STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's thesis in economics
Academic year 2016–2017

Underinvestment in Education: The Effects of Grades on Student Motivation and Performance

Arash Aslfallah (22795)

Abstract:
This paper examines the effect of grades on student motivation and performance. As such, the purpose of this paper is twofold. First, it helps to explain how the current educational model, by its assessment of students through grades, affects student motivation and performance, which is fundamental for understanding the underinvestment problem in education. Second, this paper test whether incorporating the concept of identity economics in economic decision making, can better predict student behaviour than the standard models of education. This was done by a randomized field experiment being conducted on 372 sixth graders in Sweden, where the treatments being evaluated were designed to investigate the effect of grades, when it serves as a starting point on which students' future performances and outcomes are dependent on. The results indicate a negative average causal effect of grades on student motivation and performance. This is the case regardless of whether the grading system in place is criterion-based or norm-based. The negative effect of grades also differs across the grade distribution, and with respect to gender, where students with lower grades are affected more negatively, and where girls seem to be affected more than boys. Thus, the result of this paper call into question, the role and application of grades, in the current educational model. Furthermore, this paper suggests that the concept of identity economics, as incorporated in the educational context, can in a better way predict student behaviour than standard models of education.

Keywords: Extrinsic and intrinsic motivation, Randomized experiment, Identity economics, Behavioural economics of education

JEL: I20, I21, D03, C93

Supervisor:	Martina Björkman Nyqvist
Date submitted:	16 May 2017
Date examined:	29 May 2017
Discussant:	Petter Svärd
Examiner:	Anders Olofsgård

Acknowledgements

I would like to thank Martina Björkman Nyqvist and Anna Dreber Almenberg for their incredible support and guidance, during both the experimental phase and the writing process. I also want to thank all the teachers and students who participated in this study. Without you, this study would have not been possible. Finally, I would also like to thank my family and friends for their support.

Contents

1. Introduction

One of the most important economic decisions most of us face during our life, is how much to invest in education, both in terms of effort and time. Despite this, observations of decisions and outcomes in regards to investments in education in the real world, seem to contradict the standard economic theories that have been developed so far. These standard models of education, fail to explain why a relatively large proportion of students, drop out of school just when the returns to education seem to be at their maximum (Heckman et al., 2006). As such, the design of the school system is an important national topic, nonetheless in Sweden. Much of the educational debate in Sweden concerns the role and application of grades, which is often argued to motivate students, but also to measure and indicate student performance and ability. Many people advocate an earlier introduction of grades amongst students, i.e. earlier than the current introduction in the 6th grade. However, if grades do not motivate students to exert effort in school, its purpose of indicating student performance and ability as such, becomes insignificant. It is therefore important to understand the effect of grades on student motivation and performance, and to identify through which mechanisms these effects channel through. This is particularly important among young children. If the choice of how much effort to invest in education is affected negatively by grades, it may have long-term effects on subsequent outcomes, considering that learning is cumulative.

So far, previous research investigating the effect of absolute grading (criterion-based grades) on student motivation and performance have shown no general results, but have, however, shown that grades have a differentiating effect, where low-to-medium-performing students are affected negatively (Harlen and Deakin Crick, 2002; Jalava et al., 2015; Klapp et al., 2016). In contrast, previous research investigating the effect of norm-based grading (relative rankings) on motivation and performance, have shown general positive effects, but where low-performing students, yet again, have shown to be negatively affected (Azmat and Iriberri, 2010). As such, this paper aims to build upon a growing body of research, to further study the effect of grades on motivation and performance. The purpose for this paper is thus twofold. First, understanding how the current school system, by its assessment of students with the help of grades, affect student motivation and performance, is fundamental for understanding the underinvestment problem in education and for designing effective school policies. Second, I test whether a standard model of education incorporating the concept of identity economics, is better at predicting student behaviour. By doing this, I hope to combine aspects of behavioural economics and identity economics in order to provide a new unique perspective to the economics of education.

Previous experimental papers have mainly investigated the short-term effect of grades on performance looking forward. The main idea of this paper is, however, to reverse this process and investigate the effect of grades, when it serves as a starting point on which students' future performances and outcomes are dependent on. This was accomplished by a field experiment being conducted on 372 sixth graders

in Sweden. Briefly, the experiment randomly assigned students in each class, to three groups (control, treatment 1, treatment 2), where they subsequently were given a math test to solve during 10 minutes, and to fill in a survey afterwards. Students in all three groups received two rewards if they performed adequately, as specified in their instructions. The treatments as such, were that the instruction on the first page of the test, concerning how their performance was assessed, differed among the three groups. The instructions for each group were: (Control group) students were informed that if achieving a test score equal to or above a pre-determined threshold, they would receive the rewards, (Treatment group 1) in addition to the instructions per the control group, students were informed that their previous grade would affect their outcome, (Treatment group 2) students were also in this case informed that their previous grade would affect their outcome, but that instead of a pre-determined threshold, only the top three performing students would receive the rewards. This additional treatment group was included as to assess the effect of previous grades, if the grading system were to be norm-based. These treatments were designed to simulate the current educational system in Sweden. They also served the purpose, of assigning students in the treatment groups, to social categories corresponding with their previous grades. This aspect of the treatment, is linked to the student utility model as specified in this paper, that incorporates the concept of identity economics. In short, this model implies that there are different social groups with corresponding prescriptions, that dictate the ideal characteristics and behaviour of its members. As such, the different behaviour of individuals from different social categories, can be predicted from the prescriptions (Akerlof and Kranton, 2002).

I find that the average causal effect of grades, when serving as a starting point, on which a student's future performance is dependent on, is negative on student motivation and performance. This is the case regardless of whether the grading system in place is criterion-based or norm-based. The negative effect of grades also differs with respect to gender, where girls seem to be affected to a larger extent than boys, although this difference is not statistically significant. In addition to this, the average causal effect of the treatments is also higher, the lower grade the student has. Combined with the results of previous research, indicating no general effect of grades as a non-monetary incentive, the results of this paper shed further doubt on the application of grades in schools (Jalava et al., 2015; Klapp et al., 2016). The use of grades in school could also potentially lead to differentiating effects on girls and/or low-performing students, leading to an increased gap between students, and as such, increased inequality.

Furthermore, the mechanisms as identified in this paper, that the effect of grades could potentially channel through, are student self-confidence, that grades can constitute social identities affecting behaviour, and the notion that grades with their current design unfairly assess student performance. Subsequently, I argue that, due to the potential negative effect of grades on self-confidence, and as such on intrinsic motivation, both which are increasingly stable over time, the negative effect of grades on student motivation and performance, can constitute a long-term one. More importantly, this paper

suggests that the concept of identity economics, as incorporated in the educational context, can in a better way predict student behaviour than standard models of education.

This paper is organized as follows: Section 2 covers the previous literature on the behavioural economics of education. Section 3 covers the conceptual framework of which, the experiment of this paper is based upon, as well as the concept of identity economics as incorporated in this paper. Section 4 contains the methodology of the paper, and as such gives a detailed description of the experimental design, as well as its empirical implementation. Section 5 presents the data collected through the experiment. Thereafter, section 6 presents the empirical results, which are then discussed in section 7. Lastly, section 7 presents concluding remarks.

2. Literature Review

One of the most important economic decisions most of us face in this modern time during the course of our life is how much to invest in education. Not only does education improve life-time earnings, it also improves health, reduces crime and increases voting and democratic participation (Heckman et al., 2006; Lochner, 2011). Despite this, observations of decisions and outcomes in regards to investments in education in the real world seem to contradict the standard economic theories that have been developed so far. The standard human capital models fail to explain why a relatively large proportion of students drop out of school just when the returns to education seem to be at their maximum, or why girls seem to avoid entering competitive settings or avoid math classes altogether, resulting in an overall underperformance in math when the future returns are large (Heckman et al., 2006; Niederle and Vesterlund, 2010; Oreopoulos, 2007). One path to understand these observations and to develop alternative theories that might explain investment decisions in education better is to delve into the emerging field of behavioural economics. In this section, I will briefly review the emerging literature on the behavioural economics of education with a focus on intrinsic and extrinsic motivation. I will conclude with what I hope will be the contribution of this paper to the ongoing development of this specific area.

2.1 Behavioural Economics of Education

Standard models of economics tend to oversimplify reality and often make strong assumptions, such as the fact that individuals are entirely rational and act accordingly to maximize their lifetime welfare. In contrast, behavioural economics attempts to integrate and incorporate insights from psychology, sociology, and neuroscience into standard economic theory to better predict and understand human behaviour. This line of research is often complemented by experimental economics in order to capture causal effect and map actual behaviour. Behavioural economics suggests that individuals do not act rationally all the time, but suffer from a variety of different non-standard preferences and beliefs and as such engage in non-standard decisions making. A few examples are time-inconsistent preferences,

reference dependent preferences, and the effect of framing of choices on decision making (DellaVigna, 2009).

Since its inception, behavioural economics has been successfully applied to a wide range of areas. One area, however, which has so far received less attention, is education. This is surprising since the insights gained from behavioural economics could be especially valuable given the interest in long-run decision making and the propensity for the youth to make poor ones (Lavecchia et al., 2014). As such, there is a strong need to incorporate concepts from behavioural economics into economics of education to better understand educational outcomes. For example, we need reference to time-inconsistent preferences to understand why people underinvest in education. In addition to that and what will be the focus of the rest of this literature review, we need references to behavioural theories of motivation and self-confidence to understand why people underinvest in education.

2.2 Incentives – Intrinsic and Extrinsic Motivation

One possible solution to underinvestment problems is the provision of adequate incentives for educational attainment and performance. Before we dig into greater detail of what form these incentives could take, we need to establish a distinction between intrinsic and extrinsic motivation because of their fundamental difference, especially in an educational setting. Motivation that is coming from within the students themselves, driven by an interest and enjoyment of the task itself is what is referred to as intrinsic motivation. In contrast, extrinsic motivation relies on, and is driven by external forces.

2.2.1 Extrinsic Motivation

Extrinsic motivation relies on, and is driven by, external forces that come in many different forms. In the field of education, two mutually exclusive and collectively exhaustive extrinsic incentives are monetary and non-monetary rewards. I will first briefly discuss the effects of monetary rewards on educational achievement before I turn to the effects of non-monetary rewards where I will focus in more detail on the role of grades and ranking on performance and motivation.

Monetary and Non-Monetary Rewards

The effects of monetary rewards as incentives have been widely studied in a wide range of areas and nonetheless in educational outcomes. Several experiments have been conducted analysing the effect of payments on motivation, finding positive results on educational performance (Bettinger and Slonim, 2007; Bettinger, 2012; Eisenkopf et al., 2015; Fryer Jr, 2010; Levitt et al., 2016). While the research on monetary rewards is more extensive, there is a growing area of research focused on the implementation of non-monetary rewards demonstrating that their effects can be considerable too (Ashraf et al., 2014; Frey, 2007; Jalava et al., 2015; Kosfeld and Neckermann, 2011; Levitt et al., 2016).

The effects of non-monetary rewards on educational performance could operate through a range of possible mechanism, such as self-image and status concerns or relative performance feedback. The form

that non-monetary rewards take can vary broadly from rewards, such as awards and trophies, to grades and ranking. Awards and trophies have, for instance, been shown to have a significant effect on motivation in the workplace as they arguably yield non-material benefits in the form of status and improved self-esteem (Ellingsen and Johannesson, 2007; Kosfeld and Neckermann, 2011; Weiss and Fershtman, 1998). Levitt et al. (2016) further investigates the effect of non-monetary rewards as it directly compares it to the effects of monetary rewards on student performance in a school setting. In a large-scale field experiment, Levitt et al. (2016) explore the power of behavioural economics to influence the level of effort students exert in a low-stake test by introducing key concepts and ideas from the field such as loss aversion, non-monetary rewards, and hyperbolic discounting. They find that both high monetary incentives ($20) and non-monetary rewards (in this case a trophy) improve test performance, while low monetary incentives ($10) do not. However, this substantial impact on test scores is only noticeable when rewards are delivered immediately and completely gone when rewards are delivered with a delay.

Grades and Ranking

Although not comparable to, for instance, a trophy as a non-monetary reward, grades and rank could arguably constitute and function as an extrinsic motivational mechanism and as such motivate students, improving their performance. This could, for example, be due to the impact of grades on future educational and labour market outcomes, especially when students approach higher levels of education. To get into competitive programs and universities, one needs to obtain high grades and this becomes more important with seniority. Grades and ranking can, however, operate through other mechanisms that also increase, or in some cases reduce, the motivation of students. This is where educational psychology and in recent years, behavioural economics, have played their part. Grades can for instance affect students' status, self-image and self-confidence (Koch et al., 2015).

Harlen and Deakin Crick (2002) survey studies that examine the role of absolute grading (criterion-references assessment) on the motivation of students to learn. Firstly, they conclude that low-performing students' self-esteem is negatively affected by absolute grading, that students do not like absolute gradings, and that they develop more superficial and performance-oriented strategies to learning. Secondly, they found that low-performing students were disadvantaged twice as absolute gradings labelled them as failures or less able which affected their already low self-esteem. Harlen and Deeakin Crick argue that this turns into a negative spiral where low-performing students, due to their low self-esteem, will put even less effort in school in the future. They argue that the usage of absolute gradings lead to an increased gap between students and increased inequality.

Jalava et al. (2015) not only builds on and complements the paper by Levitt et al. (2016) by introducing several non-monetary rewards (such as grades, relative ranking, prizes and diplomas), but they also incorporate different ways of grading (absolute grading and norm-based grading equating relative

rankings). They do this by conducting a field experiment in Swedish primary schools whereby they examine the effects of their treatments on test scores (math test) among 6th graders. They find that relative to their control group that did not receive any incentives, all their non-monetary rewards except for absolute grading increased performance among students. Furthermore, the design and implementation of their experiment highly inspired the approach and strategy of this paper.

A second paper analysing the effect of absolute grading in Swedish primary schools is a quasi-experimental study by Klapp et al. (2016). Due to the occurrence of a natural experiment between 1969 and 1981, municipalities could themselves decide whether to grade students in the 6th grade or not. This allowed the authors to empirically analyse the effect of grading in the 6th grade on students' results one year later. The authors conclude that there were no general effects of grading on future performance, but that there were differentiating effects. Low-to-medium-performing students received lower grades if they had been subject to grading in the 6th grade. Klapp (2015) conducted a follow-up study where she examined how grades in the 6th grade affected the students' performances in grade 7,8 and 9. She found significant negative effects of grading in the 6th grade on future performance.

Although the effects of grades on motivation and educational performance is inconsistent, the effects of relative rankings among peers by performance seem to produce consistent positive effects. According to previous studies, rank seems to be a major motivational force and has a measurable impact on behaviour (Tran and Zeckhauser, 2012). In the educational setting, rank can operate through a norm-referenced grading where the assigned grades depend on the relative performance of other students. Rank could reward students with tangible and non-tangible benefits through mechanism such as the opportunity to impress others or to increase students' self-esteems. Azmat and Iriberri (2010) find that the provision of relative performance feedback led to an increase in short-term performance across the whole distribution in regards to ability. This mechanism can however work the other way around where low rankings reduce motivation and performance for low-performing students (Jalava et al., 2015; Levitt et al., 2016).

