of students having a subjects SE mark equal to or higher than a 5.5 and lower than a 6.5. The Two subjects not showing this pattern are again biology and management and organization. In the last category are three subjects not showing the expected pattern: again biology and economics and newly geography.

The expected pattern is seen when not individual subjects, but average marks of students are compared. Students with the same SE perform better in 2012 than in 2011.

The results of the tests for hypotheses one till eight are used to test hypothesis nine and ten. Hypotheses nine and ten together form the third set of hypotheses. The hypothesis is that students who



have a higher SE mark will still need to get a high CE mark in 2012. This was not needed in 2011 because low CE marks could be compensated by high SE marks. The trend shown by the results of the analyses for hypotheses nine and ten are clear. Students having a high SE mark do score higher CE marks in 2012 than in 2011. Most effect sizes are small according to Cohen (1988). The strength of this research is that there is no sampling: data of all students enrolled in VWO in the years 2011 and 2012 was available and therefore used.

The results of the analyses done to test hypotheses one till ten are used to answer sub question one. The answer to sub question one is that there is a difference in level and variation between school examination marks and central examination marks per subject after the change in rules and regulations.

Multilevel analyses are used to test hypotheses eleven and twelve. These two hypotheses together form the fourth set of hypotheses. Testing hypotheses eleven and twelve made it possible to answer sub question two:

Is there a difference between school examination marks and central examination marks per subject caused by variation in school level and by variation on student level after the rules and regulations change?

Changes in the amount of variance which can be attributed to the school in students CE results did become lower in 2012 than in 2011. The most interesting result is the change in the amount of variance which can be attributed to the school in students SE-CE marks. The results indicate that there is a difference between 2012 and 2011. The amount of variance in students SE-CE marks which can be attributed to the school is lower in 2012 than in 2011 for ten out of fourteen subjects. These four subjects are English Language, Dutch Language, geography and mathematics C. The difference between 2011 and 2012 for English language is almost negligible. The analyses done on all subjects of a student's shows a clear picture: 14.04 percent of variation in students SE-CE marks could be attributed to the school in 2011 and 10.69 percent in 2012.

Finding strategic behavior of schools was not a goal of this study. The results in table 13 and table 14 did however found evidence of strategic behavior of schools

The results of these zero-models indicates that the hypotheses formed about the answer of sub question two do not have to be discarded: there is a difference between school examination marks and central examination marks per subject caused by variation in school level after the rules and regulations change.

### ***Main research questions.***

The main research question is:

Can discrepancies between results on school exams and central exams be explained by strategic behavior of students?

It is possible to answer this question after answering the two sub questions. The answer to the main research question is: yes, discrepancies between results on school examinations and central examinations can be partly explained by strategic behavior of students. A change in rules and regulations provided an opportunity which made it possible to derive strategic behavior of student by observing a change in patterns of relations between SE marks and CE marks of students.

This opportunity makes it possible to bypass the limit of not being able to measure students' strategic behavior directly. The CE marks of students scoring high on their SE were higher in 2012 than in 2011. This change is remarkable and interesting because schools tend to have the same amount of difference between SE and CE marks every year (De Lange, M., & Dronkers, 2007).

The evidence of strategic behavior of student is indirect, but the nature of this research and the tested hypotheses will make it difficult to come up with an alternative explanation for the observed patterns.

Strategic behavior regarding their approach to their central examinations seems the most plausible explanation. Comparing the zero-models for 2011 and 2012 found less variance in SE-CE results being attributed to the school in 2012 than in 2011. It should be noted that this is stated as a percentage. It is possible that a lower percentage of variance in students' results could be attributed to the school, if a school put in more effort (strategic behavior) in getting higher CE marks for their students and students put in relatively more extra effort (strategic behavior).

*Alternative explanations and recommendations.*

Other explanations for a change in patterns could be that students' motivation was higher for other reasons. An example could be the amount of student grants available for new students in college. However, this change in patterns would still be strategic behavior, but for a different reason.

