An excellent label of acknowledgement but a disappointing label of excellence

Master Thesis submitted in partial fulfilment of the requirements for the degree of MSc Policy Economics at the Erasmus School of Economics

Submitted by:	Mauro Stel
Supervisor:	Aart Gerritsen
Second reader:	Dinand Webbink
Date of submission	05-08-2018
Master:	Economics & Business

Abstract

In 2010, the Dutch government proposed a range of policy measures to stop the observed decline in verifiable performance of both primary and secondary students in internationally comparable tests and final exam grades. One of the introduced policy instruments was a label of excellence for high-performing schools. The label can be regarded as a non-financial acknowledgement and reward of effort and a signal of quality. Whereas initially schools were mainly selected on the basis of their objectively verifiable performance on school results, currently a school can become excellent if it distinguishes itself through its unique didactic approach. I have analysed the (un-)intentional effects of the policy on exam scores and student inflow amongst excellent schools, in order to determine to what extent the policy has affected the objective performance of students and intervenes with the school choice of future students. Moreover, the paper in its entirety addresses the concern that the label of excellence has increased inter-school and socio-economic inequality. For the latter claim, no evidence has been found. The results from the fixed-effects estimation indicate a negative causal relationship between the label of excellence and exam scores in both primary and secondary schools. Students in secondary school score 0.0615 points less (-0.24SDs) on the final nationwide exams. This significant finding is highly robust to different model specifications. Students in primary school score almost 1 point less (-0.22SDs) on the Cito test in the year the school has become excellent, although this coefficient is less consistently estimated. The decline in test scores is attributable to the bureaucratic hassle of obtaining the label of excellence. Furthermore, there is tentative evidence that student inflow goes up (with a lag) in the wake of becoming excellent. More research is needed on the persistency of the effect of the label of excellence on student outcomes and the impact of the label on the performance of students in non-excellent schools, however, it seems plausible to assume that the label of excellence has failed to boost the average performance of students in the Netherlands.

Keywords: excellence in education, test scores, school quality, school accountability, student performance, sorting, inter-school competition, inequality of opportunity, educational policy

Section 1: Enhancing (verifiable) school quality

1.1. Introduction

Shortly after the instalment of the Rutte-I government in 2010, the incumbent Minister of Education, Culture and Science, Marja van Bijsterveldt, proposed a wide range of policy measures – in its entirety referred to as a plan of action to convey the urgency of the message – which the ruling Coalition was planning on implementing in order to enhance the quality of both Dutch primary and secondary education. The Minister had noticed that despite the Netherlands' consistently high-ranking position (10th in reading, 11th in math and science in 2009) in international leaderboards of educational accomplishment, performance was gradually stagnating. Reading, math and science scores on PISA tests[1] declined relatively to other countries as well as that final exam grades were steadily decreasing too across all subjects over the course of 5 years (from a 6.4 to a 6.3 on average) (Inspectorate of Education, 2011). So as to turn the tide, Dutch education was to focus more on the 'excellent' rather than 'weak' student. For too long a period of time, it was a school's core business to merely bring all students up to a certain benchmark level of knowledge and capabilities. Some students, however, possess the cognitive qualities to reach far beyond the established benchmark. It was the Minister's wish that these exceptionally talented, highly able students would receive the attention and means necessary to thrive and live up to their talents (Mooij & Fettelaar, 2010).

A rationalization of education through close monitoring of student performance was vital in achieving these newly set educational goals. Schools were to track their students' achievements more extensively, in order to provide an ambition-sensitive, student-tailored and thoroughly individualised didactic service. At the same time, budgetary cuts were still deemed necessary in the post-crisis years, which compelled the Minister to be resourceful and efficient in spending its (financial) means. Within this context, the plan to confer schools with a 'Label of Excellence' first emerged.

Previously, high-performing schools were not rewarded for their efforts. As long as you provided a baseline level of quality, the Inspectorate of Education would keep their levels of scrutiny to a minimum, allowing schools a fair degree of freedom to operate. An incentive to excel beyond the set baseline targets by the Inspectorate was therefore practically non-existent. The Minister reasoned that in order to *"maximize the results of students"*, acknowledgement of a school's outstanding achievements was warranted (Ministry of ECS, 2011a). This led to the conception of a new, honorary title: 'Excellent'. From 2012 onwards, both primary and secondary schools could apply for the 'Label of Excellence'. Since the introduction of the reward, approximately 750 labels have been handed out to (non-special)[2] primary and secondary schools. In January 2017, excellent schools comprised a little over 3% of all (non-special) schools. As the label suggests, excellent schools distinguish themselves (in many ways) from non-excellent schools. For instance, both students in excellent primary and secondary schools score much higher on nationwide final exams, despite their student population consisting of a disproportionate number of students from disadvantaged and migratory backgrounds or lower socio-economic status (Regioplan, 2016). Furthermore, on many other indicators of verifiable student performance excellent schools also outperform non-excellent schools (see Table 3 & Appendix VII).

This is, of course, rather unsurprising, because the school's verifiable achievement in maximizing student performance is one of the main criteria upon which assessment of a school's

[1] In 2000, the international average at the PISA tests was set at a score of 500 for reading skills. Dutch students scored 513 in 2003, but dropped back to 508 in 2009. In 2003, the international average for mathematics was set at 500. Dutch students consistently perform above average on math, but the skill advantage has decreased slightly. In 2003, Dutch students scored 538, whereas in 2009 it was only 526. A convergence of test scores between the Netherlands and the rest of the world is also observed in science (Cito, 2010).

[2] Special schools are non-standard schools for children with cognitive disorders or mental and physical disabilities.

alleged excellence is based. However, it is definitely not the sole criterion. In particular post-'Excellent Schools 2014', an applicant school's 'profile of excellence' – *an inspiring, innovative or motivating curriculum or a distinctive approach to education for a specific group of students* – has gained in importance in the judges' determination of a school's eligibility for excellence. Whereas before 2015 the Ministry was responsible for assigning labels of excellence to applicant schools, the Inspectorate of Education took over in 2015. This body, responsible for the supervision of quality in Dutch education, explicitly lists a school's profile of excellence as an essential prerequisite for the label's receipt. Even though pre-2015 Dronkers (2014) already determined that a large fraction of excellent schools scored 'merely' above average in terms of verifiable performance criteria (and barely occupied the top of the national rankings), the current headline role for a school's profile of excellence has rendered the selection procedure seemingly opaque and arbitrary. Moreover, **(1)** the sustained ambiguity regarding the policy's *actual* objectives, **(2)** the confusing application of the term 'excellent' and **(3)** the widespread critique (on potential side-effects of the label) have inspired me to turn 'Excellent Schools' into the subject of this master thesis and dig deeper into the (un-)intentional consequences of the 'Label of Excellence'. In the ensuing paragraphs, I will elaborate upon these three claims in numerical order.

(1) The policy's overarching goals are vague and inconsistent. Since the label of excellence was explicitly introduced as part of a package of measures to retake the top position in the international standings on educational achievement, it could be suspected that the policy is as an attempt to indirectly induce competition between schools. A school can signal its high quality towards its future students if it bears a label of excellence, which may subsequently convince them to enrol into the school. Since a school's funding is dependent on the total number of students, the label of excellence can be an indirect way through which more funds (and prestige) can be acquired. If many schools would strive to obtain the label of excellence, their efforts to qualify for the label are likely to result in better student outcomes. Moreover, excellent schools will potentially receive more students and therefore thrive at the expense of weak, non-excellent schools which may gradually disappear. Although slightly magnified, one can clearly discern an enhanced market mechanism here. With the creation of the label of excellence, schools can suddenly distinguish themselves on more than plain statistical indicators of educational quality[3].

It is debatable whether this is a favourable development. Education is a public good in the Netherlands, in order to ensure that every individual born in this country has equal access to opportunity. Nevertheless, in 2017 the Inspectorate of Education warned that inter-school differences in quality are increasing (Inspectorate of Education, 2017a). This implies that conditional on the school of choice of two identical children, their learning outcomes should be alike. However, equally talented students will wind up at different levels of secondary education depending on their preferred school of choice. This wastes a lot of talent. Furthermore, one year prior, the Inspectorate already signalled that gradually the inequality of opportunity is increasing along socio-economic boundaries (Inspectorate of Education, 2016a). One of the main causes is that higher-educated parents tend to be better informed about a school's relative level of quality compared to other schools and henceforth send their children to *on average* better schools. More recently, in early 2018, the Inspectorate reiterated its concern by referring to Dutch education as increasingly being segregated, thereby inflating *"bubbles of likeminded people"* which are notoriously difficult to escape from (Inspectorate of Education, 2018a). All these simultaneous developments undermine the equality of opportunity. If, in addition to those developments, schools will increasingly be branded as excellent (or good) and weak (or bad), it is likely that these alarming trends will only be reinforced. Therefore, I view the policy's goals as paradoxical.

[3] Needless to say, schools can also distinguish themselves in many ways unrelated to school quality, such as the school building, learning facilities, student exchanges and much more, but these factors are not directly indicative of student performance.

(2) Although the Inspectorate has (either deliberately or inadvertently) tried to steer away from the incentivizing market forces through adherence to a broader definition of quality and more attention for a school's profile of excellence (after they took over from the Ministry in 2015), it has, in its execution, misapplied the term 'Excellent'. This would not be a big deal if the public was aware of the misnomer. However, parents do not know that an excellent school will not inevitably maximize their child's educational performance. In fact, the increased importance of the profile of excellence implies that practically any school which excels in the provision of a certain course, the facilitation of an environment for children with issues to thrive in or the fulfilment of civic duties can be eligible for the label, as long as the baseline level of quality is also 'good'[4]. In 2016, Sander Dekker, State Secretary of Education, Culture and Science, informed the Dutch parliament that 'excellence' is explicitly *not* a judgement of quality (Sikkes, 2016). However, in the public perception this is not understood well, because 'excellence' is regarded as a superlative of 'good'. According to Paul Rosenmöller, chairman of the Dutch Council for Secondary Education, the label of excellence is therefore unnecessarily confusing and suggests that an excellent school provides better education than a non-excellent school, which is not always reflective of the truth. As a matter of fact, he has on multiple occasions plead for a withdrawal of the label of excellence (van Walsum, 2018). Although the upside of a broader definition of excellence is that the polarizing effects of the label are presumably less significant, the sustained misinterpretation of the public of what excellence entails implies that the *inflow*-effects of the label will nevertheless be likely to persist.

(3) Arguably most importantly, a successful educational policy should ultimately be to the benefit of the children and students in both primary and secondary schools. Any long-term effect is expected to manifest itself through inter-school competition, which I previously dubbed the market mechanism. However, testing the hypothesis that increased school competition has raised student performance is virtually impossible. Not only because there have only been five waves of label conferral ceremonies (from 2013 to 2017), which implies the long-term consequences have yet to fully pan out, but also due to the presence of a myriad of confounding variables which affect nationwide student performance, which will make strict identification of the *competition*-effect particularly complicated. Therefore, I resort to the identification of short-term effects, (more) directly related to the policy and identifiable at the (excellent) school-level. Many performance-related variables are likely to be impacted by the receipt or mere application for the title of 'Excellent School'. Some of these alleged effects have also been invoked to criticize the label of excellence (Regioplan, 2016; Bouma, 2018).

The first and foremost is the impact on verifiable and objective student performance in excellent schools; a topic upon which no research has been carried out up until today. Although I suspect that in the long-term there will be *compositional* effects on final nationwide exam grades (in secondary education) and Cito scores (in primary education), in the short-term these changes in the student population are unlikely to be reflected in exam scores (and as such reinforce inter-school inequality). However, students in excellent schools could still experience grade changes due to *efficiency* or *motivational* effects from the receipt of the label. Throughout the procedure to obtain the label, schools receive lots of feedback on their functioning. Hence, weaknesses can be addressed, which is likely to boost student performance. Furthermore, the collective effort to obtain a coveted acknowledgement for one's achievements can subsequently contribute to "*a 'culture of education'[5] in which all stakeholders naturally strive for improvement and share their expertise"*. If all stakeholders are more motivated, this may have positive implications for student outcomes. Furthermore, some excellent schools also report that the approach towards students is more 'differentiated' or ambition-sensitive (Regioplan, 2016). On the one hand, these developments seem to result in higher grades

[4] With 'good', a sufficient baseline level of quality is implied. From 2017 onwards, 'good' has become a separate quality category, although formally it is a mere 'acknowledgement' of quality. In the *data*-section, more information can be found about the Inspectorate's quality classification scheme.
[5] In Dutch: *onderwijscultuur*

post-receipt of the label of excellence. On the other hand, schools also report that the amount of paperwork required to fill in in order to obtain the label of excellence is a heavy burden; a lot of time and energy is devoted to the trajectory to become excellent which cannot be spend on something else (i.e. opportunity costs). Furthermore, the expectations from parents and other external parties rise in the wake of the receipt of the label of excellence. This is accompanied by a fair amount of pressure, which can have adverse effects on performance. Additionally, post-receipt of the label, schools may slack off because they lose the urge to sustain an 'excellent' level of performance as well as that the public acknowledgement of excellence could crowd-out the intrinsic motivation of staff members. Moreover, ever since the profile of excellence has moved to the forefront of the assessment in 2015, schools have been perversely incentivized to dedicate a disproportionate amount of time on the development of their particular profile. However, high-quality music education, innovative civic classes or thorough attention for (a small subset) of highly able students might also divert the didactic focus from its core business. To determine which effect prevails, I raise the following research question:

In what way does the 'Label of Excellence' as conferred to primary and secondary schools by the Dutch Inspectorate of Education affect objective student performance?

I have panel data on the final exam scores and excellence status[6] for all non-special primary and secondary schools from respectively 2012 and 2008 up to and including 2017. In total, this amounts to 7057 panels of primary schools consisting of 37.602 observations and 3528 panels of secondary schools consisting of 28.805 observations. I have restricted myself to fitting fixed effects models to the data. This implies that only the causal effect of time-varying variables on exam scores could be estimated. As a matter of fact, the persistent effect of the label of excellence on the exam scores of consecutive cohorts could not be determined. A staggered Dif-in-Dif analysis would have resolved this, but could not be executed because (1) treated schools have kept the label for varying lengths and (2) the available data on covariates is too limited (in particular for primary schools) to construct suitable control groups through propensity score matching. Nevertheless, the identification strategy remains useful, as the main efficiency and/or motivational effects of the label of excellence on exam scores are to be expected shortly after conferral anyway. In addition to panel and year fixed effects, I control for a set of covariates plausibly correlated with both 'becoming excellent' and the dependent variable.

Moreover, I check for both leading and lagging effects of excellence. As indicated before, the application and selection procedure for excellence spans at least a year. In the time leading up to the conferral ceremony, the school is continuously trying to improve itself in order to obtain the label. Potentially already before it has filed an application. Therefore, leading effects have been taken into account. Similarly, it is very well possible that a school peaks in performance at the point of conferral, but slacks off in the months and years after[7]. If an effect exists, I will have captured it through the inclusion of a lagged independent variable. Although grade changes at the level of the school may seem small, for individual students minor variations can make the difference between a pass or a fail (and a mandatory resit or additional year in education).