2.2.2 Intrinsic Motivation

Having covered key elements of extrinsic motivation in the educational context, it is now time to turn to intrinsic motivation. Not only is it important to explore further what the key elements are that constitute intrinsic motivation, but we also need to be wary about the possibility, that extrinsic motivation can have a detrimental effect on students' educational achievement and crowd out intrinsic motivation. Moreover, intrinsic motivation can be composed by mechanisms such as curiosity, the joy of learning, self-confidence, and even self-image and identity (Koch et al., 2015). To survey all aspects of intrinsic motivation could be the sole purpose of another paper by itself, hence I will solely survey the literature that in some way relates to the area of education.

Self-confidence

Self-confidence can play a key role in increasing intrinsic motivation. Bénabou and Tirole (2002) were among the first to develop a theoretical model in economics investigating the role of self-confidence. They argue that since ability and effort are complementary to each other in educational performance, an overly positive view of one's ability could be a strong motivational factor. Moreover, they demonstrate that for individuals with self-control problems, it could be optimal to selectively process information and increase one's self-confidence. An increase in self-confidence will make individuals prone to believing that their effort will be more productive resulting in higher motivation.

Furthermore, self-confidence is positively related to academic intrinsic motivation (Gottfried, 1990). This is interesting as further research in educational psychology show that academic intrinsic motivation is a stable construct over time, and increasingly so with advancement in age (Gottfried, 1990; Gottfried et al., 2001). This is rather important, as it places children with low levels of motivation early in their schooling at risk.

Self-confidence in the educational settings can also have surprising effects depending on the school and grading system. Wang and Yang (2003) investigate this relationship in a theoretical paper where students care about both their grades and their own perception of their ability. The grading system in place depending on its function determines how much information a grade conveys about a student's ability. This information affects self-confidence, which later affects the choice of effort induced by the student. In a school where students care more about their perceived ability, a school system with relative grading can lead to low effort across the whole distribution of student abilities. Relative grading is more competitive as it limits the number of good grades which in turn reduces the probability that a student receives good feedback regarding their ability if he or she works hard. A student can then, to protect a positive self-image exert low effort which makes the grades relatively uninformative about ability and allows the student to maintain her positive self-image.

A final theoretical paper around the same area is by Filippin and Paccagnella (2012) where they in a model explore how a small initial difference in self-confidence can result in diverging paths of human capital accumulation. What is interesting is the fact that this is the case even when students start off with the same level of initial ability highlighting the importance of self-confidence.

Identity Economics

Questions regarding identity dominates thinking and behaviour in preadolescence. Questions such as "Who am I?" and "How do other people like me behave?" are powerful reference points for how individuals choose to behave. These questions concerning behaviour can also have significant effects on how much students decide to invest in education. Akerlof and Kranton (2002) argue that students care about the degree to which their individual behaviour differs from the prescribed behaviour of their social grouping. These social groups can be based on sex, race or other categories such as academic

success. Investment in education such as effort exerted in school, are in this context not only dependent on individual benefits such as grades received, but also on social benefits such as whether the level of effort exerted individually, is in line with the behaviour of one's social group. If a social group occupies itself with just having fun and retain from exerting any effort in school, the individual will feel the pressure to do the same.

Students can also identify themselves as failures or as less successful than others. Extensive research by Dweck (2008) show that individuals' beliefs about themselves that they bring to new situations can affect the degree of their learning and performance. Students that believe that most of the factors contributing to ones' success are innate are also more likely to be intimidated by initial failures. Students that in contrast believe that effort matters most, view failures as an indication that they need to spend more time and increase their effort in a task.

2.3 Limitation to existing literature

Koch et al. (2015), in their review of the emerging area of behavioural economics of education, outline some caveats for future research. Firstly, they argue that there exists a lack of use of experimental studies to gain insight into the economics of education. Among the few lab and field experiments that exist, most use convenience samples in a low-stakes environment. In addition to this, tasks in the lab are sometimes artificial in comparison to how decisions are taken in real-world setting. Finally, they argue that results from experimental economics so far have only shed light on short-term effects missing the more important aspect of what the effects will be in the long-term.

2.4 Contributions of this paper

The previous literature covered, motivates and formulates the research question and purpose of this paper. In short, previous studies have shown that absolute grading has no general short-term effect on performance, while it consistently has differentiating effects where low to medium performing students are negatively affected. Norm-based grading (relative rankings) on the other hand could have positive short-term motivational effect on students, but also results in differentiating effects where low-to medium-performing students are negatively affected. At the same time, self-confidence and self-image/social identity at a young age, stipulates important aspects for the intrinsic motivation of students and can have ever lasting long-term effects. This, combined with the propensity of young individuals to make short-term decisions, begs to answer the question of what the effect of grades are on self-confidence and self-image/identity?

Moreover, it might help us understand the long-term effects of grades and rankings as these mechanisms have been shown to have long-term effects on motivation and performance (Dweck, 2008). Thus, one of the aims of this paper is to identify the mechanisms which grades and rankings channel through, which could potentially constitute long-term effects on motivation and performance. Secondly, this paper investigates how the current educational model, by its assessment of students through grades,

affect student motivation and performance, which is fundamental for understanding the underinvestment problem in education and for designing effective school policies. Finally, this paper test whether an adjusted version of the student utility model developed by Akerlof and Kranton (2002), that incorporates the concept of identity economics in educational economic decision making, can better predict student behaviour than the standard models of education.

This paper as such contributes to the literature by filling the gaps identified so far by focusing on the effect of grades on self-confidence, self-image/social identities and their potential long-term effect on motivation and educational performance. It does so by conducting a field experiment on a general sample distribution, simulating how the school system looks and operates in its current form in Sweden. To the best of my knowledge, this is the first paper to directly assess the effect of absolute grading and norm-based grading, when it serves as a starting point on which students' future performances and outcomes are dependent on, and to do so with the use of a general sample distribution.

3. Conceptual Framework & Theory

Education has so far, as indicated in the previous section, received less attention in the domain of behavioural economics. Hence, this paper aims to contribute to this field by combining a set of mechanisms previously shown to have important effects on motivation and performance to further uncover the effect of grades. In this section, I will first cover the conceptual framework upon which the experiment of this paper is based on. Next, I will introduce a theoretical model which explains and predicts how individuals could make decisions (in this case how much effort to exert during a math test) whose predictions I will test in this experiment.

When it comes to the effect of grades, the starting point for previous research has been to investigate its short-term effect on performance looking forward. For example, studies have looked at whether students perform better if being graded compared to having no incentives at all (Jalava et al., 2015). The main idea of this paper is to reverse this process and investigate the effect of grades when it serves as a starting point on which students' future performances and outcomes are dependent on. This, I argue, will capture a different dimension than what has been previously been explored and could help us understand the potential long-term effects of grades on motivation and performance. The long-term effect of grades on motivation and performance is in reality difficult to capture, especially if students are only assessed once as in this paper. I argue, however, that if grades distort the optimal effort-level of students, through mechanisms such as self-confidence, social identity and their notion of unfairness, its long-term effect can be deduced. This due to the long-term effect these mechanisms can have on long-term motivation and performance as discussed earlier. In addition to this, this paper is of relevance for policy reforms as it essentially simulates how the school system assesses students in its current form.

The idea behind this experiment is largely based on how, many schools function and operate around the world, particularly in Sweden. For example, in primary school, the final grade of the semester is a weighted average of 2-3 previous exams. Depending on the outcome of the first exam, a student might change her or her effort accordingly for the final exams.

To illustrate this exactly, assume that grades are in fact an extrinsically motivating factor for students. Assume that a student optimizes her choice of effort per her utility function in the first test that she has ever been graded on. However, after the first initial exam occasion, the student by bad luck receives a low grade. Looking forward, if grades have an impact on the student's notion of what is unfair, on her self-confidence, and furthermore constitute a social identity, it could potentially affect the student's level of effort exerted in the upcoming exams. First, she might still be equally motivated by grades, but considering that the final grade is a weighted average of all previous exams, she might give up due to a lower probability of getting the grade she wants. This translates into a lower level of effort exerted by the student at the next exam occasion. This is something that has been reported amongst teachers in the 6th grade amongst low-performing students, in a report highlighting the implications of introducing grades in the 6th grade (Skolverket, 2017). Here on, I refer to this effect as the *unfairness effect* which means that students find the fact that their previous grade affects their chances to achieve a certain grade/reward at future occasions unfair. Secondly, the student might lose confidence in her own ability, if believing that the grade communicates her true ability. As such, she might believe that it is not possible for her to get a high grade no matter how much effort she exerts. In addition to this, the low grade can label the student, assigning her to a social category of fellow low-ability students. The assignment to this social category can in turn change the student's behaviour according to what she thinks is optimal considering her social identity. The loss of confidence and the new social identity then leads to lower effort being exerted in the upcoming exams. Combined with the notion of unfairness, these effects translate into a negative spiral of reinforcing forces that opt the student to exert low levels of effort, all due to either initial bad luck or low-inherited ability. The second effect however (loss of confidence and new social identity), I argue, could potentially constitute a long-term one as individuals tend to bring to new situations beliefs about themselves that can affect the degree of their learning and performance (Dweck, 2008). This could result in diverging paths of human capital accumulation and an increased gap between students.

In short, the experiment in this paper investigates the following: assuming students are extrinsically incentivized to exert effort, will their effort change if their utility is also dependent on their previous grade simulating how the school system currently operates, through a labelling effect where students are assigned to different grade-categories (even though it might or might not increase the difficulty of attaining the specified reward)? I explore this by testing whether an adjusted version of the student utility function developed by Akerlof and Kranton (2002), that incorporates the concept of identity in educational economic decision making, can better predict exerted student effort-levels.

3.1 Student Utility Model

Akerlof and Kranton (2002) develop and specify a student utility model where identity/self-image is salient. Their model essentially incorporates the concept of identity economics, which they had developed earlier, into the standard model of education (Akerlof and Kranton, 2000). In short, their utility model includes a standard model of education, where a student's utility depends on effort and pecuniary returns to her effort, and an identity payoff, where a student's utility depends on how well her characteristics and behaviour match with that of her social identity (from here on, the word social identity and social category are used interchangeably). In this section, I specify an adjusted version of their student utility model considering the purpose and context of this paper. I start by introducing a standard model of education which is followed by a specification of an identity payoff model. These two models combined constitute the final student utility model specified for this experiment, $U_i = (U_{1,i}, U_{2,i})$. I argue that this model is better at explaining as to why students underinvest in education than the typical standard models of education.

3.1.1 Standard Model of Education

The standard model of education is a utility model built on a simple version of Becker's Woytinsky lecture model (Becker, 1967), as specified in the paper by Jalava et al. (2015). Students are in this case endowed with ability (math skills and logical reasoning) and are given the choice of how much effort to exert in a low-stake test. As such, test scores are a function of their ability (α_i), effort (e_i) and a random term that captures factors such as luck (ϵ_i):

$$TS_i = \gamma_0 + \gamma_1 \alpha_i + \gamma_2 e_i + \epsilon_i \tag{1}$$

In short, $\gamma_1 \; and \; \gamma_2 > 0$ as test scores are increasing in both ability and effort. Ability is assumed to be fixed as teachers are not informed about the nature of the test. Neither are students informed about the test until the day of the experiment. Next, I introduce a reward that students can get if they achieve a test score above a predetermined threshold, $TS_i \geq \overline{TS}$. This model is only applicable to the control group and treatment group 1 and provides an easy understandable example of how the extrinsic incentives, in the form of a reward, work in a standard utility model.[1] Thus, students choose effort to maximize utility per the following specification:

$$U_{1,i} = \max\left\{\left(1 - F_\epsilon(\overline{TS} - \gamma_0 - \gamma_1 \alpha_i - \gamma_2 e_i)\right)R - 1/2(e_i)^2\right\} \tag{2}$$

[1] This is slightly different in treatment group 2 where the three students with the highest test scores will receive the rewards, representing norm-based grading. Here, the threshold will be endogenously determined as it depends on the ability and effort of peers.

subject to $e_i \geq 0$, where F_ϵ denotes the cumulative distribution function (CDF) and f_ϵ the probability density function (pdf) of ϵ. Two factors are essentially in play here: the benefit of achieving a high-test score in the form of a reward (R) and the cost of exerting an effort, $1/2(e_i)^2$.[2]

For treatment group 1, the same standard utility model as above applies except for one exception, the predetermined threshold for the test score is affected by the individual's previous grade (PG), $TS_i \geq \overline{TS}^1 + PG_i$, and students choose effort to maximize utility per this slightly changed specification:

$$U_{1,i}^* \max\left\{\left(1 - F_\epsilon\left((\overline{TS}^1 + PG_i) - \gamma_0 - \gamma_1\alpha_i - \gamma_2 e_i\right)\right)R - 1/2(e_i)^2\right\} \quad (3)$$

In both specifications, the benefit for the students is driven from the probability of receiving the reward when achieving a test score above the threshold, $\overline{TS}$, and the cost comes directly from exerting effort. If both groups face the same threshold, $\overline{TS} = \overline{TS}^1$, the treatment effect for treatment group 1 will only assess how different thresholds, as affected by the individuals' previous grade, affect their effort and performance. What is key in this experimental setup however, is that the threshold for treatment group 1 will be lower than the control group, $\overline{TS} > \overline{TS}^1$, which will make the effect that the previous grade has vary and not necessarily result in a higher threshold than the one for the control group.

In summary, the specified threshold for treatment group 1 will be designed such that even though it is affected by individuals' previous grades, it might be lower, equal or higher for students in different grade-categories than their respective counterparts in the control group. As such, it serves the purpose of labelling/assigning students to different grade-categories, capturing how students react and perform when they are told that their previous grade serve as a starting point for how their future performance is assessed. Students with A-C as previous grades have a lower threshold, students with grade D have the same threshold as the control group, and students with E-F have a higher threshold. What this achieves is to test how powerful the concepts of self-confidence and self-image/social identity are. Are they so great that they affect effort-levels exerted by students in the treatment groups even though their thresholds are the same as the students in the control group? The overall effect of the treatment will depend on the grade distribution among students, but I will cover the specifics of this setup in greater detail in Section 4.2 and 4.3.

3.1.2 Identity Payoff Model

In order to illustrate more exactly how identity and self-confidence can affect optimal effort-levels, I introduce a second model, the identity payoff model, that is based on the work of Akerlof and Kranton (2002). In their paper, they develop a model of how students act as decision-makers whose primary motivation are their identity.

[2] The form of the cost function here is replaced to the one specified by Akerlof and Kranton (2002) but essentially has the same properties as the one specified by Jalava et al. (2015) as it is twice continuously differentiable, increasing and convex.

I begin with specifying a set of social categories, C (these categories can be anything from gender to race), which in this model will be different grade-levels hereby referred to as grade-categories. Second, we have prescriptions, P, that give the ideal characteristics and behaviour for each grade-category. A student, i, is assigned to a category and we denote this as c_i. The students' self-image/identity payoffs (here on, self-image and identity payoffs are used interchangeably), I_i, depends on the match between her characteristics and behaviour with the ideals for her category, $I_i = I_i(e_i, c_i; \varepsilon_i, P)$, where ε_i is $i's$ characteristics. Akerlof and Kranton (2002) argue that this model describes behaviour and how assigned social categories influence behaviour.

