BUDAPEST WORKING PAPERS ON THE LABOUR MARKET
BWP – 2015/5

Sticky assessments – the impact of teachers' grading standard on pupils' school performance

TAMÁS KELLER

INSTITUTE OF ECONOMICS, CENTRE FOR ECONOMIC AND REGIONAL STUDIES
HUNGARIAN ACADEMY OF SCIENCES
DEPARTMENT OF HUMAN RESOURCES, CORVINUS UNIVERSITY OF BUDAPEST
BUDAPEST, 2015

Budapest Working Papers on the Labour Market
BWP – 2015/5

Institute of Economics, Centre for Economic and Regional Studies,
Hungarian Academy of Sciences
Department of Human Resources, Corvinus University of Budapest

Sticky assessments – the impact of teachers' grading standard on pupils' school performance

Author:

Tamás Keller
TÁRKI Social Research Institute
and Research Centre for Educational and Network Studies,
Hungarian Academy of Science
email: keller@tarki.hu

July 2015

ISBN 978 615 5594 03 8
ISSN 1785 3788

Sticky assessments – the impact of teachers' grading standard on pupils' school performance

Tamás Keller

Abstract

This paper argues that school grades cannot be interpreted solely as a reward for a given school performance, since they also reflect teachers' ratings of pupils. Grades therefore contain valuable information about pupils' own – usually unknown – ability. The incorporated assessment in grade might be translated into self-assessment, which could influence the effort that pupils invest in education. Getting discounted grades in year 6 for a given level of math performance assessed using a PISA-like test has a positive effect on math test scores in year 8 of elementary education and also influences later outcomes in secondary education. The empirical analysis tries to minimize the possible bias caused by the measurement error in year 6 test scores (unmeasured ability) and employs classroom fixed-effect instrumental variable (IV) regression and difference-in-difference models. The main analysis is based on a unique Hungarian individual-level panel dataset with two observations about the same individual – one in year 6 (12/13 years old) and again two years later, in year 8 (14/15 years old) of elementary education. The data for three entire school cohorts is analyzed – approximately 140,000 individuals.

Highlights

- Examines the impact of teachers' grading standards on pupils' school performance
- Takes advantage of having two different measures of pupils' math knowledge: teacher-given grades and centralized test scores
- Assumes that grades are more than test scores, since they incorporate teachers' ratings
- Tries to estimate teachers' grading standards and minimizes unmeasured ability bias by employing IV regression and diff-in-diff approaches
- Finds that year 6 grades positively influence year 8 test scores and year 10 outcomes
- Argues that teachers' assessments translate to self-assessment, which influences pupils' effort
- Concludes that grading standards in elementary school accompany pupils to secondary school

Keywords: School performance; Inflated school grades; Feedback, Good teacher; Educational panel data; Hungarian National Assessment of Basic Competencies

JEL classification: I20, I21, J24

Acknowledgement:

This work was supported by a grant from the OTKA (Hungarian Scientific Research Fund). Grant number: PD 105976. Suggestions from Zoltán Hermann, Dániel Horn and Balázs Muraközy are warmly acknowledged. All remaining errors are solely mine.

A diákokhoz nőtt osztályzatok – a tanári osztályzás hatása a diákok iskolai teljesítményére

Keller Tamás

Összefoglaló

A tanulmány érvelése szerint az iskolai osztályzatok nem kizárólag egy adott iskolai teljesítmény objektív mérései, hanem egyúttal szubjektív tanári visszajelzések is. Az iskolai jegyekben visszatükröződő tanári értékelés kifejezetten fontos információt adhat a diákoknak saját – sokszor nem pontosan ismert – képességükről. A tanári értékelések ugyanakkor könnyen befolyásolhatják a diákok önértékelését is, amely meghatározza, hogy a diákok mennyit fektetnek a tanulásba. Eredményeink szerint, ha valaki hatodik osztályban az Országos Kompetencia teszten elért matematika teljesítményéhez képest diszkontált matematika jegyet kap, ez pozitívan befolyásolja nyolcadik osztályos teszt-eredményét, sőt középiskolai mutatót is. Empirikus elemzésünkben igyekszünk minimalizálni az abból fakadó torzítást, hogy a hatodik osztályos teszteredmények jelentős mérési hibával mérik a diákok tényleges képességeit, ez pedig torzíthatja az iskolai osztályzatokban vélt tanári értékelés hatását. Ezt a problémát kezelendő, osztály fix hatásokat tartalmazó instrumentális becslések, illetve difference-in-difference modellek eredményeit vizsgáljuk. A tanulmány fő elemzésében az Országos Kompetencia Mérésből azokkal az általános iskolás diákokkal foglalkozunk, akik hatodik és nyolcadik között nem váltottak osztályt. Három teljes iskolai kohorszot vizsgálunk, ami nagyjából 140 ezer megfigyelést jelent.

Tárgyszavak: Iskolai teljesítmény; Diszkontált/inflált osztályzatok; Tanári visszajelzés, A jó tanár; Iskolai panel vizsgálatok; Országos Kompetencia Mérés

JEL kódok: I20, I21, J24

1. INTRODUCTION

Grading pupils' performance in school is a well-known practice among teachers. It is established that grades reflect pupils' achievement; however since the variation within teachers' grading practices is large, there is a considerable heterogeneity in how teachers weight the importance of different grading principles (McMillan et al. 2002). On the other hand, little is known about the impact of teachers' grading style on pupils' later school performance. Particularly little research has investigated the consequences of getting inflated grades for a given level of performance. Recently Terrier (2014) analyzed a similar question at classroom level, and found that those classrooms where teachers rewarded girls with better grades than boys for a given academic achievement are also classrooms where girls progressed more.

The main analysis uses individual panel data for three entire school cohorts in Hungary, and analyzes the change in individual math test scores for more than 140,000 pupils between year 6 and year 8 of elementary school. It takes advantage of having two measures from the same time about pupils' math knowledge: the one is teacher given and the other is assessed by a centralized PISA-like test, developed and conducted by the Hungarian Educational Authority. It is assumed that since both measures reflect pupils' math knowledge, in a regression equation where later math knowledge is regressed on lagged math knowledge and lagged grades, the impact of grades captures the effect of a teacher's rating. The paper will also go beyond this assumption and will try to minimize the bias arising from the fact that lagged test scores are only an imperfect measure of pupils' math knowledge, and therefore unmeasured ability might bias the estimated parameter for grades. To deal with this issue, instrumental variable and difference-in-difference (hereafter diff-in-diff) methods will be employed.

It is argued that the teachers' assessments contained in the grades could be transformed into self-assessment. Since pupils usually do not know their true level of ability, feedback about their performance – contained in grades – could be an important source of knowledge about their own skills. If this feedback is positive – e.g. pupils receive better grades for a given level of school performance – it will boost their self-assessment. Higher perception of own ability might contribute to the choice of the optimal level of effort, which might influence later school performance. Since school performance is assumed to be a combination of ability and effort, and effort is costly, if pupils are assured that the effort will be worthwhile (because they are more able than their peers), they will likely invest more effort in education, which could increase their later school achievement (Azmat and Iriberri 2010).

Answering this question about the possible outcomes of grades might be important, since there is considerable literature (Rockoff 2004; Wright, Horn, and Sanders 1997) showing that teachers do influence pupils' academic achievement. Much less attention has been devoted to the specific teacher characteristics that could increase pupils' achievement. Previous research on this topic has shown that the predictive power of teachers' observable characteristics – like age or type of degree – have little or no power to explain pupils' academic achievement (Rivkin, Hanushek, and Kain 2005). Teachers' grading style, however, could be an important characteristic, especially if it influences pupils' school performance.