As a complement to the research question above, I have also taken a close look at the change in the influx of new students towards schools that have recently become excellent. Even though (long-run) *compositional* changes cannot be studied, increased enrolment in the short-run could indicate that sorting takes place. From a revealed preferences perspective, this would imply that sole receipt of the label of excellence raises the school's value relative to other schools as perceived by the child's parents. Even in the absence of an immediate effect of the label on objective student performance, a bigger influx of students could indicate that the parents expect the label to foreshadow better performance in the long-run (but possibly on unverifiable aspects of education) as well as that inter-

[6] This is equivalent to the treatment status, which may span multiple years.

[7] This is especially relevant post-'Excellent Schools 2014', because from then on an obtained label was valid for at least 3 years.

school quality differences will worsen. However, the trouble with inflow as a performance-related indicator of quality is that if parents are imperfectly rational (and therefore uncritical), the label of excellence could also simply be a 'nudging device' that entices parents into picking an excellent school in the absence of any *real* effects. The latter is quite plausible, as long as the term 'excellence' remains a misnomer.

The model I estimate is a reduced-form model. The channels and pathways through which the label of excellence affects objective student outcomes is therefore unclear. However, through the inclusion in the main specification of a dummy of formerly excellent schools which failed to retain their label of excellence, a little more can be said about the underlying mechanisms. Furthermore, as robustness checks to the main fixed effect estimations, I have conducted first difference analyses as well as included the decentralized 'school exams' as a dependent variable to check for spurious correlation and/or manipulation of test results.

Lastly, I have noticed that a disproportionate number of excellent schools are situated in the most-populated provinces (Regioplan, 2016). On the basis of this finding, I hypothesize that the projected gains from the label of excellence are largest in highly concentrated regions, where schools need to compete more feverishly for the future student's favour. Since the incentive for schools to participate in densely populated regions is stronger, it is a logical corollary that this equally applies to the amount of effort and time schools put in in order to obtain the label of excellence. Moreover, I also hypothesize that the Inspectorate would want to keep the observed regional distribution of excellent schools approximately in line with the expected distribution of excellent schools based upon the number of students in each province. Although I emphasize that this is a supposition, I would not be surprised if in practice the Inspectorate is slightly more lenient in its provision of the label of excellence to an applicant school from an underrepresented region (e.g. Groningen) than to a school from an overrepresented region (e.g. the Randstad). Therefore, if the label of excellence is associated with both student exam performance and inflow, I would expect the treatment to interact with the degree of competition a school faces within 3 and 4-digit ZIP-code areas. This hypothesis is tested through reapplying the main model to a slimmed down sample of schools which share a ZIP address with an excellent school as well as the inclusion of an interaction term which denotes the status of excellence of a particular school and the degree of competition it encounters (i.e. either low, medium or high).

The results indicate a negative effect of the label of excellence on exam scores for both primary and secondary schools. Conditional on exam and cohort fixed effects as well as a set of time-varying controls, students in secondary schools which have become excellent score 0.0615 points less on the final nationwide exams, which is equivalent to -0.24SDs. Applying the same model to primary education (with a different set of controls) yields similar results. Students in primary school score almost 1 point less on the Cito test in the year that their school has become excellent, which equates to -0.22SDs. These findings are highly robust to different model specifications and sample configurations, including taking first differences. However, it must be noted that the inclusion of a 1-year lagged dependent variable to control for mean reversion did render the (reduced) excellence-coefficient insignificant for primary schools at the 15%-level.

The label of excellence does not significantly interact with the level of education in secondary schools. Nor is the size and direction of the change in exam scores affected by the inclusion of leads and lags of the independent variable. Whereas a modest 1-year positive lagged effect of the policy is observed amongst secondary schools, a much stronger 1-year positive leading effect is observed amongst primary schools. Although this suggests that part of the direct, negative effect of the policy is offset, the shrunk sample size, high sensitivity of the estimates to varying combinations of leads and lags and other sample-related issues make it hard to draw definite conclusions. Nevertheless, in all SE-models the label retains its adverse effect on exam scores, even if one adjusts the estimates for region-specific unobservables correlated with the treatment. Although the overall effect is then reduced, it remains well below zero (-0.031***). No regional trend was observed amongst primary schools and

with the exception of the inclusion of the LDV all PE-models also yield negative point estimates of the effect of the label of excellence on exam scores. Lastly, there is no irrefutable evidence that the relationship between the label of excellence and exam scores is moderated by the degree of competition an excellent school faces.

The relationship between annual inflow into the school and the label of excellence is less well-established. Receipt of the label of excellence by a secondary school for a particular level of education is associated with a significant but modest increase in inflow of 6.5 students (0.06SDs) the following year and a school expansion of 10 students (0.07SDs). Furthermore, the effect is moderated by the level of education: more difficult educational levels tend to benefit more from the label of excellence. However, no effect was found for primary schools for any of the two inflow variables, although tentative evidence suggests that student inflow is more strongly related to 1 or 2-year lags of excellence. Furthermore, the first-differenced coefficients are never negative, which is contrary to the trend amongst non-excellent schools. Confounding regional trends in inflow were not observed for both primary and secondary schools.

1.2. Related literature

According to the Inspectorate, *"making quality visible"* is one of the core objectives of assigning labels of excellence to schools. As said before, one can regard this as an attempt to stimulate inter-school competition. When viewed in that light, my research is connected to studies by Dijkgraaf et al. (2013) and Noailly et al. (2012) on Dutch school competition. Using secondary school site data from 2002 to 2006, Dijkgraaf et al. (2013) find a negative relation between competition and educational outcomes as measured by nationwide exam grades, graduation rates and the percentage of students which obtain their diploma within the designated time. Although the effect is small and insignificant at times, it is never in any way positive. The authors' preferred explanation for finding a negative relation between competition and school quality is that schools compete fiercely on many other aspects of education, which consumes both time and money that cannot be spent on enhancing objectively verifiable school quality. In a similar vein, the opportunity costs of upholding a profile of excellence could have adverse implications for nationwide exam scores in excellent schools. Noailly et al. (2012) have determined that, unlike in secondary schools, competition amongst primary schools does slightly raise student achievement from 1999 to 2003. The effect is yet again very small: a one standard deviation increase in competition increases final nationwide exam scores by only 5% to 10% of the mean standard deviation. Whereas the authors attribute the small effect to the unavailability of quality rankings[8], Dijkgraaf et al. (2013) argue that competition amongst primary schools is limited by the (greater) value that parents attach to the average distance a school is located away from one's place of residence in picking an appropriate school for their children.

In addition to an intensification of school competition, receipt of the label of excellence could be perceived by schools as a relative increase in the average quality ranking of an excellent school vis-à-vis all other schools. In that respect, it is related to research by Koning & Van der Wiel (2010) on the impact of fluctuations in school quality rankings (annually published by a national newspaper) on the responsiveness of aggregate secondary school performance. They have estimated a fixed effects model similar to the one I adopt and determined that a higher ranking in one year[9] depresses nationwide final exam scores the following year by approximately 5% of the mean standard deviation in exam scores. In the long-run, the effect of a strong increase in a school's ranking (arguably comparable to a one-off receipt of the 'Label of Excellence') can even induce a decrease in exam scores of up to 30% of the variable's standard deviation. The authors ascribe both these transitory and more

[8] In recent years, this has changed. For instance, at scholenopdekaart.nl and zoekscholen.onderwijsinspectie.nl schools can nowadays be easily compared in terms of objective student performance and many other school characteristics.

[9] The authors investigate the effect on various learning outcomes immediately after the rankings have been published, although the rankings are compiled from 3-year old data.

long-lived effects to faltering effort[10] and reduced investment by the respective schools. My analysis is restricted to exploring transitory effects of excellence through similar causal pathways.

This paper also fits within the broader literature on school accountability and student performance on standardized tests. Under the assumption that ultimately students are to benefit from the label of excellence (i.e. on average nationwide student performance goes up), the literature on school accountability may shed light on the various mechanisms that could tie the label of excellence to student performance. As extensively discussed by Figlio & Loeb (2011), school principals, managers and teachers may not comply with the interests of stakeholders, such as parents or taxpayers. In order to align interests, discipline the school board and resolve the principal-agent conflict, the government sets targets on verifiable aspects of education, which serve to inform the stakeholders about the performance of a particular school relative to a certain benchmark. In most countries without an Inspectorate of Education, so-called test-based (i.e. 'high-stakes') accountability is 'consequential' (Ladd, 2012). This implies that funding can be withdrawn if performance is weak or that financial rewards are provided to schools which exhibit exceptional results. As a matter of fact, if student performance remains consistently below the indicated threshold, a school will eventually perish. Generally, this raises student outcomes: in a meta-analysis of 14 papers on test-based accountability Lee (2008) found that most studies report modestly positive effects on math and reading scores. However, many pitfalls to standards-based testing exist. For instance, Ladd & Zelli (2002) provide evidence that schools shift their means towards subjects which are tested at the expense of students' proficiency in non-tested fields of study. Furthermore, Burgess et al. (2005) and Neal & Schanzenbach (2010) both find that mainly students in the middle of the achievement distribution have benefitted from test-based accountability. Teachers have an incentive to focus on a large group of mediocre students, because this will on average yield the largest gains in verifiable student performance. It is plausible that in a school's attempt to obtain a label of excellence, similar (but opposite) mechanisms are at play. A disproportionate degree of attention for one's profile of excellence and/or a particular subset of students could have an adverse impact on the average student performance on tested subjects.

Although in the Netherlands, the school quality classification scheme is much more elaborate (see the sidebar in the *data*-section) and not directly related to school funding (but instead to operational autonomy), the collection and publication of test scores also *indirectly* incentivizes and disciplines schools to optimize their performance. This is what is called 'low-stakes' accountability. The Netherlands has had a long history of reporting school quality rankings by national newspapers or magazines, often in collaboration with educational researcher Prof. Dr. Jaap Dronkers. More recently, certain websites also publish final exam scores, albeit with a single or multiple year lag. Nevertheless, the information provision is often incomplete or restricted to only a few indicators and dispersed over multiple platforms, which complicates inter-school comparison. For that reason, I expect the label of excellence to enhance low-stakes accountability. Related to the market mechanism described above, the provision of information on school performance enables parents and future students to make better judgements as to which school delivers the best results. If parents and students cannot distinguish schools on quality at all (i.e. 'naming-and-shaming' is impossible), schools will have little incentive to continuously improve themselves vis-à-vis their nearest competitors. However, if the public does receive sufficient information on school quality, a high position in a school ranking (partly based on standardized test scores) or attainment of a label of excellence can be very valuable to a school[11] and *ex ante* an extrinsic stimulus to raise student performance. The latter is illustrated by quasi-experimental research conducted by Bevon & Wilson (2013) on the development of exam scores

[10] Even though this is, to my knowledge, the only paper that has examined the relationship between school quality rankings (i.e. a non-financial 'reward') and verifiable performance, the authors present convincing empirical evidence that schools may slack off out of complacency in response to positive feedback on their quality.

[11] Even though the term itself may be a misnomer.

in England and Wales following the sudden abolition of school league tables in Wales. Having controlled for time-varying factors such as funding, they find systematic evidence that school effectiveness went down strongly in Wales, causing a 0.21 standard deviation decline in exam grades per year per school. This finding is corroborated by research from Nunes et al. (2015) on Portuguese schools identified as low-performing[12]. Not only did the authors determine that schools which score poorly will receive fewer enrolments the following year, but also that the probability of closure of these schools increases, thereby mechanically raising average student performance. Rockoff & Turner (2008) find similar effects for primary and secondary schools in New York City. Interestingly, within as few as four months, students in low-performing schools managed to score up to 0.1 standard deviation higher on comparable exams.

Although not undisputed due to the unintentional side-effects associated with school accountability, the literature finds largely positive effects on student outcomes (Figlio & Loeb, 2011). Furthermore, it is incredibly cheap if the policymaker's objective is to raise measurable student performance. Hoxby (2002) calculates that a fully-fledged accountability system is 7053% less expensive than a 10% reduction in class size across the entire United States or 5011% less expensive than raising teacher salaries by 10 percent. These are huge numbers, which makes one wonder if the causal effects of the various policies are of similar magnitude. I do not know of any studies in which the benefit-to-cost ratios are juxtaposed, but findings by Hanushek (2011) and Chetty & Rockoff (2011) indicate that the economic value of teachers is humongous (and very likely to outweigh any improvements caused by an accountability system)[13]. Hanushek (2011) finds that replacing the bottom five to eight percent of teachers with average teachers yields verifiable gains in student outcomes for an average class size of 20 students equivalent to a present value of $100 trillion (!), whereas Chetty & Rockoff (2011) report that replacing a teacher with a value-added (i.e. the unique contribution of a teacher to a student's performance on a standards-based test) in the lowest 5% of all teachers with an average teacher throughout a student's educational career increases a student's lifetime income by at least $250.000. One major advantage of improving teacher quality or effort vis-à-vis implementation of an accountability system is that the effect of the former is persistent, whilst the latter only has a one-off (fixed) effect or at most intermittently in the case of a bad rating. This could explain why the monetary estimations of the economic value of (better) teachers turn out to be astronomical. Although you cannot simply replace underperforming teachers with better ones, higher salaries or pay-for-performance could also serve to improve test scores. Britton & Propper (2016) determine that a 10% reduction in the wage gap between teachers' salaries and private sector salaries raises exams scores in secondary education by 2% to 5%. Furthermore, Lavy (2002) evaluated an Israeli experiment in which teachers received pay-for-performance if their students scored relatively highest on standardized tests. He found that both in the first and second year after the pay scheme was introduced, students obtained significantly more credits and substantially higher grades. However, the downside to pay-for-performance schemes is that it may crowd-out intrinsic motivation, which could hurt student test scores (Kelley et al., 2002; Weibel et al., 2007). If extrinsic non-financial acknowledgement of effort can also crowd-out intrinsic motivation, it is even conceivable that once a label of excellence has been received the school staff loses its motivation to sustain their high levels of effort at the expense of student outcomes. In addition to enhancing teacher quality and/or upping their effort levels, class room size reductions have also been shown to instigate significant improvements in standardized test scores (Angrist & Lavy, 1999; Krueger, 1999; Schanzenbach, 2014).

[12] These papers on published school rankings are related to the aforementioned research by Koning & Van der Wiel (2010). However, Koning & Van der Wiel (2010) look at the impact of changes in school rankings on student outcomes, whereas the papers reported on in this paragraph take into account the absolute ranking and not necessarily the development over time. Since I perceive the receipt of the label of excellence as a 'shock' to a school's quality, I expect my findings to be most in line with those by Koning & Van der Wiel (2010).

[13] This is also illustrated by the case of Finland, where an accountability system is practically non-existent but teachers are highly educated and experienced, which consistently puts Finland at the top of many international education-related leaderboards (Sahlberg, 2007).

Krueger (1999), for example, presents quasi-experimental evidence that the first year that students move into a small class, their exam scores already increase by 4 percentile points. For every subsequent year in a small class, exam scores go up by another 1 percentile point per year. Although these estimates were for kindergarten, Chetty et al. (2011) find that the effects of class size reductions persist into (at least) adolescence. They report that American students who had been randomly assigned to smaller classes up until third grade (i.e. till age 8-9) are not only more likely to go to college and graduate, but also own more property, save up more for retirement and marry more.