In this experiment, social categories, (C), consist of grades in mathematics from the previous semester, $C = [A, B, C, D, E, F]$. As an example, a student can belong to the category of having received the grade A in mathematics last semester or C. For each of these categories, prescriptions give the ideal characteristics. In this context, it is the degree of ability in math (which is fixed and exogenous), a, where a higher grade-level corresponds with higher ability, such that $a(A) > a(B) > a(C) > a(D) > a(E) > a(F)$. This assumption is based on the concept of that grades serve the purpose of communicating the ability of the individual. In addition to this, prescriptions also dictate ideal effort levels on the math test with $e(A) > e(B) > e(C) > e(D) > e(E) > e(F)$. This tells us that students with higher grades should generally exert higher levels of effort than students with lower grade. As such, a student's identity payoff depends on the extent to which her own characteristics and behaviour match with her category's ideals. For example, a student that is assigned to $c_i = A$ earns an identity payoff of $I_A - t(a(A) - a_i)$ where t is a positive parameter scaling the identity loss from I's distance from her ideal. Instead, a student assigned to $c_i = C$ earns an identity payoff of $I_C - t(a(C) - c_i)$. In this paper, I assume based on the paper by Harlen and Deakin Crick (2002), that $I_A > I_B > I_C > I_D > I_E > I_F$, meaning that having a rewarding self-image is positively correlated with higher grade-levels. Finally, for the model to be complete, a student will lose utility $1/2\ (e_i - e(c_i))^2$ for deviating from the prescribed effort-level of her category and this is essentially the outcome variable in this experiment. Hence, students choose effort to maximize their identity payoff per the following specification:

$$U_{2,i} = I_i = \overbrace{I_i - t(a(c_i) - c_i)}^{i's\ characteristics} - \overbrace{\frac{1}{2}\left(e_i - e(c_i)\right)^2}^{i's\ behavior} \tag{4}$$

Combining the standard model of education (2) with the identity payoff model (4), we get the final student utility model of this paper specified as:

$$\begin{aligned} U_i = \left(U_{1,i}, U_{2,i}\right) = p\left[\max\left\{\left(1 - F_\epsilon(\overline{TS} - \gamma_0 - \gamma_1\alpha_i - \gamma_2 e_i)\right)R - 1/2(e_i)^2\right\}\right] \\ +(1-p)\left[I_i - t(a(c_i) - c_i) - \frac{1}{2}\left(e_i - e(c_i)\right)^2\right] \end{aligned} \tag{5}$$

Where $0 < p < 1$ denotes the weights, students put on each utility aspect of their utility function. For example, $p = 0$ describes behaviour when students only care about their identity payoff. According to Akerlof and Kranton (2002), ethnographies suggest very low values of p indicating that students put a much higher emphasis on their identity payoffs than the utility extracted from academic life. They argue that the design of the school system or any sort of school policies affecting social parameters, will almost always influence educational outcomes, as long as p is not equal to one. As explained in the next section, the student utility model above allows for an assessment of how the current school system, affect social categories and prescriptions amongst students. As such, I test whether incorporating the concept of identity economics into the utility function of students, can better predict student behaviour than the standard models of education.

4. Methodology

In this section, I will cover the selection of the sample, the experimental design, the implementation process and the empirical specification.

4.1 Sample

In 2012, the Swedish government decided to introduce grades earlier in the school system and thus grading started in the 6th grade. Previous research has been criticized to conduct experiments on older students which they argue poses a selection bias problem of only selecting a student sample of high-achievers. These high-achievers have not only thrived with the grading systems that are currently in place in so many countries, they have also voluntarily chosen to continue to pursue a higher education. To avoid the risk of selection bias, I limited the prospective school years to those that of primary school. This because of the Swedish law that makes it mandatory for individuals to stay in school until the 9th grade. Ultimately, the choice fell on 6th graders as they pose as an interesting group of students to evaluate the effect of grades on. The 6th grade is the first year when students start to get grades and hence the effect of grades could substantially be more important in how it shapes their behaviour. Moreover, Levitt et al. (2016) discovered that elementary school students turned out to be more responsive to incentives in general, particularly to non-monetary rewards, than older students. Due to the design of the experiment as will be explained further down, it is of preference to have a sample group that are more prone to be incentivized by a non-monetary reward. In addition to this, Jalava et al. (2015), conducted their field experiment on 6th graders in Sweden four years ago and conducting a field experiment on a very similar sample lends space to interesting comparisons of results.

The experiment was carried out in the Stockholm municipality amongst 22 classes and 10 schools, except for 3 classes in 2 separate schools that was carried out in Österåker municipality. This was due to initial easy access to schools in that region which allowed me to test the feasibility of the experiment. Later, the focus was limited to schools in the Stockholm municipality. A list was created that assigned numerical values to each school. With the help of a randomization function, schools were selected and

contacted. Teachers were contacted by phone and asked to participate in the experiment. This was done per a script as to ensure consistency in communication. Out of 13 schools in the municipal of Stockholm that were contacted, 8 accepted. The sessions were usually scheduled and carried out in one to two weeks after the initial contact. Teacher who declined to participate referred to heavy workload as the reason for not participating. I do not, however, suspect that this resulted in a biased sample of schools as the school-level characteristics between those who accepted and those who declined were similar. The sample of schools is diverse with respect to socioeconomic and geographic factors.

4.2 Experimental Design

As previously mentioned, this is a field experiment that is carried out in the Stockholm region. Schools are randomly selected and teachers are appropriately contacted to be asked if they have an interest in participating in this study. Teachers are then given a very brief information package outlining what is required from them in terms of time and cooperation during the field day and are asked not to tell their students about the experiment in advance. They are given restrictive information into the nature and specifics of the experiment and are not told anything about the test questions or the treatment effects. In short, the experiment randomly assigns students to three group (control, treatment 1, treatment 2), where they subsequently are given a math test to solve during 10 minutes. The treatments as such are that the instruction on the first page of the test concerning how their performance is assessed differ among the three groups.

The math test itself is designed based upon the math test used by Jalava et al. (2015) as well as some additional questions and adjustments. I chose to include all four questions that they used in their field experiments. This was since their questions were designed to be suitable for the level of skills students in the 6^{th} grade have. I did not see any reason for not including their questions as they had already been proven to work in a sufficiently good way. Upon further discussion with the authors, I decided to add two more questions as they would have preferred to get an even larger variation in their results. This was done with the help of teachers that were not participants of the experiment. Furthermore, the decision to choose a math test in the beginning rests upon the fact that the evaluation of such questions can be done quite objectively as questions have rather specific correct answers. It is also a side-bonus that math is rather considered a very important subject in school. Sweden has for example struggled to keep a high math proficiency among its student population in comparison to other European countries. The questions in the math test can be found in Table 18 (English) and Table 19 (Swedish) in the appendix.

Trial Run

Before discussing the implementation process that was conducted, it is important to mention that before the experiment was initiated, a trial run was conducted. This trial run was initially planned to mark the start of the experiment but since I encountered difficulties, I decided to use the opportunity to analyse how the implementation process could be improved. For the trial run, 40 students from 2 classes in the

same school participated. The difficulties that made me decide to remove these observations from the rest of this study was that students were very closely seated next to each other. This enabled them to figure out that they had been given different treatments which caused a reaction in both classes. Moreover, adequate measures were not in place as to stop students from looking at each other's solutions which made the results questionable. Furthermore, the presentation of the test (leaving out the fact that the result of the test did not affect their real grades) made students nervous which was not the intention of the experiment. This experience did however provide insight in how to improve the implementation process during future sessions.

Implementation

Sufficient amount of math tests and surveys were brought to each class (divided equally amongst the control group and the treatment groups). Students were randomly assigned to control and treatment groups that differed in information concerning how their performance were assessed. In turn, students had no choice regarding participation as all of them had to sit through the entire duration of the session. Before tests were handed out, there was an initial introduction. For this, I used a script as to assure that all classes received the same type of information in a consistent manner. It should be noted that all the sessions were conducted with the presence of the teacher of the class. After a short introduction of myself, students were told that they would have to complete a 10 minutes' math test followed up by a short survey regarding school. They were immediately told that the test result would not affect their grades but that they instead could win a diploma and a prize if they performed adequately. The nature of the prize was not revealed until the end of the class, but was a pencil. Next, they were informed that their test result and their answers to the survey questions would be entirely anonymous and coded, which was strongly emphasized. Short administrative guidance was communicated such as the time constraint of the test, calculators not being allowed, the importance of solving questions individually and that they would have to sit quietly during the 10 minutes even if they were done with the test early. This was followed up by clear instructions that they could write their name and read the instructions on the first page (the treatment) as soon as I handed out the exam but that they were not allowed to turn the page and start with the test until they were instructed to do so. This was repeated several times as to ensure that everybody had understood these rules. Students were also told that any questions they had would have to wait until they were done with the test and the survey. After the brief introduction, students were told to sit according to such a way that ensured that they could not see each other's tests. This step was performed as to ensure that students with different instructions would not sit next to each other and find out that they had received different treatments. This was done either by distancing students from each other or by putting up covers between them. An important next step was the fashion the tests were handed out. Tests to the treatment groups were handed out first followed by the control group. This is due to the fact that the instructions for the treatment groups are longer and require a longer time to read. I wanted to make sure that the students had sufficient time to read and grasp the instructions before allowing them to start with the test. After all tests were handed out, I took a moment and ensured that

everybody had read their instructions and filled in their names, upon which students were told they could start with the test. After 10 minutes, tests were collected in and surveys were handed to the students. There was no specific time constraint set for them to answer the questions but 2 minutes were put aside for this component. After approximately 2 minutes, surveys were collected in and I moved to another room to correct the exams (per a pre-defined answer sheet). Subsequently I would return to the classroom to hand out the diplomas and prizes to the winners, as well as to briefly explain the purpose of this study and allow interested students to look at their test scores.

4.3 Treatment Variables

The treatment variables chosen for this experiment serve to assess the effect, the introduction of information that previous grades affect student's outcomes, can have on motivation and performance. Motivation is an essential factor affecting performance, where Wise and DeMars (2005) for example show that higher motivation is associated with higher test scores.

To achieve this, the experiment will need to have a reward that with a high probability incentivises students to exert effort. If students are not incentivised to exert effort, the varying effect of the treatment (the effect of previous grades on their utility) will be harder to capture or be completely insignificant. Since this paper investigates the effect of grades, it would have been optimal to have grades as a non-monetary incentive itself. However, previous papers show indecisive result regarding the effect of absolute grading on motivation (Harlen and Deakin Crick, 2002; Jalava et al., 2015; Klapp et al., 2016). Jalava et al. (2015), for example, show that absolute grading did not have a significant effect of increased motivation amongst students compared to no incentive at all. Due to this, I chose instead to combine two non-monetary rewards that have previously been shown to have significant effects on student motivation, a diploma (certificate) and a prize (Jalava et al., 2015; Kosfeld and Neckermann, 2011). The reason for why I choose two non-monetary rewards instead of one is to ensure that as many students as possible find the rewards incentivizing.

As indicated in Table 1, the incentive in the form of the rewards in this experiment are the same for the control group and for both treatment groups. The treatment of the experiment however is found on the first page of the math test where instructions regarding how students are being assessed differ (questions in the math test are however the same for all groups). Treatment group 1 differs compared to the control group in one area, namely that the previous grade of a student affect her outcome. Treatment group 2 differ in an additional area as well, namely that the grading system is norm-based. Instead of receiving the diploma and the prize when subjects reach a pre-defined threshold, only the top 3 performing students receive the rewards. This treatment is included as to assess the long-term effect of grades if the grading system were to be norm-based. The additional effect of treatment group 2 is analysed by comparing it to treatment group 1, which serves the purpose of being a control group in this case.

***Table 1**: Control group and treatment groups with corresponding treatments*

Group	Previous Grade	Absolute grading	Relative grading	Diploma/Prize
Control		x		x
Treatment 1	x	x		x
Treatment 2	x		x	x

4.3.1 Treatment Effects

Table 2 demonstrates exactly what the different instructions are for the three groups in the experiment. The original Swedish version of the treatments can be found in Table 17 in the appendix. The control group will receive information as to let them know that the only way they can receive the rewards is to score 20 out of 30 points on the test. Treatment group 1, however, will also be instructed that their previous grade will affect their final outcome. Essentially, this design builds as previously mentioned on how the current school system functions and operates in Sweden, as well as in other countries. For example, in primary school, students receive a final grade in math for each semester. This grade is essentially based on a weighted average of approximately 2-3 previous exams, usually with higher weights on the last exam as it encompasses all the learning objective of the course. As an example, an individual who receives a D on the first exam of the semester compared to an individual who receives a B, needs to perform much better on the second/final exam to receive the same final grade. As such, the threshold to achieve a higher utility (assuming grades are rewarding in an increasing order) has increased which relates back to the notion of unfairness stated earlier. Loss of motivation and subsequently a lower performance in treatment group 1 could as such depend on whether students find the assigned impact of previous grade fair or not. Students in the treatment group might find the fact that their final outcome is dependent on their previous grade, which is greatest at the lowest grade-level, unfair and change their behaviour. Although this may sound as something elementary for someone who is used to this assessment system, one must question the long-term effect of it on the motivation and performance of students.

Table 2: Information regarding test assessment/Treatment effects

Control group
On this test you can obtain a total of 30 points.
If you obtain 20 points or more you will receive a diploma and a prize.

Treatment group 1
On this test you can obtain a total of 30 points.
You will however receive deductions (negative points) based on your previous grade in mathematics from the previous semester.

- If you received A in mathematics last semester, you will receive 0 negative points.
- If you received B in mathematics last semester, you will receive 1 negative points.
- If you received C in mathematics last semester, you will receive 2 negative points.
- If you received D in mathematics last semester, you will receive 3 negative points.
- If you received E in mathematics last semester, you will receive 4 negative points.
- If you received F in mathematics last semester, you will receive 5 negative points.

If you in **total** obtain 17 points or more you will receive a diploma and a prize.

Exempel: If Kalle scores 19 out of 30 points on the test and received a C (2 negative points) in mathematics last semester, his **total** score will be 17.

19 – 2 = 17

Treatment group 2
On this test you can obtain a total of 30 points.
You will however receive deductions (negative points) based on your previous grade in mathematics from the previous semester.

- If you received A in mathematics last semester, you will receive 0 negative points.
- If you received B in mathematics last semester, you will receive 1 negative points.
- If you received C in mathematics last semester, you will receive 2 negative points.
- If you received D in mathematics last semester, you will receive 3 negative points.
- If you received E in mathematics last semester, you will receive 4 negative points.
- If you received F in mathematics last semester, you will receive 5 negative points.

If you are among the three with the highest score in **total** in the class, you will receive a diploma and a prize.

Exempel: If Kalle scores 19 out of 30 points on the test and received a C (2 negative points) in mathematics last semester, his **total** score will be 17.

19 – 2 = 17

The original Swedish version can be found in Table 17.

In treatment group 1, previous grades change the threshold level that students need to achieve on the math test in order to get the rewards. Table 3 illustrates what the threshold is for each student in the different grade-categories, taking into account the negative impact of previous grades as it simulates the current school system. The impact essentially lowers the threshold for students in grade-categories A-C, increases it for students in E-F and makes no impact for students with a previous grade of D compared

to their counterparts in the control group. In an optimal experimental setting with access to a computerized process, it could have been possible to tweak the threshold level on an individual basis as to have the same threshold level for each grade-category group. This would have allowed me to exclude the "threshold" effect, and only measure the potential effect of previous grade on self-confidence and whether it constitutes a social identity. This was, however, not possible considering the resources available. Instead, the decision of how large the impact of previous grade would be for each grade-category rested on a combination of two factors, namely the grade distribution in mathematics for 6th graders in Sweden and previous research that indicate that low-performing students suffer the most in terms of self-confidence and motivation as a result of grades. I wanted the impact to change the test score threshold to a degree where approximately 50% of the students would have a lower threshold, while the other 50% would have the same or a higher threshold, meaning that the threshold effect would on average be the same as the control group. According to Nydahl and Ridderlind (2016), students in grade-categories A-C constituted approximately 49% of the total population, while students with D-F constituted the rest. At the same time, I wanted some of the lower-grade categories to have a lower or equally high threshold as their counterparts in the control group to exclude the threshold effect. This is the case for grade-categories C and D, where a potential difference between the control group and treatment group 1 will yield an interesting comparison.