1.1. SELF-FULFILLING PROPHECIES AND RELATIVE PERFORMANCE FEEDBACK

Prior research about grading standard (Betts and Grogger 2003; Terrier 2014) analyzed the impact of grades on subsequent achievement as a classroom-level characteristic and not as an individual-level characteristic, such as how being over- or under-rated could influence later school performance. If teachers' biased grading standards influence pupils achievement, this could act as a self-fulfilling prophecy (e.g. those who are over-rated will perform better).

There is an extended literature on how self-fulfilling prophecies induce students' later achievement. Rosenthal and Jacobson (1968) conducted an experiment in which they told elementary school teachers that some pupils in the class would soon demonstrate a large gain in school achievement. Although they told the teachers that this information about their pupils came from the results of a test they had run on the children, in fact the test was non-existent and the researcher had randomly assigned pupils to the treated and the control groups. One year later, pupils predicted to be 'bloomers' showed larger progress in IQ than the control group.

However the findings of Rosenthal and Jacobson's research were not universally acclaimed (Snow 1969). They provoked scientific debate, which was reviewed by Jussim and Harber (2005). The criticism mainly concerns the fact that after the first year not only the experimental group, but also the control group, experienced considerable progress in IQ, and in later years the difference between the two groups averaged out. Research on the impact of self-fulfilling prophecies showed that they have a weak positive impact on later school achievement (Jussim and Eccles 1992), which varies according to social status and ethnicity (Jussim, Eccles, and Madon 1996). It has also been shown that positive prophecies were more powerful than negative ones (Madon, Jussim, and Eccles 1997).

On the other hand, feedback could be received not only in the form of grades or in how teachers treat pupils, but also more indirectly, with information gleaned by pupils about their own relative performance. Taking advantage of natural experiments, it has been shown that if pupils are informed of the average grade point in their classroom (Azmat and Iriberri 2010),

or if they know their own relative position on a nationwide test (Goulas and Megalokonomou 2015) the information provided positively affects their later school achievement. Since later performance is a combination of ability and effort, and since individual effort is costly, if pupils realize that it is worth their while (because they are good relative to others), they might invest more effort in studying hard, which could be translated later into better school performance.

However, self-fulfilling prophecies and relative performance feedback have different implications about the role of teachers. Both mechanisms help pupils to decide on the optimal level of effort. But in self-fulfilling prophecies teachers play an active role, giving signals to pupils; whereas if pupils receive relative performance feedback, then the teachers are passive actors, and pupils are basically inspired by possessing new background information about the relative position of their knowledge. Therefore there is a need to have a closer look at why grades could influence later achievement.

1.2. POSSIBLE MECHANISMS FOR WHY GRADES MIGHT INFLUENCE SCHOOL PERFORMANCE

Previous research (Betts and Grogger 2003) has argued that the higher grading standards of schools increase the achievement of pupils with the best performance. Weaker-performing students, on the other hand, may perceive themselves to be falling behind on a relative basis, even if their performance increases over time. Since pupils tend to perceive their performance in relative (rather than absolute) terms, higher grading standards could decrease the perceived probability of future educational success among those who have good performance in absolute terms, but only average performance in relative terms.

Research into self-concept also demonstrates that the achievement of peers could negatively influence the way in which pupils evaluate themselves. The big-fish–little-pond effect (Marsh and Hau 2003) suggests that, in terms of self-concept, a talented pupil will gain more if he is in a less-competitive classroom (little pond) than if he is put into a more competitive environment, with only big (or bigger) 'fishes'. In other words, while self-concept is influenced negatively by class average performance, a pupil's own performance maintains a positive effect (Marsh and Parker 1984; Marsh and Hau 2003; Marsh et al. 2008).

That said, it is plausible to assume that relative within-classroom differences in grades – or, put differently, in teachers' assessments – could influence subsequent performance by boosting self-assessment. If someone assesses his own knowledge positively, he might think it worthwhile to invest effort, since the probability of failure will be small. The self-worth theory (Covington 1984) suggests that the positive perception of ability activates school achievement. Self-efficacy is known to be influenced by positive feedback about performance

(Schunk 1985). Furthermore children's belief in their own ability and their expectation of success strongly influence their educational outcomes (Wigfield and Eccles 2000).

The economic literature also establishes the positive impact of self-confidence on achievement. Filippin and Paccagnella (2011) showed in their model that those who initially overestimate their abilities (a sign of self-confidence) will follow more ambitious educational paths and can accumulate more knowledge than those with low self-confidence, who underestimate their ability. Positive self-assessment (knowledge about own ability) translates into the choice of demanding education at the secondary level (Keller 2014), and perceived probability of success increases the decision to choose to go on to tertiary education (Keller and Neidhöfer 2014; Tolsma, Need, and de Jong 2010).

Self-assessment is clearly not the only way in which better grades could increase later achievement. It is argued, however, that other channels – such as a pupil's acceptance by peers (Wentzel and Caldwell 1997) or interest in the subject (Trautwein et al. 2006) – basically also have the side effect of increasing self-assessment. Therefore throughout the analysis this causal mechanism will be assumed to lie behind the impact of teachers' ratings.

An opposite interpretation would, however, be to assume that better grades do not motivate pupils to invest more effort, but rather having achieved a good grade they reduce their efforts, which could translate into negative subsequent school performance. Some kind of efficiency optimization is already established in the literature, but not in a classroom environment. It is known, for example, that researchers' scientific achievement declines once they achieve academic tenure; this is interpreted as indicating that if external sanctions are removed, and if people already have some kind of work security, their motivation to increase their performance is less pronounced (Holley 1977; Park and Gordon 1996). Other analyses show that employees who spend longer in a given job report boredom more frequently, which could hinder performance of the job (Ng and Feldman 2013) and could be explained by the loss in motivation and stimulation in a given work activity.

Deciding whether inflated grades have a positive or a negative effect on subsequent school performance is partly an empirical question. Therefore there is a need to undertake a more profound analysis and to investigate this question at the individual level (unlike in prior analysis). This is what will be done in this analysis.

1.3. THE PURPOSE OF THE ANALYSIS

Given the gap in prior research into the impact of grades on subsequent school achievement, this paper gives insight into whether grades do induce later progress in school performance. It will be argued that teachers' biased assessments translate into self-assessment, and positive knowledge of their own abilities increases the likelihood of pupils investing in the

effort of learning. Since effort is costly, pupils who have more positive knowledge of their own ability are assumed to be more likely to make this investment, simply because they are assured – in the form of their positive self-assessment – that the investment will not be in vain.

During the empirical analysis special attention will be devoted to dealing with the measurement error of test scores. This is necessary, since in a regression setting, where test scores are explained by prior test scores and grades, grades will show teachers' ratings only if a pupil's latent ability is controlled for entirely in prior test scores. Since test scores measure pupils' knowledge with noise, the impact of grades could be biased. This bias will be dealt with by choosing an instrumental variable approach, where grades are instrumented with school behavior, and by utilizing the diff-in-diff approach, which eliminates time-invariant ability from the estimation.

2. METHOD

2.1. DATA

Data are derived from the National Assessment of Basic Competencies (NABC), which is a Hungarian micro-level educational panel dataset that targets the full cohort of year 6, 8 and 10 students and measures their mathematical skills and reading literacy, using a PISA-like test. Pupils take a math and a reading comprehension test on the same day. The test takes four sessions of 45 minutes, and pupils have a 10 minute break after each session. The aim of the test is to assess how pupils are able to use the knowledge learnt at school in new situations, which have a practical focus. The test itself is written in the usual classroom and pupils are instructed by their teachers. However, the test questions are developed by the Hungarian Educational Authority, and the correction of the test is also organized by that authority.

The main analysis focuses on the change between year 6 and year 8, by three different cohorts of NABC. *Cohort 2008* completed year 6 in 2008 and year 8 in 2010; *cohort 2009* completed year 6 in 2009 and year 8 in 2011, while in the case of *cohort 2010* there are observations for 2010 and 2012.