To reiterate, most evidence points at a minor but positive relationship between school accountability and student performance on standardized tests. Hence, if one regards the provision of a label of excellence as an elaboration or extension of school accountability (to the public), it is to be expected that there will be positive consequences for student performance. However, there may also be unintentional consequences of the label of excellence, as with other indicators of school accountability. For instance, similar (but with an opposite effect) to 'teaching to the test', schools may disproportionately spend time and effort on developing their profile of excellence or on aiding a particular subset of students if this would increase their odds of obtaining a label, which would come at the expense of declining test scores[14].

Under the assumption that ultimately (in the long-run) all students ought to benefit from the label's provision[15], it is questionable whether a label of excellence is really a cost-efficient policy tool. Since the policy is very targeted[16], the gains are likely to be only a fraction of what is reported by studies which have looked at the total (positive but modest) impact of a (generally quite cheap) accountability system. Moreover, due to the bureaucratic nature of the application procedure the costs can be quite high[17]. Other policies aimed at raising student test scores might therefore be far more cost-efficient in the long-run. Furthermore, you would avoid the threat of collectively crowding out intrinsic motivation amongst schools which have become excellent. However, it also has to be noted that insights from organizational economics have shown that in the presence of (mutual) appreciation or reward of one's achievements, 'encouraging expectations'[18] boost individual motivation and effort (Dutch Council of Education, 2011; Wieringen, 2011). One of the current objectives of the Inspectorate through the provision of the label of excellence is to establish an educational atmosphere[19] in which these encouraging expectations are self-fulfilling; as such student outcomes could be raised (MacNeil et al., 2009). Additionally, the label should instil schools with a sense of academic optimism. Hoy et al. (2006) have found that academic optimism – an umbrella term comprising a school's emphasis on performance, self-efficacy and trust in students and parents – significantly raises student achievement too. It remains to be seen which effect empirically prevails.

The paper is organised as follows. In section 2 I extensively describe the label of excellence and the associated application and selection procedure. Furthermore, I elaborate on how my dataset has

[14] Obviously, if there is a positive effect of the label of excellence on non-excellent schools' test scores and a negative effect on excellent schools' test scores, the former effect will dominate because there are simply far more non-excellent schools. However, this would also paradoxically imply that in the wake of becoming excellent, excellent schools would have become absolutely and relatively less excellent (as defined by a school's performance on verifiable student outcomes).

[15] As stated before, the ambiguity regarding the policymaker's objective with respect to the label of excellence makes it unclear as to what the definite goal of the label's provision actually is. It is also possible that the policy's effect on test scores has fully been disregarded in its conception or that an average decrease in test scores is willingly traded off for gains in other areas deemed important, which may or may not be verifiable.

[16] Not every Dutch school will respond to the policy for a variety of reasons. For instance, schools may be unfamiliar with the policy or face no direct competition from potentially excellent schools. Moreover, not all parents and (future) students will be familiar with the policy either.

[17] Unfortunately, my request for information on the costs of 'Excellent Schools' was not honoured.

[18] In Dutch: *"aanmoedigende verwachtingen"*

[19] In Dutch: *onderwijscultuur*

been composed of various constituent datasets, compare excellent schools with non-excellent schools on a set of background characteristics and also discuss the covariates that I have incorporated in my main analyses. In section 3, the model and identification strategy are explained. In section 4, I present my main results. In section 5, I run a number of robustness checks. In section 6, I discuss the findings, relate them to some of the hypotheses and assumptions from the introduction and provide suggestions for further research.

Section 2: Data

2.1.The label of excellence

At the beginning of 2011, the former Dutch Minister of Education, Culture and Science[20], Marja van Bijsterveldt, compiled a list of measures, ideas and suggestions in order to bolster the quality of the Dutch secondary education system (with the exception of VMBO-B[21] and year 3 and 4 of all VMBO tracks[22]). Having signalled that Dutch students scored relatively less well on the internationally comparable PISA tests for math, reading skills and science as well as that nationwide final exam grades had dropped slightly over the course of seven years (from 2003 to 2009), the newly formed *Rutte-I* government (2010-2012) was keen on developing a plan of action to go from *'good to better quality [secondary] education'.* Dutch secondary education internationally still ranked fairly high (despite the signalled stagnation in performance), but very few students in the Netherlands performed exceptionally well (or exceptionally bad). Therefore, the Coalition reasoned that much could be gained in terms of educational achievement if highly able students would be motivated and encouraged to live up to their full potential. Hence, many of the newly introduced policy proposals were targeted at these underperforming students, for instance through formal acknowledgement and recognition of their exceptional effort. More generally, the entire educational system was to become more ambition-sensitive and better tailored to individual students' needs and wishes, so as to enhance overall student performance. This also entailed closer monitoring of objective student performance and a goal-oriented approach to learning by schools. If schools would repetitively set their own targets based upon clear benchmarks, student outcomes could be improved upon in an efficient fashion, according to the Minister.

Within this renewed educational framework – as embedded in the Better Performance Action Plan[23] – characterized by closer attention for individual qualities and goal-oriented school performance, the Minister, drawing on recommendations from the Dutch Council of Education (hereafter referred to as the Council), proposed to confer a label of excellence[24] to schools of outstanding quality (Dutch Council of Education, 2011; Ministry of ECS, 2011a). The Council argued that quality must not only be rewarded at the individual level, but also at the school level. Schools which deliver extraordinary quality must receive appropriate acknowledgement, otherwise the extrinsic incentive to excel is practically non-existent. Back in 2011,

Quality Judgements by the Inspectorate of Education

The Inspectorate's analysis of a school's quality is intricate and comprehensive. Out of 45 general indicators (pertaining to five distinct aspects of quality: educational outcomes, study trajectory, care for students, care for quality and rules and regulations), respectively 10 indicators in PE and 15 indicators in SE are so-called 'target indicators'. Depending on a school's performance on these indicators, a school either receives the label (in descending order of quality) baseline quality, weak or very weak. Weak or very weak schools receive additional supervision from the Inspectorate. Although the rules are fraught with exceptions, in general schools will be classified as weak or very weak if the educational outcomes (e.g. population-corrected grades) are insufficient. From 2017 onwards, 'good' has been added to the classification scheme, although formally not as an indicator of quality but as an encouragement to schools to continue pursuing their own challenging goals beyond the baseline level required by the Inspectorate. However, 'Excellent Schools 2017' are excluded from my analysis, and hence I do not have any 'good' schools in my dataset (Inspectorate of Education, 2012; Idem, 2017b).

[20] In Dutch: *OCW*

[21] See Appendix II for a description of the various educational levels within the Dutch secondary education system.

[22] In Dutch: *bovenbouw*

[23] A similar action plan – the Fundament for Performance Action Plan – was drawn up for primary education, in which a likewise plea was made for more ambition-tailored education (Ministry of ECS, 2011b).

[24] In Dutch: *excellentiepredicaat*

the Dutch Inspectorate of Education – responsible for the inspection, supervision and review of school quality across all levels of education in the Netherlands – used a three-tiered classification of school quality: very weak, weak and baseline quality. The degree of supervision by the Inspectorate is negatively related to school quality. Whereas poor-performing schools receive a lot of negative publicity as well as additional scrutiny from the Inspectorate, schools in the entire spectrum of mediocre to exceptional quality all receive the same undifferentiated verdict from the Inspectorate: baseline quality. According to the Council, this has tended to demotivate schools that in potential could deliver much better student outcomes, for example through an increase in student '*upflow*[25]' or high student grades. In order to spur creativity and ambition as well as strive for excellence, the Council therefore suggested to expand the classification scheme with two new tiers: good and excellent. Only recently, in August 2017, 'good' was added to the Inspectorate's quality classification scheme, even though formally it is a mere 'acknowledgement' for reaching an ambitious self-imposed standard of quality rather than a quality judgement[26]. Regarding a label of excellence (and promotion thereof), progress was made much more quickly. Already at the onset of 2013 (less than two years after the parliament was informed), the Ministry published the first shortlist of excellent schools in both (special) primary (hereafter: *PE*) and (special) secondary education (hereafter: *SE*).

In order to be eligible for the honorary title of 'Excellent School', schools must first apply through filing an extensive application form in which the school reflects on its own achievements and 'profile of excellence'[27]. For 3 years (i.e. from Excellent Schools 2012 till 2015) the Ministry was responsible for the ensuing selection procedure. Six judges (hailing from politics, science or education) and external experts (i.e. former deans or inspectors) adjudicated on the competences and distinctive, excellent nature of the school. On the basis of five main criteria subdivided into fourteen indicators of quality, the judges determined whether a school deserved to receive a label of excellence. These criteria encompassed (1) objective student outcomes, (2) adaptability to the circumstances within which the school operates, (3) a school-specific vision which verifiably pays off, (4) responsiveness to self-signalled developments in school results and (5) a clear profile of excellence which pervades through all facets of the school. During the time that the applications were directly dealt with by the Ministry of ECS, the school results (criteria 1 and 2) were of predominant importance in assigning a label of excellence. However, ever since the Inspectorate of Education has taken over the responsibility from the Ministry (in 2015), the quality of the school's profile of excellence has become the dominant criterion upon which the judges' verdict regarding excellence is based.

A profile of excellence is a rather elusive concept. Since schools are pluriform entities and unique in many different ways, displaying excellence is not solely restricted to schools that on average perform best on final exams or in terms of other measurable criteria. Therefore, the Inspectorate defines a profile of excellence as a school having an *"inspiring, innovative or motivating curriculum or a distinctive approach to education for a specific group of students"* (Inspectorate of Education, 2018b). For example, for the fruitful integration of children of asylum-seekers (at the compound of an immigration detention centre) into the Dutch education system, a primary school in Ter Apel received the label of excellence earlier this year. Similarly, a primary school in Utrecht received the label in 2016 for the quality of arts and culture classes provided to highly gifted students. Furthermore, schools have also been rewarded the label of excellence for high-quality music classes, early foreign language

[25] 'Upflow' as a school quality indicator refers to the number or percentage of students which have managed to attain a higher level of education than their primary school teacher has advised. Hence, a high upflow rate is a sign that a school contributes to the student's knowledge base and facilitates the development of (cognitive) skills and abilities. In other words, the school proves its relevance. Upflow can, however, be easily manipulated through simply disallowing students to climb up the educational ladder very rapidly. The school has an incentive to act this way: if mediocre students would be allowed to 'flow up', this will put downward pressure (with a 1, 2 or 3-year lag) on final exam grades, which the school wants to avoid.

[26] For more information, read the side bar on the Inspectorate's quality judgments.

[27] In Dutch: *excellentieprofiel*

programmes, superior bilingual or multi-lingual education, (classes on) civic responsibility, the shrewd use of digital equipment in the classroom and much more. In practice, it so happens that as long as a school's objective results are sufficient (i.e. it provides baseline quality according to the Inspectorate's standards[28]) relative to its peers, any school with a sustainable, verifiable and distinctive manifestation of excellency could qualify for a label of excellence (Inspectorate of Education, 2016b).

In shifting the responsibility from the Ministry of ECS to the Inspectorate for the supervision of the application procedure, the profile of excellence took centre stage in the process of determining whether a school was deserving of the label. At the same time, the Inspectorate responded to schools' criticism by extending the validity of the label from 1 to 3 years. Previously, schools had to reapply and go through tons of paperwork each year in order to retain the label. Nowadays, unless a school deliberately abandons its excellency profile, fails to sustain its high level of quality or merges with another school, the label can be kept for 3 years (Inspectorate of Education, 2016b). This implies that schools that partook in 'Excellent Schools 2015' and have received a label of excellence in January 2016 are still entitled to the label of excellence by December 2018. The re-application in order to also be eligible for the label from 2019 onwards must, however, already be filed during 'Excellent Schools 2018' (which commences early 2018). A timeline with important dates can be found in Appendix I. The shift in responsibility for the label's provision occurred in conjunction with a tacit change in the label's objectives. In the Better Performance Action Plan, the Minister still spoke of *"that system [of judgement of excellence, which] predominantly rewards the accomplishments of schools to maximize the results of students"* (Ministry of ECS, 2011a). Although the *"broader didactic role of schools"* will also be taken into account, it is telling that the label of excellence was first proposed as part of a broader set of reforms aimed at raising objective student performance as enumerated in the aforementioned plan of action. Nowadays, the Inspectorate (2018b) describes the policy's objective as *"making quality visible [to the world] and disseminating it [amongst schools]"*. Furthermore, the label *"should contribute to the creation of a 'culture of education'[29] in which all stakeholders naturally strive for improvement and share their expertise"*. Hence, there seems to be an implicit shift in the policy's objective from initially being largely focused on explicitly upping the student's performance to mainly raising the general performance of all staff members and valorising educational initiatives tailored to students' special needs and wants, which is important to keep in mind when analysing the results of the label on objective student outcomes.

2.2. Excellent schools

The label of excellence has been conferred for six years since its inception (i.e. from January 2013 to January 2018). Since information on final exam scores or control variables for the year 2018 (i.e. 'Excellent Schools 2017') is as of now unavailable, information on the excellency status over 2018 has not been included in the dataset. In order to identify the effect of the label of excellence on exam scores in the first ever year the label was conferred (i.e. 2013), you would at least need data from 2012. Luckily, for both primary schools and secondary schools, I managed to retrieve data on exam scores obtained in 2012. Concerning secondary schools, I even have data on exam scores from up to five years before the label of excellence was established. Therefore my (unbalanced) panels range from 2008 to 2017 for secondary schools and 2012 to 2017 for primary schools.

Concerning the primary schools, the panels are grouped by BRINVEST, which is a portmanteau of BRIN and *vestiging* (in English: location). The BRIN-number is a 4-digit alphanumerical code which uniquely identifies a scholastic institution in the Netherlands. Some schools consist of multiple

[28] Onwards from 'Excellent Schools 2017', the new classification scheme from the Inspectorate is applicable. Hence, conditional on possession of the qualification 'good', a school can become 'excellent'. However, these schools only received the label at the onset of 2018, and, as a matter of fact, have not been part of this research (as excellent schools) because the required data on final grades for this year is not available at the time of writing.

[29] In Dutch: *onderwijscultuur*

locations (i.e. *a cluster*), and therefore, if necessary, a 2-digit number is attached to the BRIN-code to indicate the exact location[30]. Application for the label of excellence for primary schools must be done by cluster. However, it can be the case that within a particular cluster one location obtains the label and the other does not. Concerning the secondary schools, the panels are similarly grouped by BRINVEST. However, secondary schools (often) provide multiple levels of education. Since these various levels cannot be uncritically compared in terms of exam scores, excellence and other (control) variables, the panels are further sub-divided into BRINVEST by level of education (distinguishing between VMBO B, VMBO K, VMBO G(T), HAVO and VWO). Labels of excellence are also only issued separately per level of education, and therefore a school must file multiple applications if it would like to become excellent within multiple educational levels. There is only one exception to this rule: from 'Excellent Schools 2012' up to and including 'Excellent Schools 2014' VMBO applications for excellence were filed for the entire department.