This thesis compared and found a difference between the years of 2011 and 2012. There is a possibility that the cause of this difference has another reason than strategic behavior, even though the difference is significant according to the method used in this thesis. An interrupted time series design could be used to be more certain about the outcomes of this study. Several years before 2011 should be analyzed to be sure that the difference between 2011 and 2012 is significant when compared to fluctuations between other years in the past. Data after 2012 should be added to the interrupted time series design and be analyzed when this data is available.

Evidence found for strategic behavior of students did result in higher CE marks for students in 2012 compared to students with comparable SE marks in 2011. This pattern found makes the difference between a school's SE and CE mark smaller. The most plausible explanation for this smaller difference is a change in rules and regulations, resulting in strategic behavior of students causing higher CE marks. This behavior of students gives schools a better score on one of the indicators used by the Dutch Inspectorate of Education. It gets easier for a school to meet the standard of the indicator used by the Dutch Inspectorate of Education without exerting extra effort. The strategic behavior of students found in this study makes the three year rolling average of the SE-CE marks indicator used by the Dutch Inspectorate of Education more valid. Therefore, the Dutch Inspectorate of Education could make this indicator more important.

## Bibliography.

- Bishop, J. H. (2005). *High School Exit Examinations : When Do Learning Effects Generalize ? High School Exit Examinations : When Do Learning Effects Generalize ?*. Ithaca, NY. Retrieved from <http://digitalcommons.ilr.cornell.edu/cahrswp/4>
- Boon, C., Olffen, W. Van, & Roijackers, N. (2004). Selection on the Road to a Career : Evidence of Personality Sorting. *Journal of Career Development*, 31(1), 61–78.
- Chater, N., & Oaksford, M. (2000). The Rational Analysis Of Mind And Behavior. *Synthese*, 122(1-2), 93–131. doi:10.1023/A:1005272027245
- Cohen, J. (New Y. U. (1988). *Statistical power analysis for the behavioral sciences* (second edi., pp. 273–406). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.
- De Lange, M., & Dronkers, J. (2007). *Faculteit der Sociale Wetenschappen Departement van Politieke en Sociale Wetenschappen Groeide de ongelijkwaardigheid van het eindexamen tussen scholen verder in 2005 ? Discrepanties tussen de cijfers voor het schoolonderzoek en het centraal examen in het*.
- Dekay, M. L., & Patin, D. (2009). Better Safe than Sorry : Precautionary Reasoning and Implied Dominance in Risky Decisions. *Journal of Behavioral Decision Making*, 361(March), 338–361. doi:10.1002/bdm
- Dermitzaki, I., Leondari, A., & Goudas, M. (2009). Relations between young students' strategic behaviours, domain-specific self-concept, and performance in a problem-solving situation. *Learning and Instruction*, 19(2), 144–157. doi:10.1016/j.learninstruc.2008.03.002
- Dollinger, S. J., & Clark, M. H. (2012). Test-Taking Strategy as a Mediator between Race and Academic Performance. *Learning and Individual Differences*, 22(4), 511–517. doi:10.1016/j.lindif.2012.03.010
- Eastwood, J., Snook, B., & Luther, K. (2012). What People Want From Their Professionals : Attitudes Toward Decision-making Strategies. *Journal of Behavioral Decision Making*, 468(June 2011), 458–468. doi:10.1002/bdm
- Inspectie van het Onderwijs. (2008). *De staat van het onderwijs. Onderwijsverslag 2007/2008*. Utrecht. Retrieved from <http://www.onderwijsinspectie.nl/binaries/content/assets/Onderwijsverslagen/2009/Onderwijsverslag+2007-2008.pdf>
- Inspectie van het Onderwijs. (2014). *De staat van het onderwijs. Onderwijsverslag 2012/2013*. Utrecht. Retrieved from <http://www.onderwijsinspectie.nl/binaries/content/assets/Onderwijsverslagen/2014/onderwijsverslag-2012-2013.pdf>