2.2. SAMPLE

The sample in the analysis is restricted to those who changed neither school nor class within a school (referred as 'classrooms') between year 6 and year 8. If those who changed class were also included, it would not be possible to determine if the change in test scores was a by-

product of the change of class. Only classrooms with more than five persons are considered, in order to have a meaning behind the use of classroom fixed effects.

This restriction implies that the sample in the main analysis contains only pupils in elementary education.[*] In Hungary, pupils can move up to secondary school after the year 4, 6 or 8 of primary education. However, most pupils choose a secondary school after year 8, when elementary education ends and everybody who has not already done so must change school. Since early transition to secondary school correlates positively with social status and pupil ability (Horn 2013), restricting the sample to those who have remained in the same classroom environment means that pupils in the sample have a somewhat lower level of ability than the cohort as a whole.

On the other hand, the restriction also means limiting the sample to those who managed the two-year progression within two academic years. Pupils who dropped out, moved abroad or are simply missing from the year 8 data (c. 6.6% of all year 6 pupils) are not included in the sample. Since dropping out is the most likely explanation for why somebody is missing from the data in year 8, pupils in the sample have a higher school performance than the whole year 6 cohort.

As matters stand, even though in NABC there is no information on the teacher, it could be assumed that in the majority of cases it was the same person in years 6 and 8, since ability grouping within the same classroom is more likely in secondary schools (from year 9). Furthermore, schools usually employ different teachers in the first four years of elementary education, whereas in the majority of schools the same teacher teaches a subject throughout the second four years (unless the teacher retires or goes on parental leave).

During the empirical analysis many different approaches will be used to eliminate unmeasured ability. One of these approaches is the diff-in-diff method (to be discussed later), which leads to loss of the time sequence in the data, since it utilizes the differences between year 6 and year 8. Therefore, if the data allow, the year 10 outcomes are also analyzed. Analyzing year 10 outcomes is only possible for two cohorts (Cohort 2008 and Cohort 2009), since year 10 data were not available for Cohort 2010 at the time of our analysis.

2.3. MEASURES

2.3.1. Test scores

Pupils' math knowledge is partly measured by the NABC test. The test is written towards the end of the academic year, usually in May. Although test scores in reading comprehension are

[*] Everybody who chose the early secondary track is excluded, even those who did so after year 4 of elementary school.

also available, they are not used in this analysis, since the NABC test in reading comprehension requires skills and knowledge that differ from those taught in school in the Hungarian literature and grammar lessons (and for which pupils receive grades). For example, oral presentation is also part of the grade in Hungarian (learning a poem by heart), but the NABC contains only written tests. There is a greater similarity between the NABC math test and the tests conducted in school, simply because they are similar exercises, requiring a similar way of thinking. Even then, though, it could be that a particular pupil is good at algebra, but the NABC test contained mainly exercises in calculus, which this (hypothetical) pupil is not very good at. This scenario is thought, however, to be not very likely, since NABC basically measures applied knowledge and does not test knowledge in a particular domain of math. The test score in math is standardized with 0 mean and 1 unit standard deviation. Test scores are interpreted as a kind of blind measure about pupils' knowledge, since the person who corrects the tests does not know the pupils personally.

2.3.2. School grades

Pupils' math knowledge is also measured by grades awarded by the teacher – as a kind of non-blind measure, since teachers do know the pupils personally. Grades reported in pupils' mid-term report cards are used; these are basically the average of the grades received by pupils for tests set by the math teacher between September and January. The mid-term report card is issued in January, a few months before the test scores are measured.

In Hungary, school grades range from 1 to 5, where 1 is the worst and 5 is the best. If a pupil receives a grade of 1 at the end of the semester for three or more different subjects, he or she must repeat the whole year. If someone receives a grade of 1 in one or two subjects, he can take extra examinations to determine whether the year must be repeated. While a grade of 1 has more serious consequences at the end of the academic year, it is also a cause for concern if it appears on the earlier report card, since that could indicate the future outcome. Because being awarded the worst grade could contribute to a pupil dropping out, teachers sometimes avoid giving a grade 1 (simply avoiding the psychological consequences and the extra hassle).

2.3.3. School behavior

Pupils' school behavior is also reported in their mid-term report card. This is basically a grade ranging from 1 to 5, as with other subjects. The school behavior is evaluated by the form teacher, who is not necessarily the same as the math teacher. These grades are not part of pupils' grade-point average. They basically serve to provide information about school

behavior, but do not directly influence pupils' later educational career, or their evaluation in other subjects.

2.3.4. Other controls

There is no need to use many individual controls, since lagged math test scores capture individual factors that are stable over time and that influence pupils' performance. However, since test scores (and especially grades) might be connected to gender differences (Terrier 2014), a dummy variable showing whether the respondent is female will be included in every model.

2.3.5. Year 10 outcomes

There are three different types of outcomes in year 10. There are two measures of the educational tracks followed by someone in year 10. There is a binary variable, coded 1 if someone goes to a secondary school (from where there is direct access to tertiary education) and coded 0 if that person follows the vocational track (from where there is no direct entry to tertiary education). This is a kind of horizontal difference between the secondary tracks. Another dummy variable distinguishes between the two secondary tracks: it is coded 1 if a pupil follows the secondary general track (the academic track), and 0 if someone has been admitted to the secondary vocational track, which is basically a mixed version of the secondary and the vocational track, but offers a high-school final exam, which is necessary to enter tertiary education.

The math test score in year 10 also appears as a potential outcome (with 0 mean and 1 unit standard deviation). However, between year 8 and year 10 every pupil in the sample changes classroom, and even school – since they move from elementary education to secondary education – and so year 10 outcomes might also be influenced by the chosen secondary track (Hermann 2013). This should be considered and will be controlled for.

Lastly, the highest educational level that someone aspires to in year 10 is also analyzed. This question is used as a dummy variable, which is 1 if someone plans to go on to tertiary education and 0 otherwise.

2.3.6. Fixed effects

Classroom fixed effects are employed in every model. This basically defines the school class of year 6, but because the sample is restricted to those who did not change classroom between year 6 and year 8, it is the same as the year 8 classroom. Because every classroom has a

different ID, there is no need to control for cohort additionally. Analyzing year 10 outcomes, the combination of the year 6 classroom and the year 10 school will appear as fixed effects.

2.4. Identification

The aim of the analysis is to estimate the following equation (Eq. 1), where the math test score for the *i*th individual in the *j*th classroom in time *t* is explained with the same test score for the same individual in the same classroom in time *t-1*, controlling for the math grade that pupil received in time *t-1*. Using the lagged test score variable in the equation, every individual-level characteristic is controlled for which influenced pupils' prior math knowledge and which is assumed to be stable over time – such as ability, motivation, parental background, popularity with peers, or gender. Applying classroom-level fixed effects (π) is a parsimonious way of controlling for every factor that is shared by classmates and that has a clear effect on educational achievement – such as teachers, textbooks, class size, peer quality, etc.

$$score_{ij,t} = \alpha + \beta_1 \times grade_{ij,t-1} + \beta_2 \times score_{ij,t-1} + \pi_j + \varepsilon_{ij,t} \qquad \text{(Eq. 1)}$$

However, dynamic panel models (like Eq. 1) are known to underestimate the parameters, as was shown by Halaby (2004), since the unobservable time-invariant component of errors might also be correlated to the dependent variable and its lagged variant on the right-hand side. This feature of the model is accepted in the analysis and not treated further.