Table 1

Year label is held by BRINVEST (1 year post-application)	Applications (excluding special schools & PRO) + VMBO pooled	Labels assigned (excluding special schools & PRO) + VMBO pooled	Number of excellent BRINVESTS (by educational level) in dataset	Number of non-excellent BRINVESTS (by educational level) in dataset
2013	Total: 165[31]	PE: 32; SE: 24 Total: 56	PE: 30; SE: 34 Total: 64	PE: 5962 ; SE: 3017 Total: 8979
2014	PE: 52; SE: 64 Total: 116	PE: 32; SE: 37 Total: 69	PE: 27; SE: 51 Total: 78	PE: 6005; SE: 3070 Total: 9075
2015	PE: 60; SE: 89 Total: 149	PE: 34; SE: 51 Total: 85	PE: 30; SE: 72 Total: 102	PE: 6593; SE: 3049 Total: 9642
2016 (valid till December 2018)	PE: 62; SE: 103 Total: 165	PE: 37; SE: 61 Total: 98	PE: 35; SE: 71 Total: 106	PE: 6522; SE: 2921 Total: 9443
2017 (valid till December 2019)	PE: 46; SE: 60 Total: 106	PE: 25; SE: 19 Total: 45	PE: 19; SE: 30 Total: 49	PE: 6410; SE: 2916 Total: 9326

By January 2017 (after the labels over 2016 had been assigned), there were 184 excellent schools composed of 69[32] primary schools, 93 high schools and 22 schools (both primary and secondary) for special education. Out of the 93 high schools, 34 schools were VMBO, 22 HAVO, 24 VWO and 13 PRO. Special schools and PRO-schools have been excluded from my dataset, because the variation in excellence over the years is very limited and the sample size small. Moreover, these schools cannot be compared with the 'ordinary' schools in terms of exam results, because the children in these schools either make an entirely different exam or no exam at all (College voor Toetsen en Examens, 2018).

[30] A typical code looks like this: 03EH-01. This code belongs to OBS 'T Noorder Merk, a primary school in Noordeinde, Gelderland. A fellow location (de Wereldweide) operated by the same directorate in Wezep, Gelderland, has code 03EH-00. If a BRIN consists of only one location, *00* is the default location number.

[31] It could not be ascertained how many applications there were in 2012 for respectively primary and secondary schools.

[32] Whereas in a recent article (January 2017) the Inspectorate states that there are 69 excellent primary schools, the annual spreadsheets I have used only list 62. This discrepancy can most logically be explained by a slightly different classification of a subset of the special primary schools.

The data in Table 1 have been distilled from lists on excellent schools that the Inspectorate of Education publishes annually. As can immediately be seen, column 3 and 4 do not report identical values. The Inspectorate lists a smaller number of labels of excellence for secondary education, because they subsume multiple excellent VMBO departments within one school under one label of excellence. Since my analyses make use of the variation in exam scores across the various VMBO departments, each VMBO department is assigned its own label in column 4. Moreover, my dataset consequently includes a slightly smaller number of primary schools. This is a simple consequence of the dataset that the Inspectorate made available for usage. I managed to retrieve the final exam scores by primary school for the years 2012 up to and including 2017, but found out that in some instances schools with a location affix other than *00* were not included as a separate administrative entity in the dataset. Although I was able to verify with one school that its yearly data was merged with the data of the main administrative location (with the *00*-affix), it is beyond the scope of this research to repetitively verify this for all excellent schools which were omitted from the dataset. Moreover, even if I would be able to ascertain this, it would not help me any further, because I would need access to the disaggregated data by school. Therefore, on average 3.8 excellent primary schools per year are omitted; a little over 10% of the total data on excellent primary schools to which I have access.

In Appendix III, the data on excellent secondary schools is split up by both year and level of education[33]. As can be inferred, the sample sizes by year are relatively small for the various levels of education. In some years, there are as few as three excellent schools within a particular category (i.e. HAVO in 2013). Inarguably, this weakens the statistical power of some estimations. The main fixed effects model (see: *methodology*-section) is, however, less affected by these relatively small sample sizes within educational levels, because it makes use of the variation in excellence within all subjects (i.e. BRINVESTS) over time. In fact, over time 78 different PE-panels[34] and 130 SE-panels display variance within its excellence status. To put it differently, in the dataset there are respectively 78 primary schools and 130 educational levels within secondary schools that at least for one year have been excellent. Nevertheless, alternative model specifications may lose statistical power due to relatively large standard errors, in particular models (with variables that interact) with the level-of-education-variable. Furthermore, inclusion of leads and lags of the independent variable and covariates restrict the model to a smaller subset of observations – in the latter case due to missing data – and thereby also reduce the statistical power of the estimation. Therefore, the sample size is at all times reported and will be reflected upon whenever the number of observations becomes worryingly small.

Even though only from 'Excellent Schools 2015' onwards the label of excellence's validity was extended from 1 to 3 years, both primary and secondary schools that participated in the programme pre-2015 often already boasted the label for multiple years. The tables in Appendix IV sum up the probability of switching treatment (i.e. the transition probability), conditional on the panel's current status. As can be inferred, the probability of remaining excellent is much higher than dropping back to 'ordinary' for both primary schools (77% vs. 23%) and high-schools (86% vs. 14%). Similarly, the last column in Tables 2 and 4 in Appendix IV tell us, conditional on the panel *ever* having a particular excellence status, what fraction of the time the observations within that particular panel are in accordance with its conditional status of excellence. Concerning the primary schools that at one point in time are excellent, 41% of the total number of observations within these panels are excellent. For secondary schools this statistic is slightly lower at 27%, because the dataset on secondary schools spans a larger timeframe before the policy was implemented (onwards from 2008). So even though a

[33] Data for secondary schools split up by only level of education can also be found in Appendix III.

[34] Four excellent primary schools are solely excellent in the dataset, which brings the total of excellent PE-panels to 82. However, these panels are excluded from the fixed effects regression because their status of excellence is time-invariant.

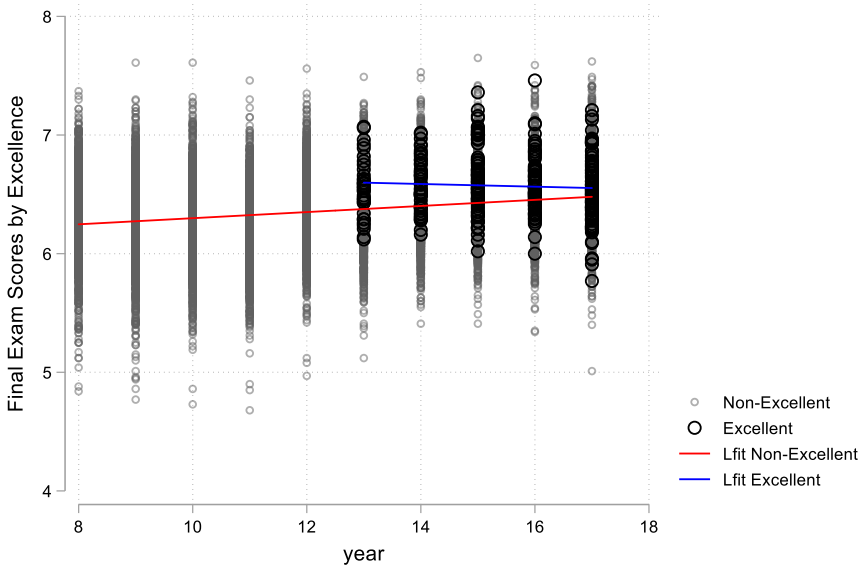
school or level of education's label of excellence is generally kept for more than a year, in both cases more than 50% of the observations belong to a time period in which the panel was not excellent.

2.3. Final exam grades

2.3.1. Secondary education

In the Netherlands, all students in their final year of high school must participate in the nationwide final exams. A student must take exams in any course in which he wishes to graduate (with a few exceptions like civics). Graduation depends for 50% on the grade received at the final exam and for 50% on a weighted average of 'school exams'. The final exams are standardized and identical for all students from a particular level of education, but differ from year to year. Even though the exams are supposed to test the level of knowledge that a student is expected to have acquired by the end of his school career, the authors of the exam cannot avoid that the level of difficulty varies over time. Despite an adjustment to the grading scheme to correct for the relative degree of difficulty, the nationwide exam average per subject changes constantly; the so-called year fixed effects. My dataset on exam scores contains information on the average of all subjects per BRINVEST from the year 2008 up to and including 2017. The data has been provided by the Dutch Inspectorate of Education through the Data Archiving and Networked Services (hereafter: DANS) platform. I have merged data from the supervisory reports[35] from 2009 to 2015 with the educational outcome reports[36] from 2016 to 2018 in order to construct a workable panel dataset. Since these reports are annual evaluations of the Inspectorate of the previous year, information on 2018 is only published mid-way 2019. As I can identify which panels are excellent in what year, it is possible to graph exam scores by excellence (graph 1a). The red and blue solid lines are lines of best fit. Excellent schools on average across the years score 0.13 grade points (0.49SD) higher on the final exams. Furthermore, the blue line suggests a tentative downward time trend in exam scores amongst the entire pool of excellent schools, but this cannot definitely be concluded, as the fitted model is insignificant. The fitted upward-sloping red line does significantly predict the observed values.

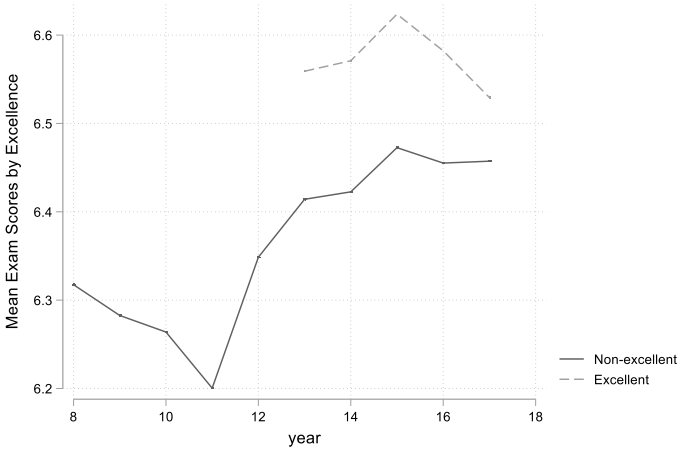
Graph 1a



[35] In Dutch: *"toezichtkaart"*

[36] In Dutch: *"onderwijsresultaten"*; previously known as *"toezichtkaart"*.

Graph 1b



Closer inspection of the time trends in mean exam scores by excellence (graph 1b) as well as by excellence and level of education reveals more (graph 1c and Appendix V). In graph 1b, you can see that not only do mean exam scores seem to vary quite strongly over time, but also that excellent and non-excellent schools move along a common trend (with the exception of 2016 and 2017). In graph 1c, you can find an overlay plot of the mean time trends for the five different levels of education by excellence. As can be inferred, the various levels of education also display roughly similar time patterns in exam scores. The sole exception that clearly stands out is VMBO-B: excellent VMBO-B schools in 2013 scored even slightly worse on the final exam than non-excellent VMBO-B schools. Furthermore, VMBO-B exam candidates consistently score much higher than any other level of education. In Appendix V, the five sub-plots which are overlaid in graph 1c are included for the sake of clarity and completeness. Moreover, in Table 2 the mean difference in exam scores over time is reported as well as the mean difference expressed in terms of standard deviations.

Graph 1c

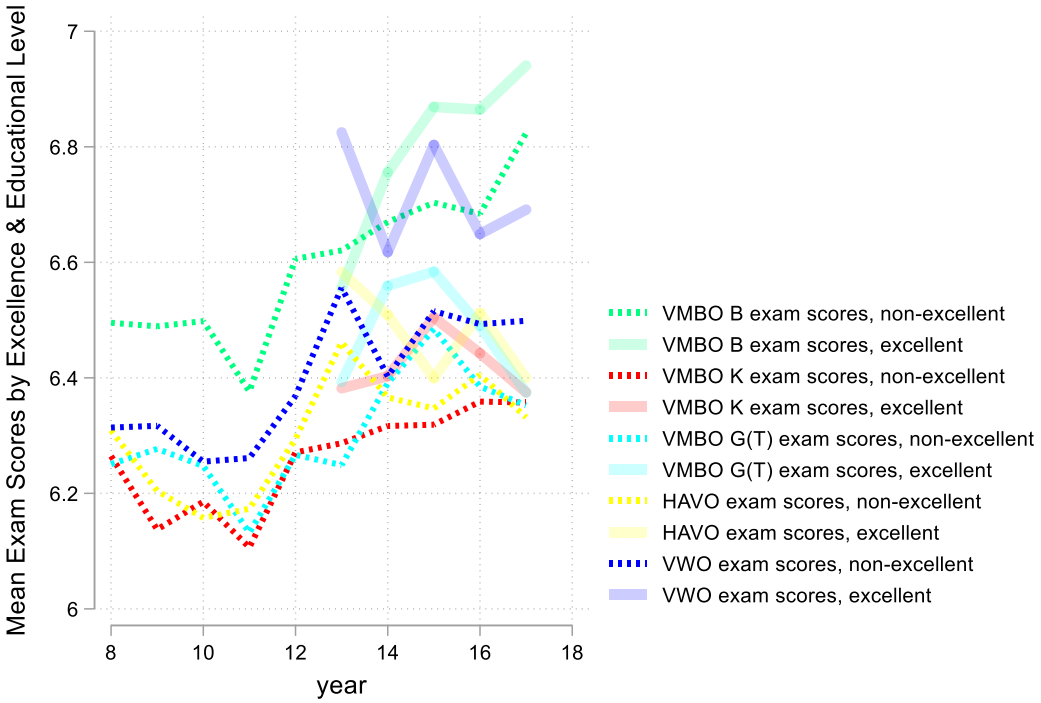


Table 2

VARIABLE	*VMBO-B*	*VMBO-K*	*VMBO-G(T)*	*HAVO*	*VWO*
Mean difference in exam scores (Ex - N.ex)***[37]	0.14	0.10	0.10	0.07	0.21
Standardized mean difference in exam scores	0.56SD	0.47SD	0.40SD	0.38SD	0.93SD

Regardless, these visual representations fail to inform us in any way about the policy's effect on final exam scores, because the composition of excellent schools is constantly in flux[38], schools receive or lose a label of excellence (i.e. *the treatment*) at various points in time and other time-varying, possibly confounding variables cannot be controlled for. Nor can we determine what the pre-policy (time trend in) exam scores of schools that have become excellent looked like, which could be used as an 'excellency' fixed effect. As a matter of fact, it is impossible to infer from these graphs what approximately the policy's impact is on final exam scores. More sophisticated panel analyses are logically required for that *(see methodology*-section*)*. This also applies to the graphs pertaining to primary schools, which are dealt with below.

2.3.2 Primary education

Unlike the data concerning (excellent) secondary schools, it was much harder to obtain the mean final exam scores by BRINVEST for primary schools. Despite my efforts, I have only been able to lay my hands on the last 6 years of exam data (from 2012 up to and including 2017). Whereas the data on high-school exam scores were provided through DANS (for which permission from the Inspectorate was required), data on primary schools' final exam scores was publicly available for the last three years through the website of DUO[39] (i.e. a government agency responsible for the execution of numerous education-related tasks). Since 6 years is the absolute minimum number of years required to fruitfully conduct my intended analyses, I subsequently contacted a Dutch news agency, RTL News, who in the past published annual rankings of primary schools using previously available exam data from the Inspectorate. Due to their benevolence, I managed to bypass the Inspectorate and obtained three additional years of exam scores classified by primary school.