***Table 3**: Treatment Group 1 – Impact of previous grades on reward thresholds*

Grade-category	Impact of previous grade	Test score threshold
A	0	17
B	-1	18
C	-2	19
D	-3	20
E	-4	21
F	-5	22

Furthermore, the treatments, in addition to simulating the current school system, can be related back to the student utility model (5) specified earlier in this paper. In the model, social categories and as such social identities equate to previous grades that were obtained in math. Although all the students in the sample have previously been assigned grades in mathematics, the treatment is supposed to enhance this "labelling"/social category identification through a framing effect. In an ideal world, decisions are based on rational thinking. Tversky and Kahneman (1985) however, show that the framing of decisions can affect individuals' actions and choices. This concept is incorporated in the treatments and I argue that it serves the function of assigning students to different grade-categories, or at a minimum reinforces/reminds them of which grade-category they belong to and as such their social identity. As such, it simulates how school policies such as the policy of assigning grades to students, as argued by

Akerlof and Kranton (2002), can change the divisions into social categories and prescriptions, thereby affecting student behavior. Students in the treatment groups will be assigned to their previous grade in mathematics which their performance is dependent on, while students in the control group will not and are unknowing of any such impact. If this assignment or reinforcement is successful, it should lead to a lower *p* value amongst students in the treatment group compared to the control group, meaning that students that are assigned to or reminded of their social identity put a higher weight on their identity payoff than students that are not. If this happens, it means that the choice of how much effort to exert depends to a larger extent on the maximization of the identity payoff than the maximization of the standard model of education, where a student's utility depends on effort and pecuniary returns to her effort. In turn, the identity payoff model specified in this paper predicts that students assigned to lower grade-categories should exert lower effort than students in higher grade-categories.

I test this theory by essentially comparing test scores between the control group and the treatment groups. For example, if students in the treatment group with the same threshold to obtain the rewards and the same grade as their counterpart in the control group have lower test scores, it potentially indicates that those students (who were assigned to or reminded of their social identity) chose their effort-levels more in preference to maximize their identity payoffs.

Students in treatment group 2 will in the same way as students in treatment group 1 be instructed that their previous grade will affect their final outcome, but they will also be instructed that only the top three performing students will receive the rewards. This additional effect is added to see whether previous grades affect the motivation and performance of students differently when they are assessed through norm-based grading. This is of interest as previous research have shown general positive effects of rankings, but where low-performing students are negatively affected (Azmat and Iriberri, 2010; Jalava et al., 2015).

4.4 Survey

After completion of the math test, students were given approximately two minutes to complete a survey. Except for their name and gender, students were asked a total of five questions. Table 4 reports the English translated version of the survey questions and the Swedish version which the students received can be found in Table 20 in the appendix. The purpose of the questions varies from solely functioning as control questions to describing how the students feel about grades and what they thought of the math test and its components. The answers to the questions are either Yes/No or are based on a likert-scale of 1 to 5. A short description of each question's purpose and functionality follows below.

Question 1 is supposed to measure how important it is for students to perform well on the test. As such, it works as a proxy for how motivating the incentives of the test are for the students. Disregarding measurement error, if students find the test design very incentivizing, it should be more important for them to perform well on the test and hence exert a higher level of effort. Question 1 can then work as a

substitute and proxy for the main outcome variable of the paper, namely test scores. The main purpose of this questions however, is linked to the identity payoff model stated earlier where prescriptions for assigned social categories dictate ideal effort levels on the math test with $e(A) > e(B) > e(C) > e(D) > e(E) > e(F)$. The treatment effects are stated in a way that assign students to their grade-category. The identity payoff model then specifies that students in turn optimize their effort-levels to the prescribed ideals of their social category. Whether this happens is something that is of interest to investigate.

Question 2a functions strictly as a control questions that checks whether students have read and understood their treatment instructions. This is important to check for as it ensures internal validity for the potential findings. Following this, question 2b measures whether students that received the control group and treatment group 1 instructions (with thresholds), found the threshold to reach in order to obtain a reward to be too high. This is linked to whether students found the test design to be unfair. Any differences between the control group and treatment group 1 is of interest.

Question 3 measures how likely students think it is that they will receive the rewards. Since this question is answered after the math test have been completed, it measures how well the students think that they performed. This in turn depends on how much effort they exerted and their belief about their own ability in math. Controlling for exerted effort-levels however, which I argued above that Question 1 could be a proxy for, Question 3 can tell us about a student's self-confidence and belief regarding her ability. If students believe that they have a high ability in math, they also believe that they have a greater probability to reach the threshold required to obtain the rewards. This links back again to the identity payoff model, this time however to the ideal characteristics that the prescriptions assign to each social category. In my model, the ideal characteristics is the degree of ability where a higher grade-level corresponds with higher ability such that $a(A) > a(B) > a(C) > a(D) > a(E) > a(F)$. Since the math test is the same for all three groups, it is interesting to compare whether the treatment that assigns students to their grade-category also affect the self-confidence and belief of students in regards to their own ability.

Finally, Questions 4, 5a and 5b serve as variables to describe the consensus amongst students regarding absolute grading and norm-based grading. Question 4 measures whether students find that grades motivate them, while Questions 5a and 5b measures whether students compare their grades and results with each other and whether it motivates them, simulating norm-based grading.

***Table 4**: Survey questions - English*

Q1. How important was it for you to do well on the test that we just did?
Not at all 1 2 3 4 5 Very important

Q2a. Was there a certain threshold that you needed to attain in order to get a diploma/prize?
☐ Yes - (over 20 or 17 points) ☐ No - (top three with the highest points receive a diploma/prize)
If you answered No, skip question 2b.

Q2b. Do you think that the threshold you needed to attain in order to get a diploma/prize was too high?
Not at all 1 2 3 4 5 Very high

Q3. How likely do you think that it is that you will receive a diploma/prize?
Not at all 1 2 3 4 5 Very large

Q4. Do grades motivate you?
Not at all 1 2 3 4 5 Very much

Q5a. Do you usually compare your grades and results with your classmates?
☐ Yes ☐ No
If you answered No, skip question 5b.

Q5b. How motivated do you become by comparing your grades and results with your classmates?
Not at all 1 2 3 4 5 Very much

The original Swedish version can be found in Table 20.

4.5 Econometrical Specification

To estimate the effect the treatments (impact of previous grades, impact of previous grades with norm-based grading) have on student performance, measured as obtained test scores, and to test the hypothesis stated in the following section, I specify an econometrical model.

The regression below is a simple ordinary least squares (OLS) model that estimates the average treatment effect, T where j specifies the treatment group, on the outcome variable, $Test\ score$, for each subject (student), i.

$$Test\ score_{ji} = \alpha_0 + \delta_j T_{ji} + \varepsilon_{ji},\ \ j = 1, 2 \tag{6}$$

The average test score for the control group is α_0, while the average causal effect of treatment j is δ_j and ε_{ji} is the error term. The estimation of the regression above gives us the average causal effect of the two treatments, where randomization assures us that the treatment effects are uncorrelated with the error term. The effect of the treatments is however highly dependent on the distribution of grades in each

group, as the treatment effects, as hypothesised below, will have different effects in scale on students with different grade-category.

In addition to this, I expect grades to have a strong positive correlation with ability, as grades per definition are meant to communicate the abilities of students. Since a high ability per definition should lead to a higher test score, I expect the impact of different grade-categories on test scores to be significantly different. To reduce noise and control for what I essentially expect to be large differences in intercepts for students in different grade-category, I control for grades in a second regression. In addition to controlling for ability, including grades indirectly also controls for the different threshold levels that students in the control group and treatment group 1 have, as their previous grades affects the threshold they need to reach in order to obtain the rewards. In the same regression, I control for gender as it is the only additional control variable on an individual level. I do not expect, however, that there will be any differences in general between boys and girls on average. This second regression takes the following form:

$$Test\ score_{ji} = \alpha_0 + \delta_j T_{ji} + \gamma_j Grade + \emptyset_j Female \quad j = 1, 2 \tag{7}$$

Furthermore, to improve the precision of the regression estimates per a similar estimation by Jalava et al. (2015), I also control for class and school-level variables in a third regression. It should be added that class size is particularly important as it has an effect when considering that it determines the number of students competing for the top three positions in treatment group 2. Hence, I control for class size (class-level control) and a vector X_s with three school-level controls; the percentage of students with foreign background, percentage of students with parents with higher education and average GPA in math amongst 6th graders at a school level. This final regression takes the following form:

$$Test\ score_{ji} = \alpha_0 + \delta_j T_{ji} + \gamma_j Grade + \emptyset_j Female + Z_j Class_{size} + \beta_j X_s \varepsilon_{ji} \quad j = 1, 2 \tag{8}$$

4.5.1 Hypotheses

Subsequently, I can proceed to hypotheses testing where I test the null-hypothesis of no average treatment effect against the alternative hypothesis that there is a non-zero average treatment effect in both treatment groups separately. I expect that the treatment effect will be negative in both treatment groups:

$$H_{0,CT1}: \overline{Testscore_C} = \overline{Testscore_{T1}}$$
$$H_{1,CT1}: \overline{Testscore_C} \neq \overline{Testscore_{T1}}$$

$$H_{0,CT2}: \overline{Testscore_C} = \overline{Testscore_{T2}}$$
$$H_{1,CT2}: \overline{Testscore_C} \neq \overline{Testscore_{T2}}$$

However, since norm-based ranking have been shown to increase performance amongst students, I expect the negative treatment effect to be lower in treatment group 2 than in treatment group 1 (Azmat and Iriberri, 2010; Jalava et al., 2015). As such, I also test for whether the two treatment effects are statistically significantly different:

$$H_{0,T1T2}: \overline{Testscore_{T1}} = \overline{Testscore_{T2}}$$
$$H_{1,T1T2}: \overline{Testscore_{T1}} \neq \overline{Testscore_{T2}}$$

Subsequently, I expect the negative treatment effects to be the largest for students in the lower grade-categories, as previous research have shown that low-performing students as a result of grades suffer the most in terms of self-confidence (Harlen and Deakin Crick, 2002). In addition to this, previous research analysing gender differences regarding math test scores show that girls tend to avoid competitive settings or perform poorly. As such, it is interesting to analyse whether the effect of the treatments will be different considering gender (Niederle and Vesterlund, 2010). If for example, the tendency to avoid competitive settings relates to self-confidence or social identity, girls might be negatively affected to a larger degree than boys.

Finally, to test the hypotheses above and attain robust statistical analyses, I estimate that the total required sample size for the experiment is 525 subjects (175 students in each group). This is based on a two-sided test, an anticipated effect size (Cohen's d) of 0.30, and a power of 0.8 with a conventional alpha of 0.05. The anticipated effect size in turn is based on a previous quasi-experimental paper by Klapp et al. (2016) that show that grades negatively affect the performance of low-to-medium performing students. The calculated effect-size of their result was a Cohen's d of 0.30. In addition to this, a Cohen's d of 0.20 is considered to be relevant in the field of education when studies are made on different performance outcomes such as test scores and grades, and can have implications for public policy reforms (Durlak, 2009). I expect the effect-size of the treatments in this paper to be equally strong, if not more.

4.6 Considerations

Due to resource constraints, all experimental studies have trade-offs. Hence, I will shortly touch upon some considerations that should be noted, in light of the experimental setup of this paper. First, this paper investigates the effect of grades, but it does so by setting previous grade as a starting point, upon which students' performances are assessed on. It does not assess, the effect of grades as a non-monetary reward itself. The reason for this can be found in section 4.2.1. Briefly, it is because in order to analyse the effects of the treatments in this experiment, it is important that students are with a high probability incentivized by a reward. As such, the non-monetary rewards in this paper are a diploma and a prize (a pencil). If the treatments, have a negative effect on motivation and performance, their effect might not be the same when grades are the non-monetary reward in place. However, I argue that this is very unlikely and most possibly of the other direction. If previous grade as a starting point have negative

effect on motivation and performance, this effect should be even bigger when the non-monetary reward promised is grades themselves. This substitution has an additional trade-off, and that is that grades as non-monetary incentives have several levels of rewards with a higher grade being equal to a higher reward. This is not the case when the reward is only a diploma/prize that can only be attained when scoring above a pre-determined threshold. This trade-off reduces the similarity of the experimental setup with how the school system operates. I argue however that it is close enough for the purpose of the study, especially since considering that the grade-levels in the middle (C-E) are arguably not even closely considered as prestigious and rewarding as the highest grades (A-B). Due to this, their differentiating effects are marginally small.

Secondly, the treatment effects in this experiment are framed negatively as students are told that there will be deductions (negative points) based on their previous grades. Optimally, I would have liked to use a neutral framing where students are told that their outcome is a weighted average of their previous grade and their test score. However, due to the age and skill-level of the students, I had to adjust the framing so that it would be easier for them understand the underlying mechanism in play. As such, it can be argued that the experimental setup in this paper does not fully simulate how students are assessed in schools, but is the closest possible.

Furthermore, treatment group 2 has two additional treatments compared to the control group, the impact of previous grades and norm-based grading instead of criterion-based grading. It would have been optimal if treatment group 2 had its own control group, with only norm-based grading, upon which the effect of previous grades could be assessed. Instead, I rely on comparing treatment group 2 with treatment group 1 as a control group, referencing the effects shown in previous research. This compromise is only due to the limited time I had left for conducting the experiment, where a larger sample size would have been necessary.

5. Data

The experiment was carried out on a total of 372 (plus students in the trial that were excluded as explained in the previous section) students in the 6th grade in the Stockholm region from 20 classes and 9 schools. This turned out to be lower than the 525 subjects (175 in each group) calculated to be required in the total sample initially. The reason for the slightly lower sample data collected was due to time constraints (spring holidays in schools) and lack of resources. Considering, however, that it is higher than most studies in the field of behavioural economics, with at least 123 students in each group, its implications are still of high interest. Out of these 372 students, 207 were boys and 165 were girls. Table 5 reports average test scores across control and treatment groups for the full sample and with respect to gender. The average test score for the full sample was 21.08 for boys, 19.60 for girls and 20.42 in total. The difference between boys and girls is statistically significant, albeit not when controlling for their

grades. The test scores range from a minimum of 2 points to the maximum of 30 points, where the test score distribution is negatively skewed, which means that students also perceived this test as relatively easy (which is interesting as it contains more questions with the same time constraint in comparison to Jalava et al. (2015).