Eq. 1 takes advantage of having two different measures about pupils' knowledge in math: one is given by teachers (*grade*) and the second (*score*) is assessed by a standardized and centralized test (NABC), organized and developed by the Hungarian Educational Authority. Both measures could be regarded as a proxy for pupils' math knowledge, which is assumed not to be observable directly. Since *score* is measured by an independent institution, it is assumed not to be shaped by teachers' ratings, but these ratings are considered to be captured in *grades*, since they are assigned by teachers. Therefore grade contains pupils' math knowledge (Ω_g) and teachers' rating (Φ) as well. Hence it is assumed that:

$$grade = \Omega_g + \Phi + \nu_g \qquad \text{(Eq. 2)}$$

The focus of the analysis is on estimating teachers' rating (Φ) on the progress in individual test scores. Having another measure about pupils' knowledge (*score*) $\beta 2$ in Eq. 1 is thought to show the impact of the teacher's rating. This premise is realistic, since *grades* are given by teachers who know pupils personally; however, the NABC test is corrected by

independent scholars, who do not know whose test they are correcting, and hence *score* does not contain Φ.

However, the assumption that *grade* shows the teacher's rating holds only if *grade* and *score* measure the same kind of math knowledge (Ω). Even though there is a similarity between the NABC math test and the school-based math test, the assumption that *score* and *grade* measure the same knowledge is not very accurate, since math knowledge cannot be observed directly, and both measures are only a proxy. One might assume, therefore, that *score* measures pupils' math knowledge also with noise. This is expressed in Eq. 3:

$$score = \Omega_s + \nu_s \qquad \text{(Eq. 3)}$$

It will be further assumed that math knowledge captured by test scores (Ω_s) is correlated only imperfectly with math knowledge reflected in school grades (Ω_g). Therefore Ψ in Eq. 4 might contain ability captured in *score* that is not reflected in *grade*. Basically Ψ is a kind of ability which is not directly observable by teachers, but which is reflected in test scores. In practical terms it could be understood as the ability to write tests; or in other words, to feel no stress and to apply knowledge learnt to solve new problems. This ability might correlate with Φ, but the correlation should not be large, since Ψ is not observable by teachers. Controlling for *score,* however, will also control for Ψ.

$$\Omega_s = \Omega_g + \Psi \qquad \text{(Eq. 4)}$$

On the other hand, the measurement error in test score (ν_s) might be correlated to *grade*. Test scores are observed only once, but teachers award grades over a longer time period and they might be sensitive to abilities that are not reflected in test scores. It is plausible to assume that there are abilities observed by teachers but not measured in test scores – such as how pupils prepare their homework, how active they are in class, etc. It is assumed therefore that the covariance between *grade* and ν_s is not zero.

$$cov(\nu_s; grade) \neq 0 \qquad \text{(Eq. 5)}$$

Unfortunately neither latent knowledge in score (Ω_s) nor the measurement error itself (ν_s) is directly observable. Therefore there is no prior assumption about whether Φ and ν_s are correlated, and whether Φ could be interpreted as teachers' grading standard or this grading standard and some unobserved ability (rewarded by teachers). Furthermore if ν_s has a direct effect on *score*, even in the case of a nil correlation between Φ and later test scores, *grade* might show an effect in Eq. 1. This is especially harmful, because it might be consistent with

the interpretation of Φ. That said, in Eq. 1 one is not able to establish whether Φ has an effect, or whether the measurement error in score (v_s) biases the results.

This problem could be ruled out if all unobserved ability were controlled for. However, this is naturally not a realistic scenario. There are, however, some possible solutions that might address this issue; these will be discussed in the following paragraphs.

2.4.1. Instrumental variable approach

Prior research has found that the age at which a child begins school has a significant effect on later school achievement (Fredriksson and Öckert 2005; Massey, Elliott, and Ross 1996; Strøm 1995). Month of birth has already been used as an instrument for ability in settings similar to this study (Terrier 2014). As Puhani and Weber (2007) argue, age at school entry could influence later academic achievement because of maturity: pupils simply have better concentration and are better able to organize themselves. Even though month of birth is a good instrument for ability, it is not an appropriate instrument for teachers' grading standard.

On the other hand, literature about the teacher's pet phenomenon (Tal and Babad 1990) has shown that pets (based on students' nomination) are usually charming and compliant pupils with social skills, but are not necessarily the best-performing pupils. Applying the Ideal Pupil and Personality checklist developed by (Torrance 1963) in a survey situation, many empirical analyses (Fryer and Collings 1991; Kaltsounis and Higdon 1977) have found that teachers regard characteristics such as being a disturbing influence on the group, talking in class, being negative, or being unwilling to accept things as the features to be discouraged most in pupil behavior. Based on this argumentation, it could be assumed that pupils' school behavior might be a good proxy for teachers' grading standard, especially in terms of receiving inflated grades for a given level of performance. It is argued that the part of the math grade which correlates with school behavior shows the part of the grade which contains the teacher's rating, independently of latent ability. Using school behavior as an instrument, only that part of *grade* will be analyzed which correlates with it. Therefore all unobserved ability will be eliminated in this way. This is shown in Eq. 6 and Eq. 7.

$$\widehat{grade}_{ij,t-1} = \alpha + \beta_1 \times behaviour \quad \text{(Eq. 6)}$$

$$score_{ij,t} = \alpha + \beta_1 \times \widehat{grade}_{ij,t-1} + \beta_2 \times score_{ij,t-1} + \pi_j + \varepsilon_{ij,t} \quad \text{(Eq. 7)}$$

2.4.2. Diff-in-diff approach

An alternative way to handle this problem would be to eliminate unobserved ability both from *grade* and from *score*. This is possible, since there are two observations (year 6 and year 8) for each. The difference in *score* and *grade* is created using the formulas below:

$$D.grade_{ij} = grade_{ij,t} - grade_{ij,t-1} \quad \text{(Eq. 8)}$$
$$D.score_{ij} = score_{ij,t} - score_{ij,t-1} \quad \text{(Eq. 9)}$$

Hence the diff-in-diff approach is calculated using Eq. 10:

$$D.score_{ij} = \alpha + \beta_1 \times D.grade_{ij} + \beta_2 \times level_{ij,t-1} + \pi_j + \varepsilon_{ij} \quad \text{(Eq. 10)}$$

Even though calculating differences is powerful, there are two limitations here: first, the relationship will not be causal (partly because the time difference between grade and score was lost), and secondly because one part of the effect of teachers' grading standard will also be eliminated if it is considered time invariant.

2.4.3. Year 10 outcomes

In order to resolve one of the shortcomings of the diff-in-diff approach (that the relationship between *score* and *grade* is not causal) year 10 outcomes are analyzed. Using Eq. 11, four different dependent variables (Y) are analyzed: secondary school versus vocational education; secondary general versus secondary vocational school; math test scores; and educational plans. The difference in grade and score (between year 6 and year 8) appear on the right-hand side, controlling (in both cases) for the level in year 6 ($c_{ij,t-1}$). Classroom fixed effects (π_j) are also employed. In the models for year 10 test scores and educational plans, some additional control variables are used: math grade in year 10 and the secondary track that the pupil is following in year 10 ($c_{ij,10}$), since in these specifications every pupil changed classroom and even school, and the secondary track might have an influence on these outcomes (Hermann 2013). In these two later models, fixed effects are defined as the combination of year 6 classroom and year 10 school (π_j), since they are assumed to be shaped by the unobserved quality of secondary schools. Lastly, in the model fitted to year 10 educational plans, the year 8 educational plans also appear among the year 10 control ($c_{ij,10}$).

$$Y_{ij,t10} = \alpha + \beta_1 \times D.grade_{ij} + \beta_2 \times D.score_{ij} + \beta_3 \times c_{ij,t-1} + (\beta_4 \times c_{ij,10}) + \pi_j + \varepsilon_{ij,t10} \quad \text{(Eq. 11)}$$

2.5. THE CEILING EFFECT, THE FLOOR EFFECT AND THE TEACHER'S MERCY MECHANISM

Since pupils are not able to achieve a grade higher than 5, if somebody performs very well at time *t-1* and subsequently increases his performance, he will still only receive a grade of 5 at time *t*; thus he is more likely to be classified as under-rated than if he had received any other grade at time *t-1* (*ceiling effect*).