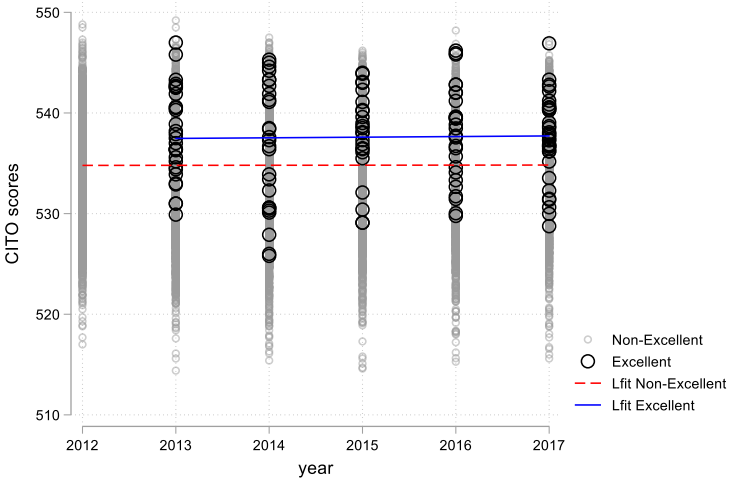
Since 2015, every student in the Netherlands in the last year of primary school (with the exception of students with severe (learning) disabilities) must take a final exam (Cito, 2018). The score obtained at such an exam serves as an objective and independent indicator of the cognitive capacities of the student and is of great importance in enrolling into an appropriate level of secondary education and school of choice. Before 2015, the exam was administered in February, a while before the teacher would give his/her advice on the presumed level of education that the soon-to-be high-school student had attained during primary school. Since the February exam put a lot of pressure on children to do well at the test (because it informed their high school advice), the law was changed and from 2015 onwards (at the same time schools were compelled to administer a final exam) the exam is taken in April/May (HP/De Tijd, 2014; PO Raad, 2014). Despite these developments, the final nationwide exam is still deemed valuable as a crude indicator of a student's capabilities. These regulatory changes may have had an effect on student performance, thereby making it vitally important to control for a yearly time trend. Besides that, it could well be that the effect of 'Excellent Schools 2012 & 2013' is harder to pick up, because the conferral of the label precedes the exam by approximately a month.

[37] All mean differences are significant at the 1% level.

[38] As exams differ by educational level, a mere change in the composition of the educational level of the schools that are labelled 'excellent' will mechanically alter the mean exam scores by excellence per year, even if this would theoretically occur in the absence of any real changes in exam scores over time.

[39] In Dutch: *Dienst Uitvoering Onderwijs*

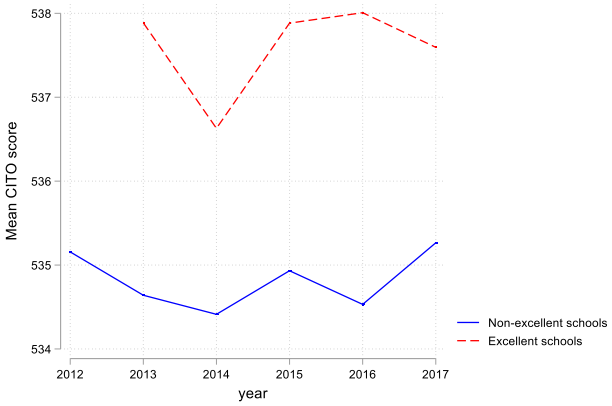
Graph 2a



In 2016, 77.0% of all students (N=139.287) that participated in the final nationwide exams took the Cito test[40] (Expertgroep PO, 2016). Before a school's participation in the final nationwide exams was made obligatory in 2015, this test was by far the most popular amongst schools to assess a child's cognitive and intellectual development. However, in response to the new law, a few new suppliers entered the privatized exam market. Therefore, the Cito test, created by the Central Institute for Test Development (Cito), a *de facto*[41] government institute, lost some ground to other providers of exams such as A-VISION (ROUTE-8 test – 6.5% market share) and Bureau ICE (IEP test – 16.5% market share). Although I have data on the exam scores for these tests from 2015 to 2017 (see Appendix VI), the number of excellent schools that have taken either the ROUTE-8 or IEP test up until now is too small; in 2017, a meagre 6 excellent schools took the ROUTE8 exam and 12 excellent schools the IEP test. Furthermore, as the exams were only introduced post-2014, two years of variation in the excellence status of schools cannot be exploited.

Cito exam scores lie in the range of 501-550. Much in accordance with what we observed for high-schools, graphs 2a and 2b shows that excellent primary schools systematically perform better on the final exam. In fact, excellent schools score on average 2.87 points (0.69SD) higher on the Cito test than non-excellent schools. Contrary to the high-schools, the blue line of best fit seems to suggest a minor increase in Cito scores over time. However, yet again the fitted model was insignificant. The red line is remarkably flat: students in non-excellent schools have been equally skilled from year to year[42].

Graph 2b



[40] Formally referred to as the central final exam, or in Dutch: *Centrale Eindtoets*.

[41] The exam market is privatized and Cito a private organization. However, the Dutch government heavily subsidizes (part of) Cito, which used to be a public organization up until 1999. Moreover, they legally have the obligation to develop a 'central final exam'. Schools may opt for an alternative, which is, unlike the Cito, not free of charge (Didactief, 2013).

[42] Cito makes use of an equalization method (or 'anchoring') in order to render scores comparable over time. Hence, a score of 535 in one year is nearly identical to a score of 535 the year after (Zijlstra, 2017).

A more detailed examination of the mean exam scores by excellence over time exposes a differential time trend amongst the treatment and control group. In addition to a more pronounced dip in exam scores in 2014, in both 2016 and 2017 the change in exam scores compared to the previous year is opposite to the non-excellent schools. The blue line is inarguably a poor counterfactual, as I noted already in the paragraph on secondary schools. Not only does the composition of excellent schools change annually, so do plenty of time-varying background characteristics which differ across excellent and non-excellent schools. Covariates that are correlated with the policy (i.e. receiving a label of excellence) must be identified and controlled for, otherwise the impact of these control variables on exam scores is wrongly attributed to the policy. Next, I discuss these covariates, how they differ across a school's status of excellence and in what sense they may confound the estimates of the model which is outlaid in the ensuing methodology section.

2.3.3. Covariates

In Table 3 and Appendix VII I present some background characteristics on the difference between excellent and non-excellent schools for respectively secondary and primary education. The variables in italics are covariates. As can be seen, I do not have information on the same number and type of variables for both primary and secondary schools, because my dataset is composed of multiple subsets with different administrative origins. The majority of data on secondary schools comes from the Inspectorate, whereas the majority of data on primary schools comes from DUO. Both governmental agencies collect and manage their own data. Moreover, since both organizations have different objectives and legal tasks, the (statistical) content they publish on Dutch schools is different too. Furthermore, my primary concern was to obtain a dataset with exam grades, separately for both primary and secondary education by BRINVEST and status of excellence. Sometimes the arrangement and nature of the data on particular variables prevented merging with the master dataset by BRINVEST[43] and required some variables to be dropped.

Furthermore, data on covariates in particular is often not publicly accessible for reasons of confidentiality. Hence, a few potential covariates which I had wished to include in my analyses were perforce excluded. For example, the (socio-)economic composition of the yearly inflow to schools is shrouded in secrecy (for both SE and PE) and I lack data on learning disabilities in primary education. At times, data was only available for a particular time period. For instance, statistics on *flow-through* and *characteristics of students taking the final exam* for high-schools from 2008 up to and including 2012 are missing (and the former indicator in its entirety for primary schools). Despite these omissions, I am confident that the most important time-varying variables are part of this dataset and many of the theoretically confounding trends can be neutralized in order to tease out causality. I will deal with these covariates in the order in which they appear in Table 3 and Appendix VII, starting off with any potential covariates related to secondary schools. Afterwards, I will discuss some of the main differences between excellent and non-excellent schools with reference to their background characteristics.

Flow-through – SE

Data on flow-through is available for both the lower and upper half of secondary school, although the flow-through in the lower half of secondary education will not be included as a covariate[44]. The flow-through refers to the number or percentage of students who manage to complete either the lower or upper half of high-school within the designated time period. Typically, exam scores are negatively associated with the flow-through at a school. If most students complete high-school very quickly, even students who are only barely sufficiently prepared to graduate will have to participate in the final

[43] This, for instance, applied to the inflow in year 1 by high-school and educational level.

[44] This is because the flow-through in the lower half of secondary education will only affect exam grades with a multiple year lag. However, inclusion of these lagged controls will eliminate too much data and has therefore been decided against.

exams. Needless to say, this is detrimental to exam scores. Yet again, if the flow-through would in any way be correlated with the receipt of the label of excellence, it will confound my estimates. Hence, it is included as a covariate too. The number of students monitored to compute the flow-through (*flow-through N*) estimator is included as an additional control. It is nevertheless not very likely that manipulation of flow-through occurs on a large scale: flow-through is one of the four indicators of school results on which the Inspectorate bases its quality judgement; the others being *upflow*, *final exam scores* and *the difference between final and school exam scores*. Therefore, the Inspectorate can easily correct the exam scores for flow-through manipulation.

Characteristics of students taking the final exam – SE

In order to control for changes in the composition and size of the student population potentially correlated with the treatment, I have information on not only the total number of students which take the final exam and the total number of courses they have engaged in but also on whether they live in an apcg-region[45] or struggle with learning disabilities. Generally, larger classes, more attended courses and an increase in the number of students from weak socio-economic backgrounds or with learning difficulties hurts exam scores.

The number of students taking the Cito – PE

For much the same reasons as the inclusion of the total number of students taking the final exam in secondary education as a control, I also control for the total number of students which take the nationwide Cito test in primary education.

Total number of students at BRINVEST – PE & SE

The total number of students by BRINVEST (and level of education) will also be controlled for in both PE and SE, even though the amount of funding a school receives increases proportionally with the student tally and part of the effect will already be captured by controlling for the number of students taking the final exam/Cito. Nevertheless, it could very well be that fixed means must now be shared by a larger number of students, which could have adverse effects on (objective) student performance. If a sudden school expansion would be correlated with receipt of the label of excellence, this could potentially confound my estimates.

Impact area – PE

An impact area is similar to an apcg (see Footnote 45), but different in two respects: (1) the migratory descent of the breadwinner is irrelevant and (2) additional funding depends on the school's location rather than the place of residence of its students (Ministry of Finance, 2017). Hence, the impact area is an indicator of the socio-economic status of a school. Although it is a rather stable variable – an area generally remains an impact area for a long period of time – it is not a fixed effect either. Therefore, I still control for changes in a neighbourhood's socio-economic classification.

(Socio-economic) weight of the school – PE

This covariate is a weighted school-specific indicator of the educational background of the students' parents. A high value implies that only a few parents have had schooling beyond pre-vocational secondary education (i.e. VMBO). If the weight is zero, this implies that all students' parents have completed a degree beyond at least VMBO. Furthermore, the variable serves as a proxy for the educational abilities of the student population. A change in the school weight could indicate that the

[45] An apcg (*armoedeprobleemcumulatiegebied*) is a geographical area with a high density of unemployment, relatively low incomes and breadwinners from a non-Western migratory background. Apcg-areas receive additional funding from the government per student.

socio-economic composition of the school is in flux, which may have consequences for exam scores. Therefore, it needs to be controlled for.

Excellent vs. non-excellent schools

Excellent schools are said to deliver exceptional quality, although this is subject to much debate and discussion, as I have previously illustrated. In order to find out how excellent schools actually compare to non-excellent schools[46], I have run a set of t-tests on some background characteristics of both groups. As can be inferred from Table 3 and Appendix VII, both groups are significantly different in almost all aspects. With the exception of the number of students with learning disabilities and the impact area they are located in, any insignificant finding is most likely attributable to the small sample size. This applies to the average score on the IEP and ROUTE8 test and the total number of students that took the ROUTE8 test per school. Furthermore, we can safely conclude that excellent and non-excellent primary schools are evenly distributed across impact areas and the number of students with learning disabilities that took the final high-school exam is approximately equal amongst schools with a different status of excellence.

Excellent schools also score (significantly) higher (but not necessarily better) than non-excellent schools (i.e. the mean difference is negative) on the majority of indicators. Not only do excellent schools contain a disproportionately large number of HAVO and VWO-schools, but also students' 'upflow' is higher. Furthermore, students complete both the lower and upper half of high-school faster, despite excellent schools being larger and containing more students from apcg-regions (even though in relative terms they may still educate less students from these areas). The mean difference between excellent and non-excellent schools of the mean difference between final and school exam scores has a positive sign, which implies that the mean difference between the school and final exam grades is smaller for excellent high-schools, which is generally perceived as a feat of quality. Unclear is whether this also applies to the mean change in exam scores from year to year, which is significantly smaller for both PE and SE excellent schools[47].

Regarding the excellent primary schools, it is very striking that they contain a disproportionate number of students from parents with a low level of education and yet score above average in terms of final exam scores. Part of this can be explained by the mere fact that excellent primary schools are larger, but as the school weight is partly corrected for school size, this observation seems to be illustrative of the common conception that excellent schools do not only bear a 'profile of excellence' but are also simply good-quality schools. This is in line with the findings by Regioplan (2015) and Regioplan (2016).

This is also corroborated by the Inspectorate's judgements on the four indicators of quality of high-schools listed above: *upflow*, *flow-through* (for both lower and upper SE), *final exam scores* and *the difference between final and school exam scores*. For these four variables, the Inspectorate calculates a panel-specific norm adjusted for the number of students with learning difficulties or from underprivileged areas. A school may underperform on one of these indicators, but if it underperforms on two or more indicators, the 'calculated judgement' by the Inspectorate concerning these indicators of objective school results becomes negative, which may have repercussions for the school's

[46] I am intentionally comparing apples with pears here to illustrate the relative strength of excellent schools vis-à-vis non-excellent schools. In order to determine if an excellent school is deserving of a label of excellence (on the mere basis of objective criteria), you would instead have to compare (a subset of) excellent schools with a carefully constructed control group with similar features and characteristics, as, for instance, executed by Dronkers (2014). This is not only beyond the scope of my investigation, but also impossible due to data limitations.

[47] Regressing this variable on excellence and a set of controls will serve as a robustness check to the main analysis. A smaller year to year change is commendable if the mean change over the years is negative, but unfavourable if the mean change is positive. The latter would imply you have performed relatively less well than the other group.

supervisory arrangement[48]. In Appendix VIII, we can see that only 1 excellent school in a certain year scored insufficiently on aggregate judgement. Such a small number is to be expected in a dataset of this size. Regarding the sub-indicators, below-the-norm scores are rare for excellent schools. With the exception of *upflow*, none of the indicators was insufficient for more than 10 excellent schools. If the share of insufficiently performing excellent schools is juxtaposed to the share of insufficiently performing non-excellent schools, we find that in all instances excellent schools score (much) better. Whereas 7.6% of the excellent schools fail to reach the *upflow*-target, this amounts to 10.9% for non-excellent schools. Similarly, respectively 4%, 0.8%, 1.7% and 0.4% of the excellent schools fail to reach their *lower flow-through*, *upper flow-through*, *mean final exam score* and *mean difference between final and school exam score*-targets, against 9.6%, 10.2%, 8.5% and 1.5% of non-excellent schools.

Another statistic which indicates excellent schools are high-performing concerns the required arrangement of supervision determined by the Inspectorate. In both PE and SE, excellent schools score better on supervision, which implies that excellent schools tend to be allocated a baseline arrangement of supervision relatively more often than non-excellent schools. In Appendix IX, you can find the frequency distribution for this variable. As expected, none of the primary schools and only two of the secondary schools that have received the label of excellence out of the entire pool of excellent schools throughout the years were classified under a different supervisory arrangement[49]; these two exceptions are most likely coding errors.