Table 5: Average test scores across control and treatment groups

	Control	T1	T2
Full sample			
N individuals	125	124	123
Test score (standardized)	0.112	-0.099	-0.014
	(0.084)	(0.093)	(0.091)
Test score (points)	21.156	19.770	20.330
	(0.554)	(0.614)	(0.600)
Boys			
N individuals	63	65	79
Test score (standardized)	0.127	-0.051	0.203
	(0.125)	(0.123)	(0.104)
Test score (points)	21.254	20.085	21.753
	(0.823)	(0.811)	(0.681)
Girls			
N individuals	62	59	44
Test score (standardized)	0.097	-0.152	-0.402
	(0.114)	(0.143)	(0.160)
Test score (points)	21.056	19.424	17.775
	(0.747)	(0.937)	(1.053)
Full sample, excluding NoTreat			
N individuals	116	109	82
Test score (standardized)	0.175	-0.051	0.186
	(0.082)	(0.096)	(0.098)
Test score (points)	21.569	20.083	21.641
	(0.539)	(0.629)	(0.644)
Boys			
N individuals	58	57	60
Test score (standardized)	0.194	-0.009	0.291
	(0.117)	(0.125)	(0.112)
Test score (points)	21.698	20.360	22.333
	(0.772)	(0.822)	(0.734)
Girls			
N individuals	58	52	22
Test score (standardized)	0.155	-0.098	-0.101
	(0.116)	(0.148)	(0.193)
Test score (points)	21.440	19.779	19.755
	(0.760)	(0.970)	(1.266)

Note: The table displays descriptive statistics of test scores and the number of students in each of the treatment groups and the control group. Both average points scored on the test and standardized test scores (with mean 0 and standard deviation 1) are reported. The first column represents the control group and columns T1–T2 represent the two treatment groups; (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. All statistics are displayed separately by gender. Standard errors are displayed in parentheses.

In addition to the data obtained through the experiment such as the test scores and survey answers, I have chosen to include a set of other variables as to control for other factors and increase the precision of the analyses. The dataset as such thus contains control variables on the individual, class and school level. The variables on the individual level include test scores, grades, gender and answers to the survey questions. The only class-level variable is an indicator for class size. For the school-level variables, I have chosen to include three measures: the percentage of students with foreign background, the

percentage of students with parents that have post-secondary education, and a measure of average GPA. All the individual and class-level variables are either measured or collected via the field experiment. In regards to the three school-level variables, both the percentage of students with foreign background and the percentage of students with parents that have post-secondary education were obtained from Skolverket's (National Agency for Education – Sweden) statistical database, Skolverkets Internetbaserade Resultat- och kvalitetsInformationsSystem (SIRIS). The measure of average GPA was, however, calculated according to the math grades obtained at the individual level. Each school were appointed an average score based on the math grade of the students in that school. This variable is supposed to be a proxy for school quality and is highly correlated with average GPA in the 9th grade for all courses at each school which is a variable collected from Skolverket's analytical tool Skolverkets Arbetsverktyg för Lokala SambandsAnalyser (SALSA). This variable was in turn not used as it missed data for 2 schools in this sample.

As a result of randomization to control and treatments within each class, I obtained a balanced number of students in each group spanning from 123 to 125 students. The exact number of students in each group is reported in Table 6 accompanied by a breakdown of grades. This equal distribution among all three groups further implies that groups are randomized and balanced with regards to other factors such as grade distribution, gender, class and school-specific factors. Table 7, a balance table, reports the mean for the control variables for the control group, as well as the difference in means between the control group and each treatment group accompanied by a breakdown of gender. All control variables seem to be equal in mean which strengthens the assumption of a clean randomization process. The only statistically significant difference is the proportion of girls between the control group and treatment group 2.

Table 6: Grade distribution in sample groups

Grade	*Control group*	*Treatment group 1*	*Treatment group 2*
A	9	5	12
B	19	19	22
C	37	41	34
D	26	33	23
E	27	21	26
F	7	5	6
Total	125	124	123

Table 7: Balance table

	Control	T1	T2	T1-C	T2-C
Full sample					
Female	0.496	0.476	0.358	-0.020	-0.138**
	(0.502)	(0.501)	(0.481)	(0.064)	(0.062)
Class size	24.064	24.363	24.033	0.299	-0.031
	(2.918)	(2.721)	(2.997)	(0.358)	(0.376)
Foreign background	19.48	18.073	20.163	-1.407	0.683
	(19.711)	(17.996)	(21.498)	(2.393)	(2.618)
Parent education	72.84	73.823	72.610	0.983	-0.230
	(13.754)	(12.683)	(14.659)	(1.677)	(1.805)
Average GPA	13.3	13.468	13.679	0.168	0.379
	(4.395)	(3.835)	(4.404)	(0.523)	(0.559)
N	125	124	123		
Boys					
Class size	24.349	24.585	24.443	0.235	0.094
	(2.737)	(2.555)	(2.510)	(0.468)	(0.441)
Foreign background	17.873	18.015	18.177	0.142	0.304
	(15.782)	(17.517)	(19.366)	(2.950)	(3.018)
Parent education	72.937	73.692	73.873	0.756	0.937
	(11.358)	(11.983)	(13.055)	(2.065)	(2.083)
Average GPA	13.929	13.346	14.652	-0.582	0.723
	(4.550)	(4.222)	(4.417)	(0.776)	(0.756)
N	63	65	79		
Girls					
Class size	23.774	24.119	23.295	0.344	-0.479
	(3.086)	(2.895)	(3.632)	(0.545)	(0.655)
Foreign background	21.113	18.136	23.727	-2.977	2.614
	(23.045)	(18.661)	(24.712)	(3.824)	(4.682)
Parent education	72.742	73.966	70.341	1.224	-2.401
	(15.920)	(13.515)	(17.096)	(2.691)	(3.236)
Average GPA	12.661	13.602	11.932	0.940	-0.729
	(4.170)	(3.388)	(3.845)	(0.693)	(0.796)
N	62	59	44		

Note: The table displays descriptive statistics for the control variables separately over treatment and control groups. The first column presents the control group mean for each variable: gender, class size, and three school-level variables. Standard deviations are displayed in parentheses. The second and third column present the same statistics for treatment group 1 and 2. Columns T1-C and T2-C report the differences in means between treatment groups and the control group. The two treatments are: (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. Standard errors are displayed in parentheses. Asterisks indicate a significant difference of means, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

In addition to an analysis of the full sample, I have also chosen to analyse a subset of the sample that excludes students that failed to give a correct answer to control question Q2a in the survey. The question asks students whether they had to score above a threshold or get a top three ranking in order to receive a diploma and a prize. I argue that students that failed to answer this question correctly may not have fully read or understood the instructions of the test which essentially is the treatment effect of the experiment. Excluding these individuals drops 65 observations in total, 9 from the control group (5 boys and 4 girls), 15 from treatment group 1 (8 boys and 7 girls) and 41 from treatment group 2 (19 boys and 22 girls). From now on, whenever a sample is referred to as excluding NoTreat, I am referring to this

subsample of 307 students. It should be noted however that for the very first session of the experiment, Question 2a was not asked. As such, 14 students out of the 65 dropped never had the chance to answer this question. I drop them regardless, as I cannot tell whether they read and understood the treatments. Table 5 also reports average test scores across control and treatment groups for this subsample, including a split by gender. A balance table for this subsample is also provided in Table 8. As reported, all control variables seem to be equal in mean in this subsample as well which strengthens the assumption of a clean randomization process.[3] In the following section, each regression is accompanied by an additional one on this subsample to assure robustness of the results. These regressions are reported in Table 14, Table 15 and Table 16 in the appendix.

Table 8: Balance table, excluding NoTreat.

	Control	T1-C	T2-C
Full sample			
Female	0.5	-0.023	-0.232***
	(0.047)	(0.067)	(0.069)
Class size	24.37931	0.336	0.218
	(0.216)	(0.299)	(0.319)
Foreign background	19.43103	-2.266	-3.370
	(1.882)	(2.486)	(2.688)
Parent education	73.40517	1.210	2.156
	(1.248)	(1.669)	(1.809)
Average GPA	13.49138	0.385	1.173**
	(0.385)	(0.503)	(0.547)
Boys			
Class size	24.655	0.275	0.045
	(0.269)	(0.378)	(0.381)
Foreign background	17.672	-1.532	-1.639
	(2.129)	(2.801)	(2.983)
Parent education	73.552	0.756	1.732
	(1.424)	(2.065)	(2.019)
Average GPA	14.310	-0.582	0.898
	(0.506)	(0.702)	(0.657)
Girls			
Class size	24.103	0.377	0.215
	(0.337)	(0.464)	(0.602)
Foreign background	21.190	-2.901	-5.053
	(3.106)	(4.169)	(5.469)
Parent education	73.259	1.164	3.060
	(2.064)	(2.793)	(3.682)
Average GPA	12.672	1.366*	0.509
	(0.563)	(0.709)	(0.982)

Note: The table displays descriptive statistics for the control variables separately over treatment and control groups excluding NoTreat. The first column presents the control group mean for each variable: gender, class size, and three school-level variables. Columns T1-C and T2-C represent the differences in means between treatment groups and the control group. The two treatments are: (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. Standard errors are displayed in parentheses. Asterisks indicate a significant difference of means, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

Finally, Table 9 reports the distribution of answers to the survey questions. As demonstrated by Question 1, students were fairly incentivized by the test design as approximately 61% of students felt that it was very important for them to perform well on the test. Over half the students (56%) that answered Question

[3] Additional t-tests were conducted for the control variables comparing the subsample to the dropped observations. The only difference that was statistically significant was the class-size variable which on average was 2 students lower for the dropped observations.

2b felt that the threshold they needed to achieve to be rewarded was too high. Lastly, Question 4 and 5b indicate that while approximately half of the students find grades motivating, only 61% of the students compare their grades with their classmates (norm-based grading) and they do not find it equally as motivating.

Table 9: Survey answer distributions

Lowest value	1	2	3	4	5	NA	*Highest value*
Q1	6,5%	7.5%	24.5%	32%	29.3%		
Q2a	Yes - 68.5%	No - 24.5%	NA - 5.7%				
Q2b	40.3%	15.6%	13.7%	5.1%	2.2%	22.9%	
Q3	11%	10.8%	27.2%	26%	13.7%	11%	
Q4	5.9%	7%	28.8%	32.5%	25%		
Q5a	Yes - 61.3%	No - 38.4%					
Q5b	10.5%	10.2%	23.9%	13.7%	5.7%	35.8%	

6. Result

In this section, the result of three regressions are reported, all which are based on the econometrical model as stated in section 4. For all regressions, the outcome variable, test scores, have been standardized to mean 0 and standard deviation 1. First, the average causal effect of the treatments is estimated based on the full sample. Second, the same analysis is reported by gender. Finally, the treatment effects are estimated by grade-categories. Each regression is accompanied by a robustness check, consisting of the same regressions, with the difference of being conducted on the subsample specified earlier, excluding Notreat. The result of these estimations can be found in the appendix. In addition to this, all regressions are estimated with cluster-robust standard errors on the class level, constituting an additional robustness check as per Jalava et al. (2015).[4]

Table 10 reports the result of the main econometrical model as stated in section 4. The first regression in Column 1 only includes the average causal effect of the treatments without any controls. The effect of both treatments is negative and in line with the hypotheses, but only the effect of treatment group 1 is statistically significant. The regression in Column 2 includes two additional individual controls, grades and gender. All the coefficients for each grade-level are positive, significant and in line with expectations. A higher grade reflecting ability is shown to have a large impact on test scores. Including these two controls does not change the effect of the treatments but improves the estimations and as such, the average causal effect of both treatments becomes statistically significant. The final regression in Column 3 includes class and school-level controls. Including these controls does not change the effect of the treatments nor the individual controls. Except for the very small coefficient for the percentage of students with parents that have higher education, none of these class and school-level controls have a statistically significant effect on test score. For all three regressions, effect of treatment 1 and 2 leads to

[4] All the regressions report the same estimations when not estimated with class level clusters.

0.25 and 0.18 standard deviation lower test score on average. In the last row, the p-value for an F-test, testing for whether the average effect of treatment 1 is equal to treatment 2, is reported. Although the effect of treatment 2 is lower than treatment 1, their difference is not statistically significant. Table 14 in the appendix reports the exact same regression estimations as Table 10, but excludes students that did not answer the control questions correctly, which arguably may have not been affected by the treatment. Hence, it works as a robustness check for the estimations with the full sample. All the estimations are of similar sign, scale and statistical significance, except in the first regression, in Column 1, where the effect of treatment 2, is no longer negative but instead close to zero. This might be due to the fact that amongst the 65 students that did not answer the control question correctly, 41 students were from treatment group 2, resulting in a large drop of observation in the group. The difference between the effect of treatment 1 and 2 is however significant, according to the F-test, when excluding Notreat, indicating that the effect of treatment 2 is less negative than treatment 1. As an additional robustness check, I ran the same regressions on the full sample with class fixed effects and got the same results.

Table 10: Impact of treatments on test scores (standardized)

	(1)	(2)	(3)
T1	-0.211*	-0.249**	-0.247**
	(0.120)	(0.0916)	(0.0927)
T2	-0.126	-0.182**	-0.187**
	(0.114)	(0.0800)	(0.0775)
Grade			
A		2.472***	2.451***
		(0.194)	(0.210)
B		2.315***	2.242***
		(0.182)	(0.219)
C		1.801***	1.751***
		(0.194)	(0.208)
D		1.408***	1.357***
		(0.202)	(0.207)
E		0.661**	0.651**
		(0.240)	(0.250)
Female		0.00865	-0.0273
		(0.0933)	(0.0919)
Class size			-0.0223
			(0.0195)
Foreign background			0.0135
			(0.00953)
Parent education			0.0343***
			(0.00868)
Average GPA			-0.0611
			(0.111)
Constant	0.112	-1.391***	-2.735
	(0.084)	(0.215)	(1.582)
N individuals	372	372	372
N classes	20	20	20
N Schools	9	9	9
F-test for T1=T2 (p -value)	0.5172	0.5264	0.5535

Note: The table reports OLS estimates on test performance of the two treatment effects; (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. The outcome variable is test scores, standardized to mean 0 and standard deviation 1. All estimations are conducted on the full sample. Column (1) do not include any controls, column (2) controls for grades and gender, while column (3) further controls for class and school-level variables. Cluster–robust standard errors clustered at the class-level are displayed in parentheses. Asterisks indicate significance, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

Table 11 reports estimates for the same regressions conducted earlier but by gender. Respondents are simply classified as girls or boys if they identify themselves as one or the other on the survey. For boys, the estimations of the treatment effects is of similar sign as previously, however weaker and not statistically significant. This excludes the effect of treatment 2 in the first regression where it is slightly positive. For girls, the treatment effects are also negative, however much stronger, and looking at the regression in Column 5 and 6, which include the control variables, both treatment effects are statistically significant. The effect of treatment 1 and 2 as such leads to almost a third standard deviation lower test scores on average for girls. Comparing the regressions in Column 2 and Column 5, where individual control variables are included, girls experience almost three times stronger negative effects of treatment 1 and treatment 2. I tested whether the difference is statistically significant by conducting an additional regression reported in Column 7. This was done by creating an interaction term between gender and the treatment effects. As reported, the different effect of the treatments on boys and girls is not statistically significant. The different effects of the treatment 1 and 2 is not statistically significant in this case either. Table 15 in the appendix reports the same regression as Table 11, excluding Notreat. All estimations are of the same sign, scale and statistical significance, except for the effect of treatment 2 in Column 4, where the effect reduces in size.