The opposite of this effect would be the *floor effect*, since pupils cannot have a worse grade than 1. This might work even in the opposite direction and lead to over-rated grades. However, a third mechanism could even be in operation. A pupil who receives the worst grade must repeat the school year, and this can have unpleasant consequences for teachers (more work, ethical qualms). In order to avoid any inconvenience, many teachers award a school grade of 2 even for very poor performance. Therefore many pupils who are spared by the teacher and receive a 2 are classified as over-rated. This *teacher's mercy mechanism* could also bias the estimated parameters since it means that the grade someone receives is in fact an already over-rated evaluation. Note that the ceiling and the teacher's mercy mechanism (or floor effect) could cancel one another out, since they work in different directions.

One possible way of finding out whether these mechanisms are in operation is to observe the pairwise difference between grades. Put differently: concentrating on only a one-unit change (comparing grade 1 with grade 2, grade 2 with grade 3, and so on), which also means gradually changing the baseline instead of fixing it at a particular grade level.

3. RESULTS

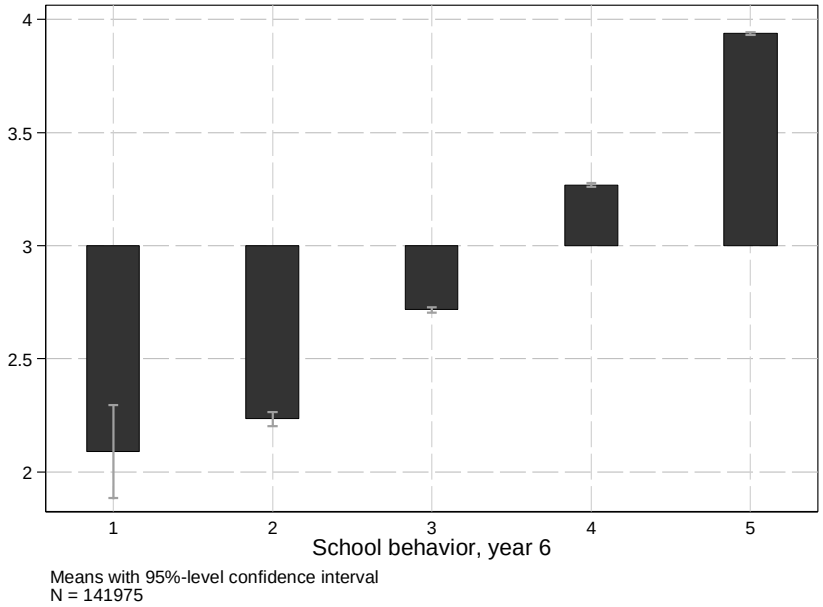
3.1. GRADES AND SCHOOL BEHAVIOR

The instrumental variable approach requires a profound analysis of the validity of school behavior, especially to find out whether it might capture teachers' biased grading standard. The argumentation for the instrument is that teachers might discriminate between pupils based on their school behavior. This assumption is plausible, in the sense that teachers may cooperate better with disciplined pupils (Fryer and Collings 1991; Kaltsounis and Higdon 1977). The raw connection between year 6 grades and year 6 school behavior (plotted on

Figure 1) seems to support this assumption. Pupils who are better at math are also those who are more disciplined. Even after controlling for prior math grade and gender (as Column 1 of Table 1 shows), those with better marks for school behavior receive better grades (assuming that grades are linearly distributed).

Figure 1.

The relationship between grade and school behavior, year 6



It is further assumed that school behavior correlates with test scores only through the math grade. This is a necessary assumption if school behavior is understood to contain no more latent ability than the math grade. As Column 2 of Table 1 shows, only outstanding school behavior has a positive effect on year 8 test scores ($b = 0.246$, $p < 0.001$). After the year 6 math grade is also controlled for (Column 3), school behavior does not have a direct effect on the year 8 test score: only the math grade does. It should be noted, however, that school behavior has a marginally significant and negative effect (in low categories) if the lagged math score is introduced as a continuous variable. This is important to remember, since the instrumental variable approach can proceed only on the assumption that the math grade is linear. However, since the effect is marginally significant and negative (which is contrary to the hypothesis that school behavior captures some latent ability), it is unlikely to bias the results later, and hence school behavior might be an appropriate instrument.

Table 1.

First step of instrumental variable approach

VARIABLES	(1) Year 6 grade	(2) Year 8 score	(3) Year 8 score	(4) Year 8 score
Year 6 grade (ref. Grade = 1)				
Grade 2			0.100**	
			(0.013)	
Grade 3			0.300**	
			(0.013)	
Grade 4			0.532**	
			(0.014)	
Grade 5			0.799**	
			(0.014)	
Year 6 school behavior (ref. Grade = 1)				
Grade 2	0.274*	-0.073	-0.101	-0.135+
	(0.103)	(0.059)	(0.058)	(0.058)
Grade 3	0.556**	-0.002	-0.076	-0.128+
	(0.102)	(0.058)	(0.057)	(0.057)
Grade 4	0.913**	0.109	-0.038	-0.097
	(0.102)	(0.058)	(0.057)	(0.057)
Grade 5	1.365**	0.246**	-0.007	-0.062
	(0.102)	(0.058)	(0.057)	(0.057)
Year 6 score (standardized)	0.648**	0.718**	0.568**	0.572**
	(0.003)	(0.002)	(0.003)	(0.003)
Female	0.008	-0.092**	-0.095**	-0.094**
	(0.005)	(0.004)	(0.004)	(0.004)
Year 6 grade (continuous)				0.226**
				(0.002)
Constant	2.423**	-0.109	-0.353**	-0.656**
	(0.102)	(0.058)	(0.057)	(0.057)
Observations	141,975	141,975	141,975	141,975
R-squared	0.482	0.581	0.618	0.617
Number of classrooms	10,644	10,644	10,644	10,644
Classroom FE	YES	YES	YES	YES
p	0	0	0	0
F	17924	17482	11990	16735

Robust standard errors in parentheses
** p<0.001, * p<0.01, + p<0.05

3.2. MAIN RESULTS

The main results are summarized in Table 2. Column 1 shows the results of the ordinary least squares (OLS) estimator while grades are introduced in discrete categories. The results show a positive effect; moreover, the size of the estimated parameter increases nearly linearly as the grade increases. Therefore in Column 2, using the same OLS estimator, the effect of grade is estimated if it is introduced as a continuous variable in the model. The estimated parameter is almost a quarter of the standard deviation of math test scores ($b = 0.235$, $p < 0.001$), showing that if someone receives a better grade for a given level of prior test score, he or she will experience progress in the test score which is approximately 25% of the standard deviation. This change is calculated as a one-unit change, which should be understood nominally – e.g. receiving grade 5 instead of grade 4.

However, this figure, estimated with OLS, is assumed to be biased by unobserved ability – or more precisely, by ability that can be observed by the teacher, but that is not reflected in the test score. Assuming that teachers' discriminatory grading standard is driven by pupils' school behavior, an instrumental variable approach is used. The results are plotted in Column 3. The estimated effect for grade is somewhat higher than the OLS results ($b = 0.301$, $p < 0.001$). One possible explanation is that abilities that may be observed by teachers, but not by the test contribute negatively to the gain in test scores. Those pupils who participate actively in math lessons and do their homework regularly are not necessarily those who will progress more. It could even be the reverse: however important and valuable these abilities are, they may compensate for a lack of understanding of the math curriculum.

Column 4 shows results of the diff-in-diff approach. The result is somewhat lower than the OLS estimation, but still positive ($b = 0.188$, $p < 0.001$). This strengthens the findings of prior models that being over-rated is associated with a positive test score. However, the result here is not causal: both grade and score are defined as the difference between year 6 and year 8 measures, hence the sequence in the data is also eradicated. The estimated lower figure obtained with this model could be explained by the fact that the time-invariant component of teachers' grading standards is eliminated from the difference. It is therefore plausible that as well as the change in teachers' grading standards, the time-invariant component might also correlate positively with test scores.