Occasionally, a variable is marked *DV*. This implies that these variables will be used as dependent variables. Either as a robustness check to the main regression or to check an interesting hypothesis related to the receipt of the label of excellence and the real effects it may induce.

Table 3: *Comparison of background characteristics of excellent and non-excellent schools in SE*

Variable	**Mean difference (N.ex – Ex); N=15.302**	**T-statistic**	**Description of variable**
Delta mean exam scores (DV)	0.0416***[50]	(3.89)	Exam scores at *t-1* minus exam scores at *t*
Supervisory arrangement	-0.0583***	(-11.82)	Ordinal variable; a high score implies you require little supervision as educational results are decent
Level of education[51]	-0.204**	(-2.69)	Ordinal variable; on average Ex. schools have more smarter students than N.ex schools
Upflow lower SE	-3.588***	(-5.11)	% of students in the lower half of SE which reach a level of education above advice in PE (i.e. upflow)
Flow-through lower SE	-0.640***	(-4.41)	% of students which complete lower half of SE within designated time
Flow-through N students	-82.96***	(-7.61)	Number of students monitored to calculate 'upflow' in upper half of SE

[48] 'Educational results' as described here form one of the pillars on which the Inspectorate's level of supervision is based.

[49] In the sidebar on page 8, a short description of the Inspectorate's examination and supervisory scheme can be found.

[50] * p<0.05, ** p<0.01 & *** p<0.001

[51] Level of education is fixed over time, therefore in a fixed effects regression it is automatically factored out. Here it is merely included to illustrate that on average excellent schools consist of relatively more smart students than non-excellent schools.

Flow-through upper SE	-3.200***	(-11.26)	% of students which complete upper-half of SE within designated time
Final exam N students	-18.00***	(-7.04)	Number of students taking final exam
Final exam learning disorder	-1.444	(-1.65)	Number of students taking final exam with learning disorders
Final exam apcg	-2.128*	(-2.35)	Number of students taking exam from apcg
Final exam N courses	-146.6***	(-7.41)	Total number of subjects taken by students at final exam
Mean score on the 'school exam'	-0.0850***	(-7.40)	As variable name
Mean difference between 'school exam' and final exam	0.0414**	(3.28)	As variable name
Total number of subjects taken by students at final and school exam	-146.7***	(-7.41)	As variable name
Total number of students by BRINVEST and level of education	-210.6***	(-7.02)	As variable name
Influx at year 1 (DV)	-39.51***	(-6.12)	Inflow into year 1, pooled by BRINVEST
Total number of students at BRINVEST by level of education (DV)	-96.18***	(-8.54)	As variable name

Section 3: Methodology

Identification of the causal effect of the label of excellence on objective student performance as measured by a final exam is not straightforward. Two issues stand out: (1) the channel through which the label may have an effect on student outcomes is unclear and (2) the conferral of the label of excellence could be endogenous to the specified model. In what follows, I will describe my main model and explain how I will try to address the issues noted above.

Firstly, let us consider a linear regression model that can be estimated through OLS[52]:

$$(1) \qquad \Gamma_{i,l,t} = \alpha + \beta \mathrm{E}_{i,l,t} + \delta \chi_{i,l,t} + \varepsilon_{i,l,t},$$

where $\Gamma_{i,l,t}$ represents the final exam grade at BRINVEST *i*, level of education *l* and year *t*. E represents the treatment, i.e. the school's excellence status. Furthermore, $\chi_{i,l,t}$ is a vector of observable BRINVEST and level of education-specific characteristics and $\varepsilon_{i,l,t}$ is the error term. The parameter we aim to estimate, β, is likely to be biased in this particular model specification, because unobservable characteristics of excellent schools are potentially correlated with the treatment and final exam grades. This implies that Cov($\mathrm{E}_{i,l,t}$, $\varepsilon_{i,l,t}$) ≠ 0, which would render our estimates inconsistent. This issue is particularly pressing here, as excellent schools are not uniquely identified through a specific rule or criterion. Although the judges at the Inspectorate who review a school's application for the label of

[52] The models in this section are equally applicable to PE and SE schools, unless stated otherwise.

excellence tick off a list of requirements regarding school results, a profile of excellence and much more, it is not an entirely transparent procedure. Therefore, it is very likely that excellent schools differ on a range of aspects from non-excellent schools which cannot be incorporated into the regression model and therefore confound the analysis.

As the model above fails to deliver a causal interpretation of β, a different approach ought to be considered. This is where the panel data nature of exam scores by BRINVEST and level of excellence for multiple years – data is available at least 1 year prior the receipt of the label of excellence – provides a feasible opportunity to isolate the effect we are interested in. Many features of schools (classified by BRINVEST and level of education in the dataset) are time-invariant. A school's level of education, location, teachers, didactic method, denomination, size and administration hardly ever change. Likewise, other characteristics of schools, such as average socio-economic background of the student population and the proportion of students from migratory descent, do change annually, albeit it slowly and marginally. Regarding the time-invariant variables, their fixed effect can be factored out through either first differencing or a so-called fixed effects (FE) regression. Although both approaches are broadly similar (and should yield roughly similar results), I have gone for the latter, so as to exploit all possible variation in the data. If one were to take first differences (FD), a year of valuable variation is lost. Nevertheless, according to Wooldridge (2002) FD is preferred over FE if serial correlation is expected. Since I have reason to believe that the residuals are indeed correlated over time (due to unobserved BRINVEST or level of education-specific shocks that linger), first differences will be taken anyhow as a robustness check. Furthermore, FD enables me to disentangle the effect on exam scores of 'becoming excellent'[53] and 'dropping back to non-excellent'.

In a fixed effects model, the analysis is restricted to exploiting within-panel variation, rather than differences between panels (i.e. between-panel variation). Therefore, the previous observation in a panel is essentially a control for the next time period. Having got rid of fixed effects, the inclusion of time-varying, policy-correlated control variables remains necessary however, for the same reasons as outlined above. Furthermore, I adjust for the time trend in exam grades (common to both non-excellent and excellent schools) through year fixed effects. The main fixed effects model I have estimated is denoted by:

$$(2) \qquad \Gamma_{i,l,t} = \alpha + \beta \mathrm{E}_{i,l,t} + \delta \chi_{i,l,t} + \psi_{i,l} + \tau_t + \lambda_l \tau_t + \varepsilon_{i,l,t},$$

where $\Gamma_{i,l,t}$ represents the final exam grade at BRINVEST *i*, level of education *l* and year *t*. E represents the treatment, i.e. the school's excellence status. Furthermore, $\chi_{i,l,t}$ is a vector of observable BRINVEST and level of education-specific characteristics. $\psi_{i,l}$ indicates panel fixed effects and τ_t year fixed effects. τ_t is also interacted with λ_l, which denotes level of education, in order to additionally capture year*level of education fixed effects. As above, $\varepsilon_{i,l,t}$ represents the error term.

The inclusion of the year*level of education fixed effects must be understood as exam fixed effects, whereas the year fixed effects aim to capture cohort fixed effects. Exams differ on a yearly basis by level of education, which implies that from one year to the next, the relative level of exams continuously changes. Through the inclusion of the interacted fixed effects, I aim to avoid that the compositional differences in terms of educational levels by excellence status will bias my estimates through a change in the relative difficulty of exams. General time dummies are not omitted from the

[53] In the data excerpts I denote these status changes as N.ex > Ex. and Ex > N.ex.

analysis, however, because it is very well conceivable that a particular cohort of students (independent of their level of education) in a certain year performs better than any of the preceding and/or succeeding cohorts.

The parameter of interest is β, which measures the average effect of the label of excellence on exam scores, disregarding the level of education. Whereas I have no evidence to suspect that the effect is level-dependent, VWO students (and schools in general) could react differently to the receipt of a label of education than VMBO students. Hence, in order to test whether the policy interacts with a school's level of education, two robustness checks will be run: one in which separate regressions are run by level of education and one in which λ_t interacts with E in the model above.

An attractive feature of Equation 2 is that it is a reduced-form model. To put it differently, β captures the effect of the policy on exam scores, irrespective of the particular pathway through which this (indirect) effect may be mediated. Upon receipt of the label of excellence (and even before that), all sorts of processes are set in motion which may or may not influence a student's exam score. However, as long as the changes in these unobserved variables[54] solely influence the exam scores of students *after* the label has been conferred, the causal chain still runs from the excellency label to the grade obtained on the final exam. As a matter of fact, model 2 is expected to resolve (or better: bypass) issue number 1. Nevertheless, β will be inconsistently estimated in model 2 if the treatment turns out to be endogenous.

Why would this be the case? This has everything to do with the application procedure. As explained in the previous section, schools file an application themselves approximately a year before the label is, if at all, ultimately conferred (see Appendix I). A school which wants to obtain the label knows well up front (i.e. before the actual application) that its school results must be outstanding, its profile of excellence clearly established and its overall performance strong. Therefore, out of strategic considerations, schools may wait a year or two before they apply to the programme. In the meantime, they put in plenty of time and effort to increase their chances of becoming excellent, which could also be reflected in higher grades for students *ex ante* the conferral of the label of excellence[55]. If this happens, the policy casts its shadow forward (i.e. there is a leading effect), which obfuscates *ex post* results. In a similar vein, it is also a theoretical possibility that an effect of the label only materializes one or two years after the label of excellence has been assigned (i.e. there is a lagged effect). In the year the Inspectorate reviews the school's application for excellence, cumbersome administrative requirements may require time otherwise spent on improving student performance. Hence, no significant change in exam grades is found in April/May, despite receiving the label of excellence in January. Instead, the effect is only picked up a year later, once the bureaucratic pressure has eased off and the school as well as the students are ready to reap the benefits from the label of excellence.

[54] My hypothesis is non-directional. Participation in 'Excellent Schools' could have a positive, negative or no effect on exam grades. This depends on the channel through which the effect, if any, is passed on. A (non-exhaustive) enumeration of a few possible positive mediators: encouraging atmosphere to perform, motivated stakeholders, attention for minute detail, weaknesses exposed by the jury are addressed and a thriving overall educational 'climate'. Furthermore, one can think of the following factors that may have an adverse effect on exam grades: bureaucratic paperwork which consumes time otherwise spent on teaching, teachers become extremely critical in grading exams and stakeholders may slack off once the label has been obtained for three years. If no effect is found, it is possible that these opposing mediating effects simply cancel each other out.

[55] Although it can never be stated with certainty that in the absence of these efforts, the label of excellence would not have been awarded, the increased effort put in by the school could render the treatment endogenous to the model. If higher grades also play a part in obtaining the label of excellence, there is even an element of reverse causality here.

Presumably, leads and lags can even occur simultaneously. Therefore, model 3 takes into account these hypothetical mechanisms:

$$(3) \qquad \Gamma_{i,l,t} = \alpha + \sum_{m=1}^{m} \beta_{-m} \mathrm{E}_{i,l,t-m} + \sum_{k=0}^{k} \beta_k \mathrm{E}_{i,l,t+k} + \delta \chi_{i,l,t} + \psi_{i,l} + \tau_t + \lambda_l \tau_t + \varepsilon_{i,l,t},$$

where *m* indicates the number of lags and *k* the number of leads; all other symbols and associated interpretations are identical to model 2.

As stated previously, I also explore a few other avenues through which the label of excellence may have (an indirect) effect on student outcomes. To test the prediction that sorting towards a school may occur in the wake of becoming 'excellent', I will include annual inflow in year 1 as a dependent variable in Equation 2[56], albeit with a different set of controls. Furthermore, I will run additional regressions with inflow as a DV on a sub-sample of schools located in the same area (to control for regional shocks to inflow correlated with receipt of the label of excellence) as well as take first differences. Moreover, I will also follow up the main regression using school exam scores as a DV rather than final exam scores. If the initial findings are robust, I would expect to identify a similar but smaller effect of the label of excellence[57]. In addition, I will present estimates of the main regression exclusively conducted on a sub-sample of schools with equal 3 or 4-digit ZIP codes as the excellent schools in my dataset, in order to pick up regional trends in exam scores which are indiscernible at the national level.

Lastly, my standard errors are at all times corrected for intra-panel correlation (in order to render the standard errors robust to serial correlation). Regarding high-schools, I have also explored the possibility that errors are correlated across higher-level clusters, for instance by BRIN or BRINVEST. However, in most instances, the results are nearly identical (and the p-values barely affected), therefore only the tables with standard errors clustered by panel will be reported upon. Since in most cases a school consists of multiple levels of education and only one of these levels is excellent, it also seems more sensible to cluster at the Panel-ID rather than BRIN(VEST)-level.

Section 4: Results

4.1. Main regression

Column 5 (3) in Table 4.1a (b) reports the point estimates and confidence intervals from fitting Equation 2 to the data on SE (PE). This is my preferred estimation, because it is all-encompassing in terms of the inclusion of covariates and fixed effects. Nevertheless, columns 1 to 4 report the estimates from slightly different models fitted to the data, in order to check if the estimates respond to excluding particular fixed effects and controls. As in the majority of regressions below, I have done this so as to check if the beta coefficients would bounce around under different specifications, which could be a sign that endogeneity is an issue. Furthermore, the inclusion of time-varying covariates is, on the one

[56] The parameter of interest is best obtained for primary schools. Information on inflow into SE is imprecise, because the data on inflow is either pooled by BRINVEST (and not sub-divided by level of education) or aggregated by all years across level of education.

[57] The score obtained at the school exam is a weighted average of tests and practical exams over three (for VWO) or two (for HAVO and VMBO) years. On the one hand, I therefore expect that if the label of excellence has an effect ex post, it is more difficult to detect, because the scores obtained as part of the school exam in the last year of high-school make up only a part of the total weighed score on the school exam (except if the label is kept over a longer time period). On the other hand, manipulation of school exam grades is easier, which would imply that if grade manipulation is correlated with the treatment a larger effect is to be found. As I expect both of these forces to work in opposite directions, I predict a similar (in terms of the sign) but smaller effect of the label of excellence on school exam scores.

hand, likely to reduce any bias, which is favourable. On the other hand, covariate data is only available from 2013 (or even 2014 in PE) up to and including 2017, which implies that some variation in the data is lost, thereby inflating standard errors and reducing statistical power. As schools have become excellent from 2013 onwards, the inclusion of covariates implies that 'Excellent Schools 2013' which have kept the label across time do not display any variance in their treatment status and are thus excluded from the analysis[58]. Although the total sample of excellent schools remains sufficiently large, the estimates will either way be less consistent.

Interestingly, the estimates in Table 4.1a & 4.1b seem to point at a negative effect of the treatment, i.e. receipt of the label of excellence, on exam scores. Although the estimates are insignificant at the 5% level if the controls are omitted from Equation 2, the coefficients do point in the same (negative) direction as when the full model is applied to the data. Although an estimate of -0.0615 ($p<0.01$) for high-schools may seem relatively small, it is equivalent to -0.24SDs on the final nationwide exams. Moreover, it is sufficiently large to conclude that a fraction of students will have failed the final nationwide exams as a result of the school's receipt of the label of excellence, regardless of the underlying mechanisms that have caused the decline in performance. Remarkably, the estimated effect in PE is nearly identical. Excellent primary schools score almost 1 point less on the Cito test, which equates -0.22SDs.