Table 11: Impact of treatments on test scores (standardized), by gender

	Boys			Girls			Interaction
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
T1	-0.178	-0.124	-0.127	-0.248	-0.350**	-0.330**	-0.155
	(0.250)	(0.161)	(0.164)	(0.181)	(0.153)	(0.151)	(0.160)
T2	0.0759	-0.0908	-0.107	-0.499**	-0.272*	-0.260*	-0.117
	(0.151)	(0.0999)	(0.103)	(0.179)	(0.146)	(0.129)	(0.105)
Female							0.0842
							(0.144)
Female*T1							-0.185
							(0.244)
Female*T2							-0.153
							(0.181)
Grade							
A		2.583***	2.651***		2.351***	2.211***	2.476***
		(0.207)	(0.219)		(0.433)	(0.411)	(0.215)
B		2.371***	2.406***		2.243***	2.009***	2.245***
		(0.195)	(0.193)		(0.351)	(0.365)	(0.219)
C		1.937***	1.969***		1.608***	1.478***	1.757***
		(0.187)	(0.175)		(0.346)	(0.362)	(0.202)
D		1.378***	1.403***		1.392***	1.281***	1.359***
		(0.172)	(0.183)		(0.379)	(0.362)	0.668**
E		0.768***	0.806***		0.567	0.514	(0.248)
		(0.209)	(0.208)		(0.360)	(0.363)	(0.363)
Class size			-0.0445*			-0.00594	-0.0227
			(0.0256)			(0.0315)	(0.0193)
Foreign background			0.0108			0.0122	0.0131
			(0.0110)			(0.0125)	(0.00946)
Parent education			0.0236			0.0417**	0.0337***
			(0.0137)			(0.0154)	(0.00860)
Average GPA			-0.0233			-0.118	-0.0599
			(0.0911)			(0.170)	(0.112)
Constant	0.127	-1.541***	-2.089	0.0967	-1.229***	-2.684	-2.757
	(0.144)	(0.189)	(1.778)	(0.102)	(0.333)	(2.222)	(1.595)
N individuals	207	207	207	165	165	165	372
N classes	20	20	20	20	20	20	20
N Schools	9	9	9	9	9	9	9
F-test for T1=T2 (p - value)	0.1792	0.8317	0.8996	0.3946	0.6905	0.6957	-

Note: The table reports OLS estimates on test performance of the two treatment effects; (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. The outcome variable is test scores, standardized to mean 0 and standard deviation 1. All estimations are conducted on the full sample by gender. Column (1) do not include any controls, column (2) controls for grades, while column (3) further controls for class and school-level variables. Cluster–robust standard errors clustered at the class-level are displayed in parentheses. Asterisks indicate significance, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

Table 12 reports estimates for regressions conducted by grade-categories. With a large enough sample, it would have been optimal to do this by each grade-category but since observations in each category

are not large enough, grade-categories are merged together.[5] For grade-categories A-B, the effect of treatment 1 is positive while the effect of treatment 2 is close to zero, none of which are statistically significant. The effect of the treatments on grade-categories C-D are both negative, but only the effect of treatment 1 is statistically significant which on average leads to almost a third standard deviation lower test score for students with grades C-D. The same is true for grade-categories E-F, however the effects are much stronger for these two grade-categories. For this group, the effect of treatment 1 leads to a 0.43 standard deviation lower test score on. As such, the impact of the treatment on different grade-categories of students seem to be in line with expectation. Table 16 in the appendix reports the same regressions as Table 12, excluding Notreat. All estimates are of the same effect sign. The negative effect of the treatments is stronger for the lower grade-categories compared to higher grade-categories. The effect of treatment 1 is close to zero for grade-categories A-B, whilst the effect of treatment 2 is negative and stronger compared to the estimation in Table 14.

Table 12: Impact of treatments on test scores (standardized), by grades

	A-B		C-D		E-F	
	(1)	(2)	(3)	(4)	(5)	(6)
T1	0.0507	0.0328	-0.296**	-0.243**	-0.435*	-0.437*
	(0.189)	(0.191)	(0.110)	(0.112)	(0.206)	(0.221)
T2	-0.0280	-0.0648	-0.181	-0.171	-0.291	-0.306
	(0.134)	(0.127)	(0.127)	(0.109)	(0.232)	(0.272)
Grade						
A		0.214				
		(0.124)				
C				0.395***		
				(0.0972)		
E						0.580**
						(0.241)
Female		-0.0769		-0.0420		0.0522
		(0.216)		(0.0809)		(0.240)
Class size		-0.00321		-0.0359		-0.00207
		(0.0207)		(0.0374)		(0.0618)
Foreign background		0.0200*		0.0203		0.00257
		(0.00978)		(0.0167)		(0.0182)
Parent education		0.0326**		0.0422**		0.0196
		(0.0128)		(0.0182)		(0.0334)
Average GPA		-0.0202		-0.0651		-0.0808
		(0.0830)		(0.151)		(0.192)
Constant	0.830***	-1.611	0.266**	-1.698	-0.764***	-1.566
	(0.0906)	(1.430)	(0.106)	(2.673)	(0.169)	(2.694)
N individuals	86	86	194	194	92	92
N classes	19	19	20	20	18	18
N Schools	9	9	9	9	9	9
F-test for T1=T2 (p -value)	0.5904	0.7029	0.3942	0.5903	0.5576	0.5930

Note: The table reports OLS estimates on test performance of the two treatment effects; (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. The outcome variable is test scores, standardized to mean 0 and standard deviation 1. All estimations are conducted on the full sample by grades. Column (1) do not include any controls, while column (2) controls for gender, class size and school-level variables. Cluster–robust standard errors clustered at the class-level are displayed in parentheses. Asterisks indicate significance, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

[5] Regression by each grade-category were conducted with the estimations of the treatment effects being of the same sign as the corresponding estimations when grade-categories are merged together.

6.1 Exploratory Analysis

In this section, I will estimate the effect of the treatments on three of the survey questions answered by students after completion of the math test. Although not part of the main analysis of the paper, an exploratory analysis can eventually provide some insight into the direct effect of the treatments. The questions analysed are Questions 1, 2b and 3 in the survey. Question 1 is a proxy for how much effort students exerted during the test and is linked to the ideal effort levels indicated in their identity payoff model. Question 2b is a proxy for how unfair students found the thresholds to be. This is, however, only applicable to the control group and treatment group 1. Finally, Question 3 is a proxy for how much effort students exerted in the math test, as well as their self-confidence in regards to their own ability. Controlling for how much effort a student exerted during the test, Question 3 can thus provide insight into the effect of the treatments on the students' belief of their own ability. This is related to the ideal ability indicated in the students' identity payoff model. More detailed description of each question can be found in Section 4.2.2.

Table 13 reports the result of the main econometrical model estimated before with the only difference that the outcome variable is substituted with the three survey questions stated above. The outcome variable is stated at the top of each column. The first regression for each survey question includes the average causal effect of the treatments without any controls. The second regression includes all individual, class and school-level control variables. It should be noted for interpretation that each survey question has a corresponding answer ranging in a scale from 1 to 5. Hence, the estimations for each question can be compared to each other with ease. The average effect of both treatments is negative on Question 1, but not statistically significant indicating that the treatments on average lowered student motivation and hence exerted effort. The effect of treatment 1 is positive on Question 2b and statistically significant. The positive coefficient indicates that the treatment effect on average made students feel that the test was unfair, as thresholds were too high. Notice that students in treatment group 2 were excluded in this analysis as they did not per say have a specific threshold. Finally, the average effect of both treatments is negative on Question 3, but only the effect of treatment 2 statistically significant. Notice that in this regression, I also include a control for Question 1, which is a proxy for exerted effort levels. Thus, the negative effect of both treatments indicates that the treatments resulted in a lower belief amongst students in regards to their own ability. This effect is substantially stronger in treatment group 2 compared to treatment group 1.

Table 13: Impact of treatments on survey questions

	Q1	Q1	Q2b	Q2b	Q3	Q3
	(1)	(2)	(3)	(4)	(5)	(6)
T1	-0.235	-0.212	0.309*	0.334*	-0.0241	-0.139
	(0.219)	(0.221)	(0.175)	(0.181)	(0.215)	(0.207)
T2	-0.174	-0.192			-0.511**	-0.557***
	(0.171)	(0.168)			(0.188)	(0.177)
Q1					0.212	0.345**
					(0.152)	(0.130)
Grade						
A		1.012**		0.111		0.639
		(0.405)		(0.904)		(0.461)
B		0.779*		-0.0859		1.042***
		(0.440)		(0.821)		(0.281)
C		0.769*		0.159		0.445
		(0.427)		(0.773)		(0.260)
D		0.429		0.183		-0.0767
		(0.369)		(0.890)		(0.244)
E		0.435		0.257		-0.194
		(0.446)		(0.790)		(0.379)
Female		0.0423		0.233		-0.401***
		(0.169)		(0.135)		(0.111)
Class size		-0.0346		-0.0627*		0.184***
		(0.0328)		(0.0327)		(0.0548)
Foreign background		-0.0116		-0.0251**		-0.0177
		(0.0119)		(0.0113)		(0.0213)
Parent education		-0.0361*		-0.0652***		0.0412
		(0.0178)		(0.0164)		(0.0292)
Average GPA		0.0658		0.254*		0.155
		(0.112)		(0.123)		(0.114)
Constant	3.816***	5.997**	1.360***	4.439**	2.032***	-7.713*
	(0.125)	(2.170)	(0.105)	(1.702)	(0.475)	(3.702)
N individuals	372	372	249	249	372	372
N classes	20	20	20	20	20	20
N Schools	9	9	9	9	9	9

Note: The table reports OLS estimates on three survey questions of the two treatment effects; (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. The outcome variable is Question 1, Question 2b and Question 3. All estimations are conducted on the full sample, except when estimating Q2b where treatment group 2 is excluded. Column (1) do not include any controls, column (2) controls for all individual, class and school-level controls. Cluster–robust standard errors clustered at the class-level are displayed in parentheses. Asterisks indicate significance, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

7. Discussion

7.1 Main Findings

The average causal effect of grades, when serving as a starting point on which a student's future performance is dependent on, is negative on student motivation and performance. This is the case, both when grading is absolute (criterion-based assessment) and norm-based, as the treatment effects indicate.

Combined with the results of previous research indicating indecisive or a negative impact of grades as a non-monetary incentive, the results of this paper shed further doubt on the use of absolute grading in schools (Harlen and Deakin Crick, 2002; Jalava et al., 2015; Klapp, 2015; Klapp et al., 2016). In terms of norm-based grading, the average causal effect of previous grades on student motivation is also negative. The difference of its negative impact is, however, not statistically significantly different from when absolute grading is used instead, and as such contradicts expectations. As such, the result of this paper does not suggest that a norm-based grading system is a better solution, in regards to student motivation and performance, than an absolute grading system. This is to some degree in contrast with previous research that show that rankings improve student motivation compared to absolute grading (Azmat and Iriberri, 2010). Furthermore, the motivational power of the treatments differs with respect to gender. Girls seem to be affected more negatively by both treatments than boys, although this difference is not statistically significant. This I argue, could be due to that the effect of the treatments, channel through self-confidence which in turn differs by gender. Previous research by Niederle and Vesterlund (2010) show that girls fail to perform well in competitions, or shy away from competitive environments, which I argue could be due to a lower self-confidence in regards to their math ability. Dreber et al. (2014) confirm this as they suggest that a gender gap exists for mathematical tasks. As such, due to an already lower self-confidence, girls are then more severely affected by the treatment effects. Finally, the average negative causal effect of the treatments is higher, the lower grade the student has. In fact, the effect is close to zero for students in the grade brackets A-B, while it is negative for students in grade brackets C-F, which together represent approximately 75% of the general student population (Nydahl and Ridderlind, 2016). This confirms previous research, indicating that low-performing students suffer the most when grading is introduced, and is in line with the report from the Swedish National Agency for Education, where teachers report that as a result of grading, low-performing students lose motivation (Harlen and Deakin Crick, 2002; Klapp et al., 2016; Skolverket, 2017). This could potentially lead to an increased gap between different student-categories, and as such result in increased inequality.

The negative effect of grades as per this paper are in line with previous research, and as such strenghten the external validity of the results. There are however several other factors to consider when analysing the external validity of the results such as the age-group analysed, varying effects in different countries and how closely the experimental setup simulates the real school setting. First, the effects found in this paper could vary with age. The age-group analysed in this paper is 6th graders (12 year olds) that are being graded for the first time in schools. The negative effect of grades could arguably be both larger and smaller on this age-group as the concept of grades is something new for them. As I will argue in the following sections, I believe that this effect will be cumulative and lead to continued negative effect of grades as it shapes student behaviour. However, to conclude anything more specific about the effects of grades on other age-groups would require its own experiment. Second, the effects found could vary in

different countries. Sweden especially is unique in introducing grades at such a late stage. As such, experiments on the same age-group in other countries where grades are already introduced in earlier stages could lead to different results. A better comparison would be to conduct similar experiments in the age-groups where grades are initially introduced per each particular country. There could however still be factors that are different and thus, it is best not to assume external validity of the results in this paper beyond Sweden and its education system. Finally, it is important to discuss whether the experimental setup simulates the Swedish school system closely enough. As pointed out earlier, the negatively framed treatments do not simulate the real school setting to a full extent. Although final grades are based on a weighted average of several exams, the impact of previous grades/exams are not as strongly emphasized and do not directly affect the outcome of a final test like in the experimental design of this paper. Instead, previous grades combined with the outcome of the final test constitute the underlying material which students are assessed upon. Whether this difference is important and affects the external validity of this paper is worth further research.

Why is the effect of grades then, when serving as a starting point on which a student's future performance is dependent on, negative on student motivation and performance? This as I argue previously in the paper, can depend on whether grades can affect the self-confidence of students, or their notion of what is unfair, or whether they constitute a social identity which affects the behaviour of students per a set of prescriptions. I thus discuss these three mechanisms, which the negative effect of grades can potentially channel through, in the next section.

7.2 Self-confidence, Social Identities and Unfairness

To decipher the impact of each of these three mechanisms, it is important to exclude the threshold effect, which depends on the individual's previous grade. As a reminder, for the treatment groups, particularly treatment group 1, previous grades change the threshold that students need to reach in order to obtain the rewards. This could as I have previously stated, lead students to feel that the test design is unfair, which could eventually change the effort they exert, referenced to as the threshold effect. I argue that including the control variable for grade, partly captures and thus controls for, the different thresholds that each individual student has. As such, the negative effect of the treatments can mostly be accrued to their effect on student self-confidence, or whether grades constitute a social identity with subsequent prescriptions. I believe this argument is further strengthened, when estimating the treatment effects for students in grade-categories of C and D. Students with either a grade of C or D in treatment group 1, have as a result of their previous grade, a lower or the same threshold than their peers in the control group. In this case, any potential negative effect of the threshold is gone but the negative effect of the treatments is still there, larger than for the general student population, and statistically significant. To further decipher whether the negative effect of the treatments, is channelled through either their effect on student self-confidence, or whether grades constitute a social identity with subsequent prescriptions, requires a discussion of the exploratory analysis.

The exploratory analysis in section 6.1 should be interpreted as an attempt, to define the mechanisms the treatments affect students through. Question 1 will be linked to ideal effort levels, as dictated by the prescriptions of social categories specified earlier, relating to the concept of identity payoffs. Question 2b assesses the students' notion of whether the test design was unfair. Finally, Question 3 relates to impact of the treatments on their self-confidence.

Although including grade as a control variables, partly controls for the threshold effect, the effect of treatment 1 seems to cause a higher average notion of unfairness in treatment group 1 compared to the control group. This means that even after controlling for different thresholds, students whose outcome is affected by their previous grade find the math test more unfair. This I argue, is because students find the concept, that they are all being affected to different degrees per their previous grade, unfair. For example, a student with a previous grade of C finds it unfair that her threshold level is higher than a student with a previous grade of A.