Table 2.

Main results, explaining year 8 test scores

	(1)	(2)	(3)	(4)
		Year 8 score		DIFF-score
	OLS	OLS	IV	DID
Year 6 grade (continuous)		0.235**	0.301**	
		(0.002)	(0.005)	
DIFF-grade				0.188**
				(0.002)
Year 6 grade (ref. Grade = 1)				
Grade 2	0.113**			0.216**
	(0.013)			(0.013)
Grade 3	0.324**			0.497**
	(0.013)			(0.013)
Grade 4	0.566**			0.784**
	(0.013)			(0.014)
Grade 5	0.840**			1.125**
	(0.014)			(0.015)
Year 6 score (standardized)	0.569**	0.572**	0.523**	-0.478**
	(0.003)	(0.003)	(0.004)	(0.003)
Female	-0.077**	-0.077**	-0.094**	-0.105**
	(0.003)	(0.003)	(0.003)	(0.003)
Missing grade (*DIFF-grade set to 0*)				-0.050**
				(0.005)
Constant	-0.420**	-0.782**	-1.005**	-0.592**
	(0.013)	(0.007)	(0.019)	(0.013)
Observations	141,975	141,975	141,975	141,975
R-squared overall model	0.598	0.598	0.593	0.226
Number of classrooms	10,644	10,644	10,644	10,644
Classroom FE	YES	YES	YES	YES
SE clustered	YES	YES	NO	YES
Lagged grade				YES
rho	0.403	0.402	0.402	0.413
p	0	0	0	0
F	19965	39033	.	4033
Wald chi2			197358	

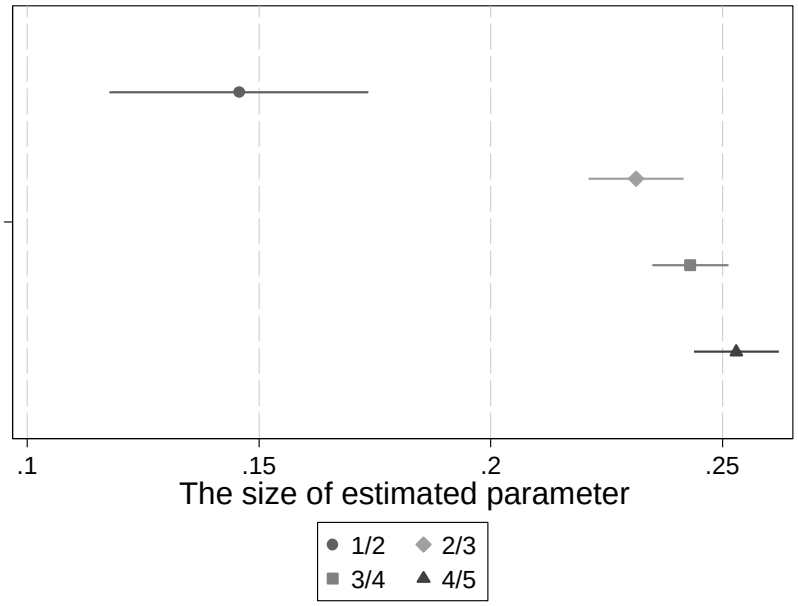
Robust standard errors in parentheses
** p<0.001, * p<0.01, + p<0.05

The results carried out with this analysis indicate that being over-rated in year 6 might increase pupils' later math test scores in year 8. The exact size might range between the upper and lower bound results provided by IV regression and the diff-in-diff approach. The results seem to be robust, even after focusing on pairwise differences between grades. As Figure 2 illustrates (full results are given in Table A1 in the Appendix) the unit-by-unit difference between the grades is similar, with the exception of the gap between grade 1 and grade 2. This means that those who received grade 2 instead of grade 1 benefited less from being over-rated than any other group that had a similar one-unit difference between (any other two) grades. This could be because pupils cannot have a worse grade than 1 or else

pupils are simply given a grade 2 as a sign of teacher's mercy. Both mechanisms translate to over-rated school performance, and hence the virtual difference between these two grades is smaller than the difference between any other two grades. Therefore pupils who receive grade 2 instead of grade 1 are less motivated to achieve more. On the other hand, the ceiling effect seems not to bias the estimations, because the effect of grades on those who receive grade 5 instead of grade 4 is not significantly different from the effect of someone receiving grade 4 instead of grade 3.

Figure 2.

The effect of lagged grade on subsequent test scores, pairwise differences between grades



3.3. SOME CONSEQUENCES IN YEAR 10 OF BEING OVER-RATED

In order to surmount one of the shortcomings of the diff-in-diff approach – lower bound estimation – other dependent variables are chosen. Using the difference in grade and score between year 6 and year 8, the year 10 outcome variables are explained. Even though these variables are partly categorical, an OLS estimator is employed, since – as previous literature has pointed out – calculating the marginal effects after logit models requires greater care (Buis 2010; Norton, Wang, and Ai 2004).

Table 3.

Explaining year 10 outcomes

VARIABLES	(1) Secondary / vocational year 10	(2) Sec. gen./ Sec. voc. year 10	(3) Math score, year 10	(4) Year 10 plan
DIFF-grade	0.080**	0.114**	0.075**	0.032**
	(0.002)	(0.003)	(0.005)	(0.003)
DIFF-score	0.029**	0.041**	0.384**	0.020**
	(0.002)	(0.004)	(0.006)	(0.004)
Year 10 math grade				
1 (worst)			Ref.	Ref.
2			0.046**	0.069**
			(0.013)	(0.009)
3			0.132**	0.147**
			(0.013)	(0.009)
4			0.223**	0.197**
			(0.013)	(0.010)
5 (best)			0.321**	0.209**
			(0.015)	(0.010)
Type of school (year 10)				
Secondary general			0.088**	0.149**
			(0.008)	(0.006)
Secondary vocational			Ref.	Ref.
Vocational			-0.280**	-0.139**
			(0.012)	(0.007)
Year 8 plan				0.368**
				(0.006)
Plan is missing				0.212**
				(0.008)
Female	0.044**	0.132**	-0.216**	0.018**
	(0.002)	(0.004)	(0.006)	(0.004)
Constant	0.277**	0.070+	-0.362**	0.075**
	(0.015)	(0.028)	(0.036)	(0.018)
Observations	72,966	62,928	69,547	69,937
R-squared overall model	0.225	0.256	0.534	0.449
Year 6 grade	YES	YES	YES	YES
Year 6 score	YES	YES	YES	YES
Number of classrooms	7,178	7,109		
Year 6 classroom FE	YES	YES		
N of year 6 classroom & sec.-school FE			28,312	28,414
Year 6 classroom & sec.-school FE			YES	YES
SE clustered	YES	YES	YES	YES
rho	0.225	0.256	0.534	0.449
p	0	0	0	0
F	1247	1300	4928	1595

Robust standard errors in parentheses
** p<0.001, * p<0.01, + p<0.05

The results are shown in Table 3. Being over-rated increases the likelihood that a pupil will choose more demanding education, irrespective of whether the horizontal (Column 1) or the vertical (Column 2) differences are considered between the secondary tracks. A one-unit increase in *grade* translates into an approximately 8% increase in the likelihood of a pupil embarking on the secondary school (rather than the vocational) track, and also increases the likelihood of a pupil following secondary general school track (instead of secondary vocational) by approximately 11%.

Receiving inflated grades also boosts the year 10 math performance. A one-unit better grade increases the math test score by 7.5% (Column 3), and increases the chances of a year 10 pupil considering tertiary education by approximately 3%. Note that these results are calculated after controlling for prior educational plans in year 8, the secondary education track that a pupil is following, and after applying year 10 school fixed effects (Column 4).