Table 4.1a

DV: average score on final nationwide exams

VARIABLES	(1)	(2)	(3)	(4)	(5)
Excellent	-0.0262*[59]	0.0314	-0.0205	-0.0241	-0.0615***
	(-0.0527 - 0.000310)	(-0.0539 - 0.117)	(-0.0465 - 0.00556)	(-0.0981 - 0.0499)	(-0.0918 - -0.0312)
Excellent#vmbo b		0		0	
		(0 – 0)		(0 – 0)	
Excellent#vmbo k		-0.0482		-0.0442	
		(-0.165 - 0.0689)		(-0.142 - 0.0540)	
Excellent#vmbo g(t)		-0.0389		-0.0349	
		(-0.139 - 0.0609)		(-0.123 - 0.0533)	
Excellent#havo		-0.114**		-0.0698	
		(-0.208 - -0.0192)		(-0.154 - 0.0141)	
Excellent#vwo		-0.0714		-0.0263	
		(-0.166 - 0.0228)		(-0.122 - 0.0697)	
Constant	6.319***	6.319***	6.319***	5.543***	5.543***
	(6.310 - 6.327)	(6.310 - 6.327)	(6.311 - 6.327)	(5.420 - 5.667)	(5.401 - 5.686)
Observations	27,239	27,239	27,239	13,196	13,196
R-squared	0.185	0.185	0.237	0.241	0.241
Number of PanelID	3,365	3,365	3,365	2,867	2,867
Group FE	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES
Year*Education FE	NO	NO	YES	YES	YES
Controls	NO	NO	NO	YES	YES

Table 4.1b

[58] In total 17 out of 130 SE panels which display variation in their status of excellence will be excluded from the analysis due to the inclusion of covariates. Concerning PE, 24 out of 78 schools are excluded.

[59] *** p<0.01, ** p<0.05, * p<0.1

DV: average score on Cito			
VARIABLES	(1) Model 1 – 6 years of data	(2) Model 2 – 6 years of data	(3) Model 3 – 4 years of data
Excellent	-0.299	-0.196	-0.914**
	(0.375)	(0.372)	(0.442)
Total amount of students taking Cito			-0.0161***
			(0.00422)
Impact area			-0.0840
			(0.497)
School weight by BRINVEST			-0.0142**
			(0.00565)
Total amount of students at BRINVEST			-0.00167
			(0.00118)
Constant	534.8***	535.2***	535.6***
	(0.00174)	(0.0367)	(0.322)
Observations	31,042	31,042	19,461
R-squared	0.000	0.010	0.012
Number of PanelID	6,415	6,415	5,986
Panel FE	YES	YES	YES
Time FE	NO	YES	YES

As it is unclear how exactly the label of excellence depresses exam scores in SE, it could very well be that the effect is dependent on a level-of-education-specific variable. Therefore, I interacted the excellence dummy with a level-of-education dummy. In column 4, none of the interactions significantly differ from the baseline dummy, VMBO-B. However, within-level of education samples sizes are rather small, especially if controls are included (which could possibly explain the insignificant findings). Nevertheless, it is reassuring that all coefficients are consistently negative, because this suggests that a common factor has caused the drop in grades. In Table 2 of Appendix XI, I ran separate regressions by level of education to find out if the effect of the label of excellence is consistently different from zero across the five distinct levels of education. With the exception of VMBO-B (p=0.544) and VWO (p=0.155), the estimates were all significantly different from zero. Yet again, this is testament to the fact that it is unlikely that a lurking level of education-specific variable is causing the decline in grades. Lastly, the covariates' beta coefficients of my preferred model (as reported upon in Table 4.1a, column 5) can be found in Table 1 of Appendix XI.

4.2. Leads and lags of the main regression

In the data section I already mentioned that schools which apply for the label of excellence have an incentive to perform exceptionally well in the years leading up to the application to increase their chances of obtaining the honorary title of 'excellent school'. Therefore, exam scores may receive a boost 1 or 2 years *ex ante* treatment is received. In the years after the label has been conferred, this incentive disappears (especially after the label's validity has gone up in 2016 from 1 to 3 years) and schools may be unable or unwilling to sustain their exceptionally high level of quality; the label of excellence has become a goal in itself. In order to verify whether such patterns are present in the data, Equation 2 has been complemented with leads and lags of the treatment to form Equation 3. The output from estimating Equation 3 through OLS using a variety of lead-lag combinations can be found in Table 4.2a (SE) and 4.2b (PE). In order to facilitate comparability to the preferred estimation of Model 2 above, I have run all regressions with a full set of controls and fixed effects.

One should immediately notice the striking similarity of the main coefficient in Table 4.2a. The dip in exam scores observed a few months after the label has been conferred persists, even if 2-year leads and lags are included. Moreover, the effect of the policy in the presence of different combinations of leads and lags is of the same magnitude as the estimated effect in Model 2. This is particularly noteworthy considering that (some) observations are dropped due to missing data if leads and lags of excellence are added to the model[60]. Leading up to the receipt of the label, there seems to be no clear increase or decrease in exam scores. Hence, no evidence is found for the conjecture that schools might influence the judges' decision on excellence through raising objective student performance. An attempt might still have been made, but in vain. Regarding lagged effects, a clearer pattern emerges. 1-year lags of excellence seem to suggest that one year after receipt of the label of excellence, part of the decline in exam scores in the previous year is offset. It could even be the case that the scores return to their original level the year after, but the 2-year lag estimate is insignificant across different combinations of lags and leads. Longer lags (or leads) could not be included, as this would severely hamper statistical inference. In Appendix XII, (some of) the regressions below are run without controls. Although this barely changes the coefficients on *'Excellent'*, they do suggest a marginally positive but significant leading effect, unlike in Table 4.2a. This can either mean that pre-Excellent Schools 2012 and 2013, leading effects *were* present or that the estimates are simply biased. Although it is theoretically possible that first and second year applicants were relatively more eager and motivated to obtain the label than applicants in later years (hence the observed leading effects), I am more inclined to attribute the significant leads to omitted variable bias (as these regressions were run without controls).

Concerning PE in Table 4.2b, the estimates for '*Excellent*' also lie in the range of the main estimate in Model 2. Whereas for secondary schools the lagging effect is more pronounced, primary schools display a stronger leading effect. This effect is approximately equal to the decline in exam grades the year after (+- 1 point on the Cito test), which would make the net effect zero. However, the lagging effect is also negative, albeit it insignificant if leads are added to the model. For that reason, it is hard to determine what the likely long-term effect of the policy is, especially since 2-year leads or lags cannot be included with controls as this reduces the sample size too much to draw valid conclusions regarding statistical significance.

Table 4.2a

DV: average score on final nationwide exams

VARIABLES	(1) L1	(2) L1&L2	(3) F1	(4) F1&F2	(5) L1&F1	(6) L1+L2&F1+F2	(7) L1+L2&F1	(8) L1&F1+F2
Excellent	-0.0667***	-0.0640***	-0.0693***	-0.0635***	-0.0698***	-0.0498**	-0.0671***	-0.0575***
	(0.0133)	(0.0135)	(0.0146)	(0.0225)	(0.0141)	(0.0219)	(0.0143)	(0.0217)
L.Excellent	0.0240*	0.0248*			0.0345**	0.0248	0.0295	0.0368
	(0.0136)	(0.0135)			(0.0173)	(0.0210)	(0.0179)	(0.0237)
L2.Excellent		-0.00253				0.0565**	0.0186	
		(0.0155)				(0.0286)	(0.0199)	
F.Excellent			0.00594	-0.0172	0.00729	-0.00329	0.0153	-0.0140
			(0.0159)	(0.0181)	(0.0161)	(0.0199)	(0.0168)	(0.0185)
F2.Excellent				0.0220		0.0315		0.0218
				(0.0208)		(0.0214)		(0.0212)
Constant	5.520***	5.554***	5.497***	5.508***	5.470***	5.497***	5.508***	5.463***
	(0.0609)	(0.0604)	(0.0667)	(0.0733)	(0.0652)	(0.0696)	(0.0662)	(0.0694)
Observations	13,079	12,847	10,492	7,825	10,427	7,528	10,195	7,760
R-squared	0.244	0.242	0.256	0.305	0.259	0.315	0.258	0.311
Number of PanelID	2,817	2,811	2,796	2,739	2,794	2,724	2,786	2,735
Panel FE	YES	YES	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES	YES	YES

[60] More specifically, observations are dropped for one and two-year leads. If a one-year lead is added, 2017 is dropped and if the two-year lead is added as well, 2016 is dropped too.

Time*Education FE	YES	YES	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES	YES	YES

Table 4.2b

DV: average score on Cito

VARIABLES	(1) Lag of excellence	(2) Lead of excellence	(3) Excellence & 1-year lag	(4) Excellence & 1-year lead	(5) Excellence and 1-year lag & lead	(6) Excellence and 1-year lag & lead
Excellent			-0.628	-1.344**	-1.186**	-0.389
			(0.439)	(0.586)	(0.598)	(0.492)
Lag of excellence	-1.003**		-0.857*		-0.185	-0.554
	(0.458)		(0.461)		(0.611)	(0.441)
Lead of excellence		0.900*		1.085**	1.066**	0.932**
		(0.468)		(0.470)	(0.475)	(0.381)
Constant	535.6***	535.5***	535.6***	535.5***	535.8***	534.8***
	(0.319)	(0.462)	(0.319)	(0.462)	(0.448)	(0.0363)
Observations	18,964	15,372	18,964	15,372	14,881	20,311
R-squared	0.012	0.011	0.012	0.012	0.012	0.007
Number of PanelID	5,883	5,956	5,883	5,956	5,821	6,139
Group FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	NO

The consistently negative point estimates in Table 4.2a of the effect of the label of excellence also largely rule out the possibility of reverse causality. It is theoretically possible that in the absence of the increased effort before conferral of the label of excellence, at the margin the label would not have been granted. In that case, the causality runs opposite to what we would expect and the policy would be endogenous to the model. However, in the presence of reverse causality you would expect either a significant leading effect (if the school improves itself before *t-1*) or positive point estimates (if the main improvements occur after *t-1* but before *t=0*). Since secondary schools do not seem to be affected by leads of the treatment or positive treatment effects, we can largely rule out reverse causality in SE. Regarding primary schools, this is different. The leads are quite strong, which does not entirely eliminate the possibility of reverse causality, at least for some schools[61]. Nevertheless, the point estimates in all specifications remain negative for primary schools too, which renders performance improvements after exam scores have been set at *t-1* unlikely. In column 2 of Table 4.2c, I furthermore check whether the leading effects are being carried over to the next year through the inclusion of a 1-year lagged dependent variable. This is important, otherwise the effect I measure may suffer from mean reversion as a source of endogeneity. I find strong evidence for persistence of the dependent variable amongst primary schools; sufficiently strong to even render the effect of the label of excellence insignificant at the 15%-level (p=0.158). No evidence for mean reversion is found amongst secondary schools and the coefficient in column 1 of Table 4.2c is in line with the coefficients in the tables above.

As a matter of fact, the negative relationship between the label of excellence and exam scores amongst secondary schools seems well-established. Regarding primary schools, we should be a little more careful in drawing strong conclusions, as the results in Tables 4.2b and c suggest that reverse causality or mean reversion could have biased the estimates in Table 4.1b. However, the

[61] The judges base their verdict on much more than exam scores, therefore it is still not very likely that reverse causality occurs, but it cannot be ruled out either, especially for schools 'at the margin' of becoming excellent.

preponderance of evidence still seems to point to a negative effect of the label of excellence on exam scores amongst primary schools (although it may be only modest).

Table 4.2c

DV: average score on final nationwide exams & Cito

	(1)	(2)
VARIABLES	SE - LDV	PE - LDV
Excellent	-0.0592***	-0.614
	(0.0128)	(0.435)
1-year lag of final exam	0.0123	
	(0.0105)	
1-year lag of Cito		-0.193***
		(0.00917)
Constant	5.417***	638.8***
	(0.0847)	(4.918)
Observations	12,740	18,387
R-squared	0.250	0.048
Number of PanelID	2,729	5,731
Panel FE	YES	YES
Time FE	YES	YES
Time*Education FE	YES	NO
Controls	YES	YES

4.3. Inflow as a DV

Under the assumption that parents regard the label of excellence as an indicator of school quality, I expect the inflow into year 1 to increase following the receipt of the label of excellence. Regioplan established back in 2015 and 2016 that the total number of students in excellent primary schools increased over time relative to non-excellent schools, whereas ambiguous results were found amongst secondary schools. Whereas it is unclear if the researchers have controlled for the total number of students that failed to graduate or had to redo an entire year (which they should have), I simply cannot, since data on these statistics is unavailable to me. This is not at all problematic, however, because I use data on annual student inflow into year 1 in PE, which is a more precise estimator as it directly measures the policy's impact. Unfortunately, similar information is unavailable for secondary schools, and I therefore had to resort to less precise accounts of year 1 inflow. Hence, the estimates for SE are likely to be lower bounds[62] (*Table 4.3, model 1-4*) or biased by time-varying unobservables I cannot control for[63] *(Table 4.3, model 5-6)*.

Columns 1-4 in Table 4.3a indicate that on average student inflow goes up by approximately 6.5 (0.06SDs) students in the wake of receiving a label of excellence. More precisely, schools which have become excellent experience a smaller decline or even a (counter-trend) increase in the inflow

[62] The dependent variable in models 1 to 4 in Table 4.3 is yearly inflow in year 1 by BRINVEST. Hence, I do not have information on the change in inflow by BRINVEST and level of education. Since most excellent panels (uniquely identifiable at the BRINVEST and level of education-level) are nested within BRINVESTS consisting of multiple levels of education, the effect of 'becoming excellent' on student inflow is averaged out across the total number of (non-excellent) panels within a unique BRINVEST. Since we have no reason to believe that non-excellent panels within the same BRINVEST benefit from spill-over effects, any (positive) estimate of the label of excellence on an excellent panel is likely to be a lower bound.

[63] Since I have no information on the number of students that leave school without having graduated or have had to redo an entire year, these estimates are likely to be slightly biased.

of new students. In terms of standard deviations, the relative increase is rather small, because the variation across both years and levels of education is large. It is even the case, as can be inferred from columns 2 and 4, that the effect of the label on inflow is moderated by a school's level of education. Whereas the insignificant estimate for VMBO-B is presumably negative, it is significantly positive for VWO-schools. Therefore, it seems as if the signalled relative increase in student influx is predominantly driven by excellent VMBO-G(T), HAVO and VWO schools. The findings in columns 1 to 4 are corroborated by the estimations in columns 5 and 6, in which the total student population by BRINVEST and level of education is regressed upon a school's status of excellence. On average, schools that become excellent expand by 10 students (0.07SDs) relatively to non-excellent schools. Whereas the fixed effects do capture the pre-treatment differences in size amongst excellent and non-excellent schools, it is unclear whether this increase is driven by a larger inflow, a smaller outflow or even both. Nevertheless, in conjunction with the findings in columns 1 to 4, it seems plausible to assume that the policy has induced at least a minor increase in student inflow, especially amongst higher levels of education.