Question 1 is supposed to measure how important it is for students to perform well on the test. Disregarding measurement error, if a student finds the test design very incentivizing, it should be more important for her to perform well on the test, and hence exert a higher level of effort. This relates to the identity payoff model stated earlier, where prescriptions for assigned social categories, dictate ideal effort levels on the math test with $e(A) > e(B) > e(C) > e(D) > e(E) > e(F)$. If grades in fact constitute social identities, and the prescriptions specified for these in the context of this experiment are correct, the identity payoff model predicts a lower effort exerted, the lower grade an individual has. As such, students with lower grades in the treatment groups, whom are assigned to these social identities, should by definition exert lower levels of effort. This is confirmed as the effect of the treatments is on average negative on Question 1, which suggests that grades constitute social identities.

Question 3 is supposed to, when controlling for exerted levels of effort (Question 1), be a proxy for student self-confidence, and their belief in regards to their ability in math. This question also relates to the identity payoff model stated earlier, where prescriptions for assigned social categories instead, dictate ideal characteristics that students should have. In the identity payoff model specified, the ideal characteristics is the degree of ability, where a higher grade-level corresponds with higher ability, such that $a(A) > a(B) > a(C) > a(D) > a(E) > a(F)$. Again, if grades in fact constitute social identities and the prescriptions specified for these in the context of this experiment are correct, the identity payoff model specifies a lower ability in math, the lower grade an individual has. Grades then function as a signalling mechanism, communicating the real ability of students, which in turn affects their self-confidence. The lowest grades could for instance, label students as failures which in turn affects their self-confidence. As such, students in the treatment groups that are assigned to these social identities, should by definition have a lower level of self-confidence (lower level of belief in regards to their ability in math). This is confirmed as the effect of the treatments is on average negative on Question 3, when

controlling for exerted levels of effort by including Question 1, which further suggests that grades constitute social identities.

In addition to students in the treatment groups finding the test design of this experiment more unfair, the analyses above suggest that grades in fact constitute social identities. The treatment effects in this paper, can be argued, assigns these social identities, or at a minimum reinforces their existence, which makes students put a higher weight on the identity payoff aspect of their utility. This in turn dictates the ideal effort levels they should exert, as well as reinforces the notion that grades communicate the degree of their ability, resulting in lower self-confidence. This for example, can explain why the average effect of the treatments on test scores are close to zero for students in grade-categories A and B. Students with grade A-B in the treatment groups are assigned to their respective social identity, which as specified in the model, dictates higher levels of effort than dictated for the rest of the students. As such, any difference between A-B students in the control group and the treatment group, is marginal. This leads me to conclude that the concept of identity economics as incorporated in the educational context, can potentially capture the behaviour of students in a better way than standard models of education. As such, it could provide additional insight into the causes of underinvestment in education.

7.3 Implications in the Real World

It should be noted that although this experiment is conducted in the students' natural environment, it is still far away from replicating the exact process of which students are assessed in and where they make decisions of how much effort to invest in education. The math test in this experiment is a low-stake test, and I argue that the cost of effort, although increasing in effort, is not that high. This because students if refusing to exert any effort, must sit throughout the duration of the session anyway, which for young children with lots of energy can be an equally if not a costlier alternative. In the real world, however, I argue that the cost of effort is much higher. In the real world, students need to attend school, do homework and study for exams, all of which require substantial amount of their time and effort up-front, with benefits that seem incremental, distant and uncertain. Instead, exciting options to procrastinate, in the form of games and smartphones are endless, all of which offer instant form of pleasures which surely makes the decision of whether to study more math difficult. Considering then that children and adolescents are shown to be more prone to short-term thinking, where a study by Bettinger and Slonim (2007) find that children's choices, are consistent with hyperbolic discounting, the results in this paper are of even more importance. If student motivation and performance in a low-stake test setting, where the cost of effort is arguably low, is negatively affected by grades, the same negative effect could be substantially larger in real school settings. This due to a combination of a higher cost of effort, lucrative alternative options to procrastinate, and a tendency for children to be more prone to short-term thinking.

A second additional aspect that should be pointed out, is that the non-monetary rewards (a diploma and a prize) used in this experimental study, function as a substitute for grades, assuming that grades in the real school setting are actually motivating factors for students looking forward. This assumption is

however a strong one, especially considering that previous research show indecisive results of the effect of grades on motivation, when it is used as a non-monetary incentive. If the average causal effect of grades, when serving as a starting point on which a student's future performance is dependent on, is negative on student motivation and performance, when students find the rewards incentivizing. What happens in the real school setting, where the incentives are replaced by grades, especially in years where they don't serve as a filtering mechanism for higher education? I argue that this constitute a second argument for, which the negative effects of grades found in this paper could potentially be even larger in the real school setting. Simply put, I argue that when the effect of grades is negative, even when students are incentivized to perform per the result of this experiment, when the reward for future performance is replaced to grades, this effect could be even larger.

Combined, the two aspects pointed out above, potentially indicates that the negative effect of grades found in this paper, could be larger in the real school settings. As such, the negative effect of grades as pointed out here requires especial attention from policy makers. On the other hand, as pointed out previously, the experimental setup does not fully simulate the school settings and as such requires further research.

7.4 Potential Long-term Effects of Grades

The long-term effect of grades is difficult to assess, especially when students are only assessed during one session, as in the experiment of this paper. I argue, however, that due to the mechanisms of which the effect of grades channel through, its long-term effect can be deduced. As previously mentioned, I argue that this paper show that grades potentially constitute social identities, which in turn prescribe the ideal characteristics and behaviour of students. Furthermore, the effect of grades as social identities, seem to negatively affect the self-confidence of students, labelling them as less able or even as failures. This is interesting because self-confidence could in turn constitute intrinsic motivation. Gottfried (1990) show that for young school children, academic intrinsic motivation is positively related to achievement and perception of competence (self-confidence). Further research in educational psychology show that, academic intrinsic motivation is a stable construct over time, and increasingly so with advancement in age (Gottfried, 1990; Gottfried et al., 2001). The stability of academic intrinsic motivation, especially with advancement in age, thus places children with low levels of motivation early in their schooling at risk. In addition to this, a study by Butler and Nisan (1986) show that intrinsic motivation is undermined after receipt of controlling normative grades.

Combined, the result of this paper and previous research in educational psychology, indicate that grades can both through a direct and an indirect way, reduce the academic intrinsic motivation of students. Indirectly through the negative effect of grades on self-confidence, which in turn affect the academic intrinsic motivation of students. Considering that both self-confidence and academic intrinsic motivation is shown to be stable over time, introducing grades at an early age, could have negative long-term effects

on student motivation and performance, especially for students that are initially endowed with low self-confidence.

Another aspect of which grades, through their construct of social identities, negatively affect motivation in a long-term fashion, is by prescribing and dictating ideal effort levels exerted by students. The part of the brain which helps individuals focus on the future does not fully mature until an individual has reached their mid-twenties (Lavecchia et al., 2014). As such, children are more susceptible to behavioural effects, which may change the level of effort they exert in a negative direction. I argue that grades constitute such a behavioural effect. As their role as social identities, grades dictate ideal levels of effort, that distort the optimal level of effort children should exert. This behavioural change becomes increasingly stable over time, as it molds the behaviour of children, which are more susceptible to these behavioural effects. This in turn leads to negative effects, particularly for children whose social identity dictates low levels of effort exerted.

The implication of the arguments in this section, is that with the current school system, grades assigned to students have negative long-term effect on their motivation and performance. As such, grades that represent extrinsic motivation risk crowding out intrinsic motivation. This effect is because grades constitute social identities, which in turn distort the optimal behaviour of students, reduces their self-confidence and as such their academic intrinsic motivation. These mechanisms in turn have long-term effects on student motivation and performance. The result of this paper as such, suggests that the current design of the school system, may lead to long-term underinvestment issues regarding student effort.

7.5 What is the alternative?

Considering that the long-term effect of grades on student motivation and performance is shown to be negative, what is then the alternative? In order to provide suggestions for alternative educational systems, I need to go back to the time when the idea of this thesis was formed. The idea of which the purpose and design of this paper was conceptualized on, was first introduced to me by Salman Khan. For most individuals around my age, he is famous as the founder of Khan Academy, an educational organization with the sole purpose of providing free and accessible education for people around the world. With the help of over thousands of YouTube videos, Salman Khan, educates students in different fields, ranging from history to algebra. In two of his TED talks, Khan introduces the idea of an educational system that emphasise mastery learning (Khan, 2011; Khan, 2015). In this section, I will briefly cover some of his ideas which essentially paints the picture for a better alternative educational model than the one currently in place. These ideas are however not tested or examined in this paper but are included in this section to provide some suggestive measures that are of interest for further research.

The current traditional academic model is designed in a way that leads to accumulated gaps of knowledge. Students are grouped by age and shepherd together at the same pace. To illustrate this, imagine students in a math class, currently studying multiplication. After a period of time with lectures

and homework, they are tested on their multiplication skills. On that test, student test score ranges from 50% to 95%. Although the test identified gaps in each students' knowledge regarding multiplication, the class will move on to the next subject, for instance division, which assumably builds on those gaps. This process continues for years, where for instance a student who didn't understand 30% of the foundational concepts of math, were assigned a grade of C and pushed forward to more advanced levels. Eventually, because of the accumulation of all these gaps, the student might hit a breaking point, where the 30% that the student did not understand before matters, resulting in a loss of self-confidence.

The idea of mastery learning is however the opposite. Instead of constraining how long students work on different concepts, upon which they are graded on, students are instead encouraged to work on concepts until they master them, upon which they can move to more advanced levels. Khan (2015) argues that in addition to making students learn better, this model also reinforces the right mindset, making students realize that getting 30% wrong on a test doesn't mean that they have a C branded in their DNA, it just means that they should keep working to bridge that gap. This model is however not just something that Khan makes up, but builds upon well-established theories developed in educational psychology (Block et al., 1971). For instance, Ames and Archer (1988) show that students in classrooms with an emphasis on mastery learning, show better problem-solving skills and have a stronger belief that success is dependent on one's effort, compared to students of whom are influenced by performance goals.

Why has this model of mastery learning then, despite its benefits, not ben implemented so far? Khan (2015) point out that one such reason has been impractibility. To offer personalized learning for each student, would have not been scalable and in turn logistically difficult. He argues, however, that it is no longer impractical, as we now possess the technology required to implement it. Through his educational platform, Khan reports positive results of mastery learning, where students who take a bit extra time on one concept, when finally mastering the concept, race ahead. As such, a student whom the traditional academic model would have labelled as a low-performing one, would have now been considered gifted. As such, I think that the ideas presented by Khan of mastery learning, and as developed within the field of educational psychology, combined with the opportunity new technologies equate to, provides us with potentially what is an interesting solution, worth investigating further.

8. Conclusion

The design of the school system, and as such the assessment of students, is an important national topic, especially in Sweden. Much of the educational debate concerns the application of grades, where they serve the purpose of both motivating students, but also to measure and indicate student performance and ability. However, if grades do not motivate students to exert effort, its purpose of indicating student performance and ability as such, also becomes insignificant. It is therefore important to understand the

effect of grades on student motivation and performance, and to identity through which mechanisms these effects channel through.

This paper finds that the average causal effect of grades, when serving as a starting point on which a student's future performance is dependent on, is negative on student motivation and performance. This is regardless of, whether the grading system in place is criterion-based or norm-based. This negative effect also differs with respect to gender, where girls seem to be affected to a larger extent than boys, although this difference is not statistically significant. In addition to this, the average causal effect of the treatments is also higher, the lower grade the student has. Combined with the results of previous research, indicating indecisive or a negative impact of grades as a non-monetary incentive, the results of this paper shed further doubt on the role and application of grades in schools. The use of grades in schools, could instead potentially lead to differentiating effects on girls and/or low-performing students, leading to an increased gap between different student-categories, and as such to increased inequality.

Furthermore, the mechanisms as identified in this paper, that the effect of grades channel through, are student self-confidence, that grades can constitute social identities affecting behaviour, and the notion that grades with their current design unfairly assess student performance. Subsequently, I argue that, due to the potential negative effect of grades on self-confidence, and as such on intrinsic motivation, both which are increasingly stable over time, the negative effect of grades on student motivation and performance, can constitute a long-term one. More importantly, this paper suggests that the concept of identity economics, as incorporated in the educational context, can in a better way predict student behaviour than standard models of education.

Thus, I believe that there are endless avenues for further research on this topic that are of great importance. One such avenue, is to strengthen the link between the impact of grades and possible mechanisms, which its effect channel through. A second avenue for further research, is to assess the impact of other educational assessment systems on student motivation and performance, as developed in the literature of educational psychology.

References

Akerlof GA and Kranton RE. (2000) Economics and identity. *Quarterly Journal of Economics* 115: 715-753.

Akerlof GA and Kranton RE. (2002) Identity and schooling: Some lessons for the economics of education. *Journal of Economic literature* 40: 1167-1201.

Ames C and Archer J. (1988) Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology* 80: 260.

Ashraf N, Bandiera O and Jack BK. (2014) No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics* 120: 1-17.

Azmat G and Iriberri N. (2010) The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics* 94: 435-452.

Becker GS. (1967) *Human capital and the personal distribution of income: An analytical approach*: Institute of Public Administration.

Bénabou R and Tirole J. (2002) Self-confidence and personal motivation. *Quarterly Journal of Economics* 117: 871-915.

Bettinger E and Slonim R. (2007) Patience among children. *Journal of Public Economics* 91: 343-363.

Bettinger EP. (2012) Paying to learn: The effect of financial incentives on elementary school test scores. *Review of Economics and Statistics* 94: 686-698.

Block JH, Airasian PW, Bloom BS, et al. (1971) *Mastery learning: Theory and practice*: Holt, Rinehart and Winston New York.

Butler R and Nisan M. (1986) Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology* 78: 210.

DellaVigna S. (2009) Psychology and economics: Evidence from the field. *Journal of Economic literature* 47: 315-372.

Dreber A, von Essen E and Ranehill E. (2014) Gender and competition in adolescence: task matters. *Experimental Economics* 17: 154-172.

Durlak JA. (2009) How to select, calculate, and interpret effect sizes. *Journal of pediatric psychology*: jsp004.

Dweck CS. (2008) *Mindset: The new psychology of success*: Random House Digital, Inc.

Eisenkopf G, Hessami Z, Fischbacher U, et al. (2015) Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland. *Journal of Economic Behavior & Organization* 115: 123-143.

Ellingsen T and Johannesson M. (2007) Paying respect. *The Journal of Economic Perspectives* 21: 135-149.

Filippin A and Paccagnella M. (2012) Family background, self-confidence and economic outcomes. *Economics of Education Review* 31: 824-834.

Frey BS. (2007) Awards as compensation. *European Management Review* 4: 6-14.

Fryer Jr RG. (2010) Financial incentives and student achievement: Evidence from randomized trials. National Bureau of Economic Research.

Gottfried AE. (1990) Academic intrinsic motivation in young elementary school children. *Journal of Educational Psychology* 82: 525.

Gottfried AE, Fleming JS and Gottfried AW. (2001) Continuity of academic intrinsic motivation from childhood through late adolescence: A longitudinal study. *Journal of Educational Psychology* 93: 3.

Harlen W and Deakin Crick R. (2002) A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1). *Research Evidence in Education Library* 1.

Heckman JJ, Lochner LJ and Todd PE. (2006) Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond. In: Hanushek E and Welch F (eds) *Handbook of the Economics of Education.* Elsevier, 307-458.

Jalava N, Joensen JS and Pellas E. (2015) Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization* 115: 161-196.