The results suggest that the grading standard used in elementary school might accompany pupils into secondary school, where their teachers and classmates are completely new, and where the elementary school grading standard should not have a direct effect. One possible mechanism by which the grading standard could influence later outcomes in year 10 is through self-assessment. As was argued in previous research, self-assessment might have an effect in educational transition – simply stimulating a pupil to dare to choose more demanding education (Keller 2014).

4. DISCUSSION

The empirical evidence of this paper supports the idea that receiving inflated school grades for a given level of performance in year 6 may have a positive effect on subsequent test scores in year 8. Even though, its exact size might not be measured without bias throughout the research, different approaches yield an equally positive impact, whether IV regression is used or the diff-in-diff model is employed. Furthermore, choosing year 10 outcomes and employing the difference in grades between year 6 and year 8 also reveals that grade maintains a positive influence on secondary track choice, secondary school performance, and further educational plans.

It is argued that teachers' assessment might increase self-assessed ability, which influences the effort that pupils invest in their own education (Azmat and Iriberri 2010). Since school performance is a combination of ability and effort, and since any investment in effort is costly, those pupils who have a more positive knowledge of their ability might be prepared to invest greater effort, since they are more certain that the investment is worthwhile. This reasoning is in line with previous findings, which show that if someone's

knowledge is confirmed by the teacher, this approval will boost the pupil's self-confidence, which could be translated into educational outcomes (Pajares and Schunk 2001).

Supporting this argumentation, the results show that being over-rated and getting a one-unit better grade in year 6 of elementary school increased math performance two years later, in year 8 (but still at the same school), by between 0.188 and 0.301 standard deviation, and year 10 math performance (in a different school and classroom environment) by 0.075 standard deviation. Since the standard deviation of math grade in year 6 is approximately one unit, and the standard deviation of DIFF-grade is approximately 0.7 of a unit, the results could be compared to the findings of Rivkin et al. (2005). They showed that a one standard deviation increase in teacher quality translates into an increase in math performance in the next year of 0.095 standard deviation. Hence, even considering that the results in this recent analysis are calculated for two years (not for one), the estimated effect of teacher's grading standard is somewhat larger than teacher quality. This might be on account of different methodology, but it could also be a consequence of other quality attributes of teachers which might decrease the impact of their grading standard.

Contrary to previous research, which argues that a teacher's assessment might act as a self-fulfilling prophecy (Rosenthal and Jacobson 1968) this paper argues that a teacher's grading standard might work through supportive feedback and encouragement of pupils, especially in relation to their academic achievement (Schunk 1983).

The findings presented here are new, in the sense that, unlike previous research (Betts and Grogger 2003; Terrier 2014), this analysis has been able to indicate the consequences of receiving inflated grades at the individual level (rather than the classroom or school level), while unobserved classroom-level heterogeneity is controlled for. Adding to prior scholarship, it is also established that being over-rated in elementary school may have consequences in secondary education as well. This is important, since elementary education and secondary education are institutionally separated in Hungary. For the sample analyzed in this paper it is shown that, even though pupils have different teachers and peers in secondary school, still a biased grading standard from their elementary education accompanies them to secondary school.

4.1. LIMITATIONS

There are some clear limitations to this analysis, and these invite a careful reading. First of all, because one of the aims was to minimize the possible bias caused by the measurement error in lagged test scores (unmeasured ability), different approaches were employed which assigned lower and upper bound estimations. But a more precise localization was not possible employing this scheme.

A further limitation is that grades and test scores might not measure the same concept. However, by restricting analysis to math test scores we basically seek to eliminate this bias; nevertheless, it could still be that latent ability biases the estimations.

Finally, throughout the analysis it was assumed that those pupils who do not change classroom over time were taught by the same teacher between year 6 and year 8. However, since there was no information about teachers in the survey, the analysis was not able to go beyond this assumption. On the other hand, it is likely that teachers' observable characteristics (especially their age and working experience) do have an influence on the grading standard that they employ.

4.2. CONCLUSION

One possible conclusion to be drawn from the results is that the grading standard used by teachers has a lasting – and in that sense *sticky* – effect for at least two reasons. First, it might influence pupils' self-assessment, which may determine the effort that pupils invest in education. Though this was not tested directly, the results showed that those who were over-rated in elementary school had more ambitious educational plans and followed more demanding secondary tracks. These findings could also indicate that a teacher's assessment in terms of grades might be influenced by a pupil's self-assessment, and especially by how ambitious that pupil is for further education.

Secondly, the grading standard used by teachers in elementary school also influences pupil outcomes in secondary school. Every pupil in the sample changed school and moved from elementary into secondary education. This means that the rating they received at elementary school accompanied them into the new school and classroom environment of secondary education.

Future analysis should cast further light on why teachers over-rate some pupils in the classroom. Previous analysis has shown that girls are more likely than boys to be teacher's pets, and that teacher's pets tend to have fairly good, but not excellent performance (Tal and Babad 1990). However, much more knowledge is needed about teachers' characteristics in order to understand their grading standards.

At this stage less is known about the possible negative effect of grades – especially whether pupils are spoilt if they receive inflated grades. Therefore the interpretation that the better the grades a pupil receives, the greater the benefits in terms of subsequent school performance is not necessarily supported by the results. Hopefully the results shown in this paper will contribute to greater understanding of how teachers employ grades, of what the long-run consequences are of receiving inflated grades, and especially of the heterogeneity in the effect of grades according to social background.

REFERENCES

Azmat, Ghazala and Nagore Iriberri. 2010. "The importance of relative performance feedback information: evidence from a natural experiment using high school students." *Journal of Public Economics* 94(7-8): 435–52.

Betts, Julian R. and Jeff Grogger. 2003. "The impact of grading standards on student achievement, educational attainment, and entry-level earnings." *Economics of Education Review* 22(4): 343–52. http://linkinghub.elsevier.com/retrieve/pii/S0272775702000596 (retrieved 3 November 2014).

Buis, Maarten L. 2010. "Stata Tip 87: Interpretation of interactions in non-linear models." *The Stata Journal* 10(2): 305–8.

Covington, Martin V. 1984. "The self-worth theory of achievement motivation: findings and implications." *The Elementary School Journal* 85(1): 4. http://www.journals.uchicago.edu/doi/abs/10.1086/461388

Filippin, Antonio and Marco Paccagnella. 2011. "Family background , self-confidence and economic outcomes." IZA Discussion Paper No. 6117: 1–25. http://ftp.iza.org/dp6117.pdf

Fredriksson, Peter and Björn Öckert. 2005. "Is early learning really more productive? The effect of school starting age on school and labor market performance." IZA Discussion Paper No. 1659. http://repec.iza.org/dp1659.pdf

Goulas, Sofoklis and Rigissa Megalokonomou. 2015. "Knowing who you are: the effect of feedback information on exam placement." Mimeo. Paper presented at the XXIV Meeting of the Economics of Education Association, Madrid, Spain, 25–26 June 2015. http://2015.economicsofeducation.com/user/pdfsesiones/011.pdf

Halaby, Charles N. 2004. "Panel models in sociological research: theory into practice." *Annual Review of Sociology* 30(1): 507–44.

Hermann, Zoltán. 2013. "Are you on the right track? The effect of educational tracks on student achievement in upper-secondary education in Hungary." Budapest Working Papers on the Labour Market BWP-2013/16. http://www.econ.core.hu/file/download/bwp/bwp1316.pdf

Holley, John W. 1977. "Tenure and research productivity." *Research in Higher Education* 6(2): 181–92. http://link.springer.com/10.1007/BF00991419

Horn, Dániel. 2013. "Diverging performances: the detrimental effects of early educational selection on equality of opportunity in Hungary." *Research in Social Stratification and Mobility* 32: 25–43.