- Jones, B. D. (2008). Journal of Applied School The Unintended Outcomes of High-Stakes Testing The Unintended Outcomes of High-Stakes Testing. *Journal of Applied School Psychology*, (02 October 2008), 37–41. doi:10.1300/J370v23n02
- Jürges, H., & Schneider, K. (2009). Central exit examinations increase performance... but take the fun out of mathematics. *Journal of Population Economics*, 23(2), 497–517. doi:10.1007/s00148-008-0234-3
- Jürges, H., Schneider, K., Senkbeil, M., & Carstensen, C. H. (2012). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, 31(1), 56–65. doi:10.1016/j.econedurev.2011.08.007
- Kennisnet, & CvTE. (2013). Normering bij de centrale examens in het voortgezet onderwijs.
- Kooreman, P. (2012). Rational students and resit exams. *Economics Letters*, 118(1), 213–215. doi:10.1016/j.econlet.2012.10.015
- Laborde, S., Dosseville, F., & Scelles, N. (2010). Trait emotional intelligence and preference for intuition and deliberation: Respective influence on academic performance. *Personality and Individual Differences*, 49(7), 784–788. doi:10.1016/j.paid.2010.06.031
- Luyten, H., & Dolkar, D. (2010). School-based assessments in high- stakes examinations in Bhutan : a question of trust? Exploring inconsistencies between external exam scores , school-based assessments. *Educational Research and Evaluation*, (December 2010), 37–41.
- Neill, J. (2008). Why use Effect Sizes instead of Significance Testing in Program Evaluation. Retrieved from <http://www.wilderdom.com/research/effectsizes.html>
- Oerke, B., Maag Merki, K., Holmeier, M., & Jäger, D. J. (2011). Changes in student attributions due to the implementation of central exit exams. *Educational Assessment, Evaluation and Accountability*, 23(3), 223–241. doi:10.1007/s11092-011-9121-7
- Rijksoverheid. (2013). Wanneer ben ik geslaagd voor het eindexamen havo of vwo? Retrieved from <http://www.rijksoverheid.nl/onderwerpen/voortgezet-onderwijs/vraag-en-antwoord/wanneer-ben-ik-geslaagd-voor-het-eindexamen-havo-of-vwo.html>
- Scheerens, J., & Bosker, R. J. (1997). Scheerens, Jaap, and Bosker, Roel J. 1997. The Foundations of Educational Effectiveness. *International Review of Education*, 45(1), 113–120. doi:10.1023/A:1003534107087
- Schildkamp, K., Rekers-mombarg, L. T. M., & Harms, T. J. (2012). Student group differences in examination results and utilization for policy and school development. *School Effectiveness and School Improvement*, (April 2012), 37–41.
- Schneider, K. (2003). The effect of central exit examinations on student achievement: Quasi-experimental evidence from timss Germany.
- Vreeburg, B. A. N. M. (2007). *Paper ORD: Leerresultaten en risicogestuurd toezicht in het voortgezet onderwijs. Validiteit en betrouwbaarheid van kengetallen voor de bepaling van het verschil tussen schoolexamen en centraal examen*. Utrecht.
- Vreeburg, B. A. N. M., Theunissen, G., & Coenen, A. (2010). *Internal report about differences between SE and CE marks*. Utrecht.

Wößmann, L. (2005). The effect heterogeneity of central examinations: evidence from TIMSS, TIMSS- Repeat and PISA. *Education Economics*, 13(2), 143–169.  
doi:10.1080/09645290500031165

## Appendix A.

The subjects that will be analyzed are shown in table A1 below.

*Table A1. Subjects to be analyzed including the amount of students enrolled in that subject*

Subject	N
Dutch language	73422
English language	73417
Mathematics (A,B,C aggregated)	73453
French language	31674
German language	37500
History	36400
Geography	22145
Science	37335
Chemistry	39857
Biology	37073
Economics	39276
Management and organization	21296