Khan S. (2011) Let's use video to reinvent education. TED.

Khan S. (2015) Let's teach for mastery - not test scores. TED: TED.

Klapp A. (2015) Does grading affect educational attainment? A longitudinal study. *Assessment in Education: Principles, Policy & Practice* 22: 302-323.

Klapp A, Cliffordson C and Gustafsson J-E. (2016) The effect of being graded on later achievement: Evidence from 13-year olds in Swedish compulsory school. *Educational Psychology* 36: 1771-1789.

Koch A, Nafziger J and Nielsen HS. (2015) Behavioral economics of education. *Journal of Economic Behavior & Organization* 115: 3-17.

Kosfeld M and Neckermann S. (2011) Getting more work for nothing? Symbolic awards and worker performance. *American Economic Journal: Microeconomics* 3: 86-99.

Lavecchia AM, Liu H and Oreopoulos P. (2014) Behavioral economics of education: Progress and possibilities. National Bureau of Economic Research.

Levitt SD, List JA, Neckermann S, et al. (2016) The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy* 8: 183-219.

Lochner L. (2011) Non-production benefits of education: Crime, health, and good citizenship. National Bureau of Economic Research.

Niederle M and Vesterlund L. (2010) Explaining the gender gap in math test scores: The role of competition. *The Journal of Economic Perspectives* 24: 129-144.

Nydahl A and Ridderlind I. (2016) Ämnesprovet i matematik i årskurs 6, 2015/2016. Stockholm University: Stockholm University, 3-4.

Oreopoulos P. (2007) Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *Journal of Public Economics* 91: 2213-2229.

Skolverket. (2017) Utvärdering av betyg från årskurs 6. Stockholm: Skolverket, 6-10.

Tran A and Zeckhauser R. (2012) Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics* 96: 645-650.

Tversky A and Kahneman D. (1985) The framing of decisions and the psychology of choice. *Environmental Impact Assessment, Technology Assessment, and Risk Analysis.* Springer, 107-129.

Wang XH and Yang B. (2003) Why competition may discourage students from learning? A behavioral economic analysis. *Education Economics* 11: 117-128.

Weiss Y and Fershtman C. (1998) Social status and economic performance:: A survey. *European Economic Review* 42: 801-820.

Wise SL and DeMars CE. (2005) Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment* 10: 1-17.

Appendix

Table 14: Impact of treatments on test scores (standardized), excluding Notreat.

	(1)	(2)	(3)
T1	-0.226*	-0.289***	-0.278***
	(0.120)	(0.0921)	(0.0932)
T2	0.0110	-0.178**	-0.187**
	(0.0940)	(0.0775)	(0.0697)
Grade			
A		2.180***	2.142***
		(0.279)	(0.327)
B		2.037***	1.996***
		(0.266)	(0.321)
C		1.533***	1.520***
		(0.287)	(0.320)
D		1.176***	1.130***
		(0.264)	(0.300)
E		0.469	0.461
		(0.307)	(0.342)
Female		0.0301	-0.0302
		(0.0964)	(0.0996)
Class size			-0.0174
			(0.0254)
Foreign background			0.0175
			(0.0118)
Parent education			0.0420***
			(0.0138)
Average GPA			-0.0859
			(0.0808)
Constant	0.175*	-1.125***	-2.909
	(0.0878)	(0.266)	(1.947)
N individuals	307	307	307
N classes	19	19	19
N Schools	8	8	8
F-test for T1=T2 (p -value)	0.0636	0.2953	0.3427

Note: The table reports OLS estimates on test performance of the two treatment effects; (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. The outcome variable is test scores, standardized to mean 0 and standard deviation 1. All estimations are conducted on the sample excluding NoTreat. Column (1) do not include any controls, column (2) controls for grades and gender, while column (3) further controls for class and school-level variables. Cluster–robust standard errors clustered at the class-level are displayed in parentheses. Asterisks indicate significance, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

Table 15: Impact of treatments on test scores (standardized), by gender, excluding Notreat

	Boys			Girls			Interaction
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
T1	-0.204	-0.156	-0.158	-0.253	-0.384**	-0.350*	-0.187
	(0.230)	(0.156)	(0.163)	(0.176)	(0.153)	(0.166)	(0.162)
T2	0.0966	-0.118	-0.141	-0.256	-0.233	-0.288	-0.127
	(0.154)	(0.102)	(0.106)	(0.184)	(0.205)	(0.196)	(0.0904)
Female							0.0720
							(0.140)
Female*T1							-0.185
							(0.247)
Female*T2							-0.136
							(0.237)
Grade							
A		2.427***	2.501***		2.038***	1.885***	2.198***
		(0.444)	(0.377)		(0.530)	(0.552)	(0.325)
B		2.246***	2.339***		1.892***	1.674***	2.022***
		(0.432)	(0.338)		(0.379)	(0.434)	(0.318)
C		1.836***	1.921***		1.251***	1.173**	1.551***
		(0.444)	(0.343)		(0.391)	(0.469)	(0.315)
D		1.262***	1.327***		1.126**	1.015**	1.154***
		(0.421)	(0.335)		(0.413)	(0.479)	(0.295)
E		0.658	0.762**		0.356	0.272	0.499
		(0.393)	(0.317)		(0.454)	(0.505)	(0.349)
Class size			-0.0142			-0.0284	-0.0190
			(0.0278)			(0.0520)	(0.0259)
Foreign background			0.0285**			0.00239	0.0169
			(0.0126)			(0.0193)	(0.0119)
Parent education			0.0461**			0.0354	0.0410***
			(0.0178)			(0.0256)	(0.0140)
Average GPA			-0.0238			-0.183	-0.0854
			(0.0650)			(0.125)	(0.0835)
Constant	0.194	-1.401***	-4.707**	0.155	-0.901**	-0.304	-2.879
	(0.148)	(0.400)	(1.983)	(0.0951)	(0.371)	(3.210)	(1.959)
N individuals	175	175	175	132	132	132	307
N classes	19	19	19	19	19	19	19
N Schools	8	8	8	8	8	8	8
F-test for T1=T2 (p - value)	0.0752	0.8191	0.9169	0.9892	0.5340	0.7857	-

Note: The table reports OLS estimates on test performance of the two treatment effects; (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. The outcome variable is test scores, standardized to mean 0 and standard deviation 1. All estimations are conducted on the sample excluding Notreat, by gender. Column (1) do not include any controls, column (2) controls for grades, while column (3) further controls for class and school-level variables. Cluster–robust standard errors clustered at the class-level are displayed in parentheses. Asterisks indicate significance, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

Table 16: Impact of treatments on test scores (standardized), by grades, excluding Notreat.

	A-B		C-D		E-F	
	(1)	(2)	(3)	(4)	(5)	(6)
T1	0.0570	0.0428	-0.371***	-0.270**	-0.487**	-0.492*
	(0.192)	(0.190)	(0.119)	(0.124)	(0.211)	(0.238)
T2	-0.103	-0.126	-0.174	-0.148	-0.215	-0.215
	(0.147)	(0.144)	(0.130)	(0.105)	(0.280)	(0.297)
Grade						
A		0.134				
		(0.148)				
C				0.411***		
				(0.0919)		
E						0.364
						(0.363)
Female		-0.120		-0.0761		0.242
		(0.226)		(0.113)		(0.271)
Class size		0.00663		0.0151		-0.160*
		(0.0378)		(0.0394)		(0.0810)
Foreign background		0.0244		0.0403**		-0.0328
		(0.0151)		(0.0173)		(0.0307)
Parent education		0.0390**		0.0705***		-0.0246
		(0.0171)		(0.0230)		(0.0483)
Average GPA		-0.0227		-0.0983		0.0198
		(0.100)		(0.0911)		(0.146)
Constant	0.823***	-2.350	0.303***	-4.937*	-0.631***	5.065
	(0.0940)	(2.670)	(0.0990)	(2.752)	(0.172)	(4.820)
N individuals	77	77	168	168	62	62
N classes	18	18	19	19	17	17
N Schools	8	8	8	8	8	8
F-test for T1=T2 (p -value)	0.2844	0.3736	0.1690	0.2839	0.3413	0.3540

Note: The table reports OLS estimates on test performance of the two treatment effects; (T1) previous grade + absolute threshold, (T2) previous grade + rank based threshold. The outcome variable is test scores, standardized to mean 0 and standard deviation 1. All estimations are conducted on the sample excluding NoTreat, by grades. Column (1) do not include any controls, while column (2) controls for grades, gender, class size and school-level variables. Cluster–robust standard errors clustered at the class-level are displayed in parentheses. Asterisks indicate significance, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

Table 17: Information regarding test assessment/Treatment effects - Swedish

Control group
På detta test kan du få totalt 30 poäng.
Om du får 20 poäng eller mer får du ett diplom och ett pris.

Treatment group 1
På detta test kan du få max 30 poäng.
Du kommer dock få avdrag (minuspoäng) baserat på ditt tidigare betyg i matematik från förra terminen.

- Om du fick ett A i matematik förra terminen, får du 0 minuspoäng.
- Om du fick ett B i matematik förra terminen, får du 1 minuspoäng.
- Om du fick ett C i matematik förra terminen, får du 2 minuspoäng.
- Om du fick ett D i matematik förra terminen, får du 3 minuspoäng.
- Om du fick ett E i matematik förra terminen, får du 4 minuspoäng.
- Om du fick ett F i matematik förra terminen, får du 5 minuspoäng.

Om du sammanlagt får 17 poäng eller mer får du ett diplom och ett pris.

Exempel: Om Kalle får 19 av 30 poäng på det här testet och fick ett C (2 minuspoäng) i matematik förra terminen så blir hans **sammanlagda** poäng 17.

19 – 2 = 17

Treatment group 2
På detta test kan du få max 30 poäng.
Du kommer dock få avdrag (minuspoäng) baserat på ditt tidigare betyg i matematik från förra terminen.

- Om du fick ett A i matematik förra terminen, får du 0 minuspoäng.
- Om du fick ett B i matematik förra terminen, får du 1 minuspoäng.
- Om du fick ett C i matematik förra terminen, får du 2 minuspoäng.
- Om du fick ett D i matematik förra terminen, får du 3 minuspoäng.
- Om du fick ett E i matematik förra terminen, får du 4 minuspoäng.
- Om du fick ett F i matematik förra terminen, får du 5 minuspoäng.

Om du är bland de tre med högst antal poäng sammanlagt i klassen får du ett diplom och ett pris.

Exempel: Om Kalle får 19 av 30 poäng på det här testet och fick ett C (2 minuspoäng) i matematik förra terminen så blir hans **sammanlagda** poäng 17.

19 – 2 = 17

Table 18: Math test questions - English

Task 1.
Emma has two sisters and three brothers. Emma's brother David has three sisters and two brothers.

Question: How many boys and how many girls are there in the family?

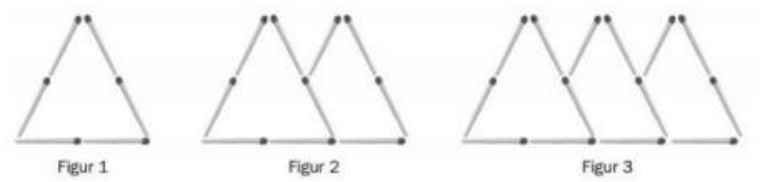
Answer: ____ boys ___ / 4 points
____ girls

Task 2.
Solve the equations.

a) x – 6 = 7
Answer: x = _____

b) 24 + 6 = x + 14
Answer: x = _____ ___ / 4 points

Task 3.



Question: How many matches in total are needed to build a new figure 4 according to the pattern above?

Answer: _____ ___ / 5 points

Task 4.
Calculate the following in the way you prefer.

a) 406 – 288 = ?
Answer: _____

b) 8 · 70 = ?
Answer: _____ ___ / 5 points

Task 5.
One evening Amanda started to read her book on page. She stopped when she had finished reading page 11. The following evening, she read until she had finished reading page 20. However, two of the pages read contained only pictures.

Question: How many pages of text did Amanda read in total during the two evenings?

Answer: _____ ___ / 6 points

Task 6.
Sara, Adam and Leo had together saved 180 SEK.
They want to buy fruit for the whole sum of money.
A fruit basket costs 30 SEK /each.

Question: How many fruit baskets does each of them get if they share equally?

Answer: _____ ___ / 6 points

Table 19: Math test questions - Swedish

Uppgift 1.
Emma har två systrar och tre bröder. Emmas bror David har tre systrar och två bröder.

Fråga: Hur många pojkar och hur många flickor finns det i familjen?

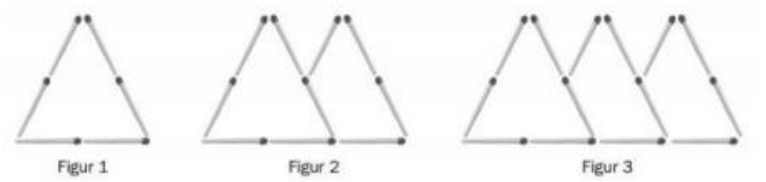
Svar: ____ pojkar ___ / 4 poäng
____ flickor

Uppgift 2.
Lös ekvationerna.

a) x – 6 = 7
Svar: x = _____

b) 24 + 6 = x + 14
Svar: x = _____ ___ / 4 poäng

Uppgift 3.



Fråga: Hur många stickor totalt behövs för att bygga en ny figur 4 enligt mönstret ovan?
Svar: _____ ___ / 5 poäng

Uppgift 4.
Räkna ut på det sätt du tycker är bäst.

a) 406 – 288 = ?
Svar: _____

b) 8 · 70 = ?
Svar: _____ ___ / 5 poäng

Uppgift 5.
En kväll började Amanda läsa sin bok på sida 4, hon slutade när hon läst klart sida 11. Nästa kväll läste hon fram till att hon läst klart sida 20, men två av de lästa sidorna innehöll endast bilder.

Fråga: Hur många sidor text läste Amanda totalt på de två kvällarna?

Svar: _____ ___ / 6 poäng

Uppgift 6.
Sara, Adam och Leo har sparat 180 kr tillsammans.
De vill köpa frukt för hela summan.
En fruktpåse kostar 30 kr / styck.

Fråga: Hur många fruktpåsar får var och en om de delar lika?

Svar: _____ ___ / 6 poäng

Table 20: Survey questions - Swedish

Namn:
Pojke: ☐
Flicka: ☐

1. Hur viktigt var det för dig att göra bra ifrån dig på testet som vi gjorde alldeles nyss?
Inte alls 1 2 3 4 5 Väldigt viktigt

2a. Fanns det en viss poänggräns som du behövde nå för att få ett diplom/pris?
☐ Ja - (över 20 eller 17 poäng) ☐ Nej - (topp tre med högst antal poäng får diplom/pris)
Om du svarade Nej, hoppa över fråga 2b.

2b. Tycker du att gränsen av poäng du behövde få för att få ett diplom/pris var för högt?
Inte alls 1 2 3 4 5 Väldigt mycket

3. Hur stor tror du chansen är att du kommer att få ett diplom/pris?
Inte alls 1 2 3 4 5 Väldigt stor

4. Blir du motiverad av att få betyg?
Inte alls 1 2 3 4 5 Väldigt mycket

5a. Brukar du jämföra dina betyg och resultat med dina klasskamrater?
☐ Ja ☐ Nej
Om du svarade Nej, hoppa över fråga 5b.

5b. Hur motiverad blir du av att jämföra dina betyg och resultat med dina klasskamrater?
Inte alls 1 2 3 4 5 Väldigt mycket