Jussim, Lee and Jacquelynne S. Eccles. 1992. "Teacher expectations: II. Construction and reflection of student achievement." *Journal of Personality and Social Psychology* 63(6): 947–61.

Jussim, Lee and Kent D. Harber. 2005. "Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies." *Personality and Social Psychology Review* 9(2): 131–55.

Jussim, Lee, Jacquelynne Eccles, and Stephanie Madon. 1996. "Social perception, social stereotypes, and teacher expectations: accuracy and the quest for the powerful self-fulfilling prophecy." *Advances in Experimental Social Psychology* 28(C): 281–388.

Keller, Tamás. 2014. “Talented but unaware? An analysis of the role of self-assessment in educational transition.” Budapest Working Papers on the Labour Market BWP-2014/9. http://www.econ.core.hu/file/download/bwp/bwp1409.pdf

Keller, Tamás and Guido Neidhöfer. 2014. “Who dares, wins? A sibling analysis of tertiary education transition in Germany.” SOEPpapers on Multidisciplinary Panel Data Research No. 713. http://www.diw.de/documents/publikationen/73/diw_01.c.492428.de/diw_sp0713.pdf

Madon, S., L. Jussim, and J. Eccles. 1997. “In search of the powerful self-fulfilling prophecy.” *Journal of Personality and Social Psychology* 72(4): 791–809.

Marsh, Herbert W. and Kit-Tai Hau. 2003. “Big-fish–little-pond effect on academic self-concept: a cross-cultural (26-country) test of the negative effects of academically selective schools.” *American Psychologist* 58(5): 364–76. http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.58.5.364 (retrieved 23 May 2013).

Marsh, Herbert W. and John W. Parker. 1984. “Determinants of student self-concept: is it better to be a relatively large fish in a small pond even if you don’t learn to swim as well?” *Journal of Personality and Social Psychology* 47(1): 213–31.

Marsh, Herbert W. et al. 2008. “The big-fish–little-pond-effect stands up to critical scrutiny: implications for theory, methodology, and future research.” *Educational Psychology Review* 20(3): 319–50. http://link.springer.com/10.1007/s10648-008-9075-6 (retrieved 25 May 2013).

Massey, Alf, Gill Elliott, and Emma Ross. 1996. “Season of birth, sex and success in GCSE English, mathematics and science: some long-lasting effects from the early years?” *Research Papers in Education* 11(2): 129–50.

McMillan, James H., Steve Myran, and Daryl Workman. 2002. “Elementary teachers’ classroom assessment and grading practices.” *The Journal of Educational Research* 95(4): 203–13.

Ng, Thomas W.H. and Daniel C. Feldman. 2013. “Does longer job tenure help or hinder job performance?” *Journal of Vocational Behavior* 83(3): 305–14. http://linkinghub.elsevier.com/retrieve/pii/S0001879113001395 (retrieved 17 November 2014).

Norton, Edward C., Hua Wang, and Chunrong Ai. 2004. “Computing interaction effects and standard errors in Logit and Probit models.” *The Stata Journal* 4(2): 154–67.

Pajares, Frank and Dale H. Schunk. 2001. “Self-beliefs and school success: self-efficacy, self-concept, and school achievement.” In R. Riding and S. Rayner (eds), *Perception*. London: Ablex Publishing.

Park, Seung Ho and Michael E. Gordon. 1996. “Publication records and tenure decisions in the field of strategic management.” *Strategic Management Journal* 17(2): 109–28. http://doi.wiley.com/10.1002/smj.4250170201 (retrieved 15 January 2015).

Puhani, Patrick A. and Andrea M. Weber. 2007. “Does the early bird catch the worm?” *Empirical Economics* 32(2-3): 359–86.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, schools, and academic achievement.” *Econometrica* 73(2): 417–58 http://www.nber.org/papers/w6691.pdf (retrieved 10 September 2014).

Rockoff, Jonah E. 2004. “The impact of individual teachers on student achievement: evidence from panel data.” *The American Economic Review* 94(2): 246–52.

Rosenthal, Robert and Lenore Jacobson. 1968. “Pygmalion in the classroom.” *The Urban Review* 3(1): 16–20.

Schunk, Dale H. 1983. “Ability versus effort attributional feedback: differential effects on self-efficacy and achievement.” *Journal of Educational Psychology* 75(6): 848–56.

Schunk, Dale H. 1985. “Self-efficacy and classroom learning.” *Psychology in the Schools* 22(2): 208–23.

Snow, Richard E. 1969. “Unfinished Pygmalion.” *Contemporary Psychology* 14(4).

Strøm, Bjarne. 1995. *Student Achievement and Birthday Effects*. Trondheim: Norwegian University of Science and Technology.

Tal, Zohar and Elisha Babad. 1990. “The teacher’s pet phenomenon: rate of occurrence, correlates, and psychological costs.” *Journal of Educational Psychology* 82(4): 637–45. http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-0663.82.4.637

Terrier, Camille. 2014. “Giving a little help to girls? Evidence on grade discrimination and its effect on students' achievement.” PSE Working Papers 2014-36. https://hal-pjse.archives-ouvertes.fr/hal-01080834/document

Tolsma, Jochem, Ariana Need, and Uulkje de Jong. 2010. “Explaining participation differentials in Dutch higher education: the impact of subjective success probabilities on level choice and field choice.” *European Sociological Review* 26(2): 235–52. http://esr.oxfordjournals.org/cgi/doi/10.1093/esr/jcp061 (retrieved 31 May 2014).

Trautwein, Ulrich, Oliver Lüdtke, Herbert W. Marsh, Olaf Köller, and Jürgen Baumert. 2006. “Tracking, grading, and student motivation: using group composition and status to predict self-concept and interest in ninth-grade mathematics.” *Journal of Educational Psychology* 98(4): 788–806.

Wentzel, Kathryn R. and Kathryn Caldwell. 1997. “Friendships, peer acceptance, and group membership: relations to academic achievement in middle school.” *Child Development* 68(6): 1198–209. http://doi.wiley.com/10.1111/j.1467-8624.1997.tb01994.x (retrieved 11 November 2014).

Wigfield, Allan and Jacquelynne S. Eccles. 2000. “Expectancy-value theory of achievement motivation.” *Contemporary Educational Psychology* 25(1): 68–81. http://www.ncbi.nlm.nih.gov/pubmed/10620382 (retrieved 24 May 2013).

Wright, S. Paul, Sandra P. Horn, and William L. Sanders. 1997. “Teacher and classroom context effects on student achievement: implications for teacher evaluation.” *Journal of Personnel Evaluation in Education* 11: 57–67. http://www.springerlink.com/content/l7q2242qnj2125w0/

APPENDIX

Table A1.

The effect of lagged grade on subsequent test scores, pairwise differences between grades

	(1) Grade 1 /Grade2	(2) Grade 2 /Grade3	(3) Grade 4 /Grade5	(4) Grade 4 /Grade5
Year 6 grade (continuous)	0.146**	0.231**	0.243**	0.253**
	(0.014)	(0.005)	(0.004)	(0.005)
Year 6 score (standardized)	0.476**	0.529**	0.567**	0.598**
	(0.007)	(0.004)	(0.003)	(0.003)
Female	-0.085**	-0.079**	-0.076**	-0.073**
	(0.009)	(0.005)	(0.004)	(0.004)
Constant	-0.697**	-0.812**	-0.819**	-0.863**
	(0.028)	(0.014)	(0.015)	(0.020)
Observations	27,289	66,053	86,377	73,267
R-squared overall model	0.287	0.386	0.455	0.516
Number of classrooms	10,644	10,644	10,644	10,644
Classroom FE	YES	YES	YES	YES
SE clustered	YES	YES	YES	YES
rho	0.551	0.479	0.446	0.437
p	0	0	0	0
F	1620	8204	15071	17132

Robust standard errors in parentheses
** p<0.001, * p<0.01, + p<0.05