Unlike excellent high-schools, excellent primary schools hardly seem to experience an increase in their yearly student inflow. Although models 1 and 3 in Table 4.3b suggest that either directly or 1 year after having received the label of excellence, excellent primary schools do expand slightly relatively to non-excellent schools, inclusion of only two controls (i.e. Cito-score & impact area) render the estimates insignificant. Although this might be due to spurious correlation, I do believe that changes in exam scores and impact areas inform parental decisions as to where their child should go to school. Viewed in that light, it is quite surprising that Regioplan (2016) found clear, unmistakably positive effects on the total number of students in excellent PE, whereas I do not find any substantial evidence for direct or lagged increases in year 1 inflow. Similarly, my (potentially imprecise) estimates for SE suggest clear positive effects of the label on the size of the student population, whereas Regioplan found ambiguous results. Inspired by Regioplan and in an attempt to reconcile these findings, I have constructed regional control groups of non-excellent schools based on the area ZIP-code in which an excellent school is located. I report on the results of these regressions in section 5 (see Table 5.3b & Appendix XIV). If the treatment is somehow correlated with a region-specific trend, this could have confounded my estimates. However, it is also plausible that differences in the composition of the sample have caused the observed discrepancy in results. For instance, Regioplan's regressions contained data up until 2015, whereas my analysis comprises two more years.

Table 4.3a

Models 1-4: DV is yearly inflow in year 1 into SE by BRINVEST
Models 5-6: DV is total amount of students in SE by BRINVEST & level of education

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES						
Excellent	6.485**	-5.594	6.541**	-5.127	10.40**	0.668
	(2.842)	(8.057)	(2.855)	(8.259)	(4.143)	(5.841)
Excellent#vmbo k		-0.205		-0.558		5.893
		(10.74)		(11.03)		(7.927)
Excellent#vmbo g(t)		13.35		12.85		2.228
		(9.300)		(9.518)		(8.228)
Excellent#havo		16.26		15.71		10.48
		(10.70)		(10.96)		(12.91)
Excellent#vwo		21.44**		21.09**		27.41**
		(9.283)		(9.568)		(13.15)
Constant	200.6***	200.6***	200.6***	200.6***	186.1***	186.1***
	(0.893)	(0.893)	(0.899)	(0.899)	(0.720)	(0.720)

Observations	16,633	16,633	16,633	16,633	16,674	16,674
R-squared	0.033	0.033	0.034	0.035	0.034	0.035
Number of PanelID	3,000	3,000	3,000	3,000	3,022	3,022
Panel FE	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Time*Education FE	NO	NO	YES	YES	YES	YES
Controls	NO	NO	NO	NO	NO	NO

Table 4.3b

DV: Inflow into PE by year 1

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(7) Model 5	(6) Model 6
Excellent	2.344**	1.608	0.357	1.461	2.659	1.814
	(1.053)	(1.992)	(1.277)	(1.961)	(1.692)	(1.992)
1 year lag of excellent			4.470***	-0.278	-0.114	-0.483
			(1.248)	(1.169)	(1.096)	(1.120)
2 year lag of excellent					1.317	1.935*
					(1.087)	(1.143)
Constant	23.84***	38.44***	23.60***	40.13***	23.51***	40.62***
	(0.0988)	(9.429)	(0.0915)	(9.683)	(0.0889)	(9.850)
Observations	36,579	19,466	29,624	18,969	22,770	18,424
R-squared	0.006	0.007	0.007	0.008	0.007	0.007
Number of PanelID	6,799	5,986	6,756	5,883	6,611	5,781
Group FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Controls	NO	YES	NO	YES	NO	YES

Section 5: Robustness checks

5.1. First differences

As indicated previously, the presence of serially correlated errors is to be expected. A myriad of variables determines one's final exam score but remain unobserved. Some of these variables will also vary over time and are therefore not eliminated in a fixed effects regression. Successive cohorts of students at the same school and within the same level of education (i.e. within a particular panel in my dataset) may be exposed to certain common developments, incidents and events which influence their final exam performance. These 'shocks' may persist over a certain period of time and then fade out. Typically, a first difference (FD) regression is better equipped to deal with serially correlated errors, because any resulting error differences will be uncorrelated, which improves the efficiency of the estimation of the parameter of interest (Wooldridge, 2002). The downside to using first differences as an estimation technique is that with unbalanced panel data (which I have), two observations are dropped for every first difference which is taken on missing data. Hence, some variation in the data is lost. In Table 5.1a and 5.1b, you can find the output from re-estimating Equation 2 using FD on schools in SE. As expected with a more efficient estimation method, the standard errors are smaller in all specifications. Moreover, the coefficients are largely similar to the FE-values reported in Table 4.1a, column 5 and Table 4.2a, column 1,3 and 5. The minor differences between the FE and FD coefficients is most likely attributable to sampling error. Concerning PE, the estimates in Table 5.1b, columns 1 and 2 are comparable to the FE-estimates in Table 4.2a, columns 2 and 3. However, the SEs are not smaller, because due to the nature of the data many observations were dropped taking FD. For similar reasons, lags or leads have not been added to the model, as this would imply practically no variation in the status of excellence of schools would have been left.

Columns 5 and 4 of respectively Table 5.1a and 5.1b and graphs 3a, 3b and 3c (separate graphs by level of education are available in Appendix XIII) shed light on another interesting phenomenon. In addition to exploring the effect of 'becoming excellent' (i.e. N.ex > Ex), I also looked into the disparate impact of losing one's status of excellence (i.e. Ex > N.ex). Remarkably, the effect on exam scores of 'abandoning excellence' is almost perfectly opposite (although not significant within PE). Since schools which have lost the label of excellence consist both of schools which did not reapply for the label or schools which failed to meet the requirements for excellence in the ensuing year, it is hard to interpret the significance of this finding. However, if, one the one hand, most schools who dropped back to non-excellence were schools who decided not to reapply, it would suggest that the application procedure leading up to the receipt of the label puts downward pressure on exam grades. On the other hand, if most of the previously excellent schools have reapplied, but were simply deemed ineligible by the judges (to keep the label), it would suggest that it is anything *but* the application procedure that has had an adverse effect on exam grades, otherwise we should have observed a negative effect on exam grades for the *Ex > N.ex*-schools as well. As it is beyond the scope of my inquiry, it is up to future research to investigate the exact underlying mechanism which induces exam grades to decline after a school has turned 'Excellent' (and vice versa).

Table 5.1a

DV: change in average score on final nationwide exams

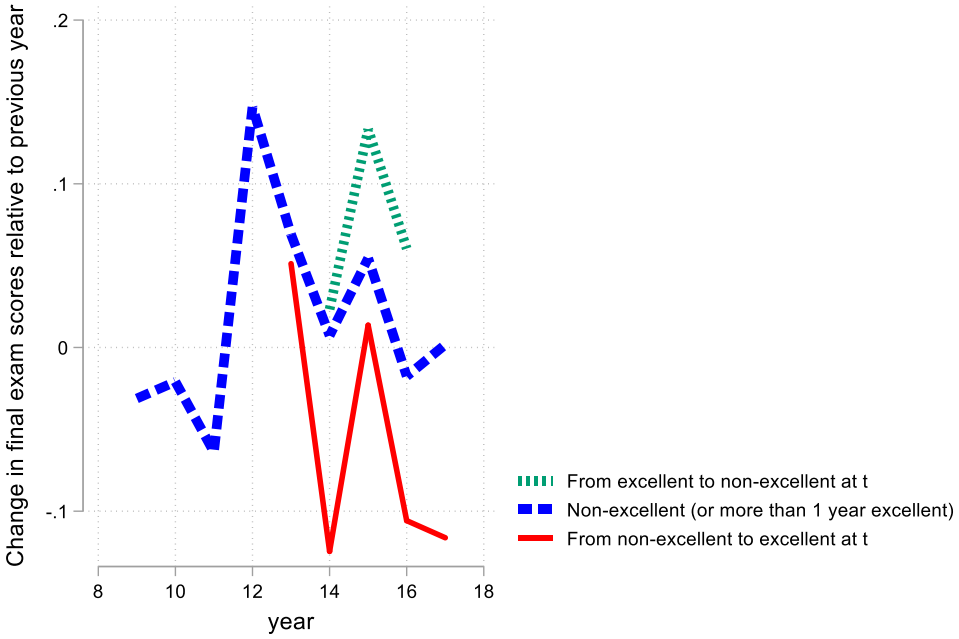
VARIABLES	(1)	(2)	(3)	(4)	(5)
D.Excellent	-0.0736***	-0.0664***	-0.0674***	-0.0584***	
	(0.0138)	(0.0134)	(0.0159)	(0.0152)	
1-year lag of D.Excellent		0.0245*		0.0242	
		(0.0137)		(0.0177)	
1-year lead of D.Excellent			-0.00993	-0.00744	
			(0.0180)	(0.0181)	
N.ex > Ex					-0.0756***
					(0.0159)
Ex > N.ex					0.0677**
					(0.0287)
Constant	0.0407***	0.0421***	0.0397***	0.0411***	0.0407***
	(0.0108)	(0.0109)	(0.0108)	(0.0108)	(0.0108)
Observations	10,249	10,190	7,644	7,585	10,249
R-squared	0.203	0.205	0.228	0.231	0.203
Time FE	YES	YES	YES	YES	YES
Time*Education FE	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES

Table 5.1b

DV: change in average score on Cito

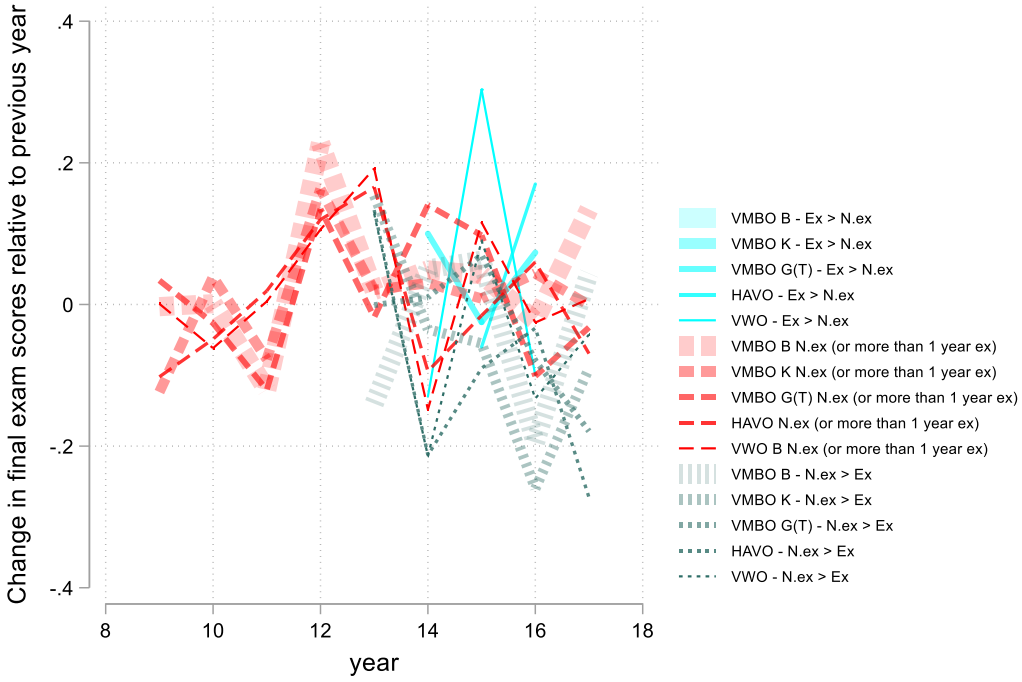
VARIABLES	(1)	(2)	(3)	(4)
N.ex > Ex			-0.418	-1.220**
			(0.430)	(0.511)
Ex > N.ex			0.00548	0.886
			(0.865)	(1.313)
D.Excellent	-0.306	-1.113**		
	(0.392)	(0.545)		
Constant	-0.464***	0.494***	-0.464***	0.494***
	(0.0540)	(0.0611)	(0.0540)	(0.0611)

Observations	24,159	13,272	24,159	13,272
R-squared	0.012	0.015	0.012	0.015
Time FE	YES	YES	YES	YES
Controls	NO	YES	NO	YES

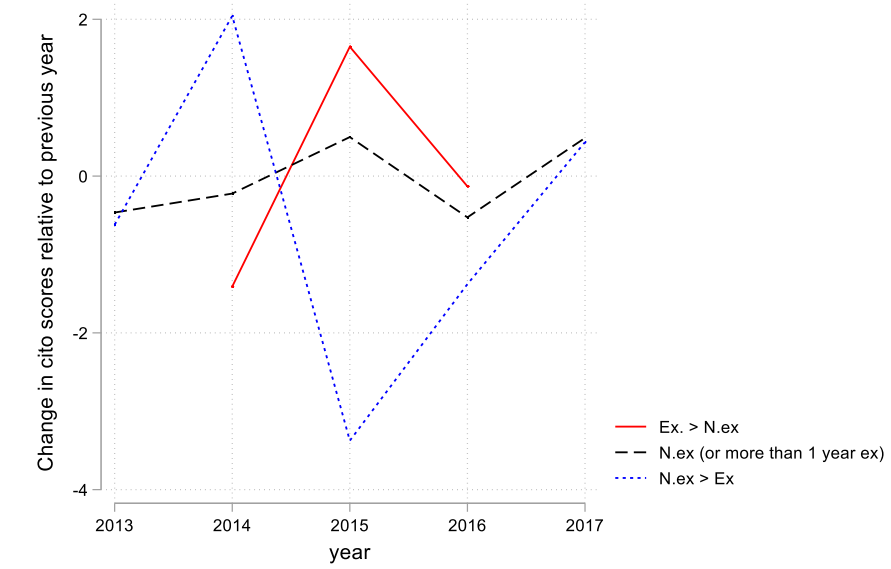


Graph 3a

Graph 3b



Graph 3c



First-differenced equations have also been estimated with the annual inflow of students as a dependent variable. Unlike the FD-regressions above, the results in Table 5.2a do not substantiate the conclusions drawn from the FE-regressions in the *results*-section. Concerning schools in SE, the point estimates for the first difference of 'excellence' are positive but insignificant and much smaller than in the FE estimation above. The FD coefficients in Table 5.2b for primary schools are more or less equal to the associated FE coefficients[64]. Even though the results hint at a lagged effect of excellence, the coefficients bump around too much to draw definite conclusions. Although part of the discrepancy in parameter estimates between FE and FD models can be attributed to sampling variance, it is also a possible signal of model endogeneity. Despite these data issues and inferential limitations, the coefficients are at all times positive and more than exceed the downward trend in inflow as indicated by the 'constant' (with the exception of column 3 and 5 in Table 5.2b). Therefore, it is still highly likely that excellent schools benefit from the label of excellence in terms of annual inflow (as is also suggested by graph 4), however, the relative size of the effect cannot be determined with certainty. Lastly, in columns 6 and 7 in Table 5.2b, I split up the FD estimation by 'receiving' and 'losing' the label of excellence. Unlike Cito exam scores, inflow is not adversely affected if a school's status of excellence is lost, which could be an indication that the parental perception of a previously excellent school's reputation remains high up until at least a little over a year after the label is withdrawn.

Table 5.2a

Model 1: DV is change in yearly inflow in year 1 into SE by BRINVEST
Model 2: DV is change in total amount of students in SE by BRINVEST & level of education

VARIABLES	(1)	(2)
D.Excellent	3.747	1.787
	(2.480)	(2.460)
Constant	-2.201	-1.125
	(2.259)	(0.936)
Observations	13,601	13,630

[64] Controls have not been included, as otherwise many of the panels were reduced to only 2 or 3 observations each.

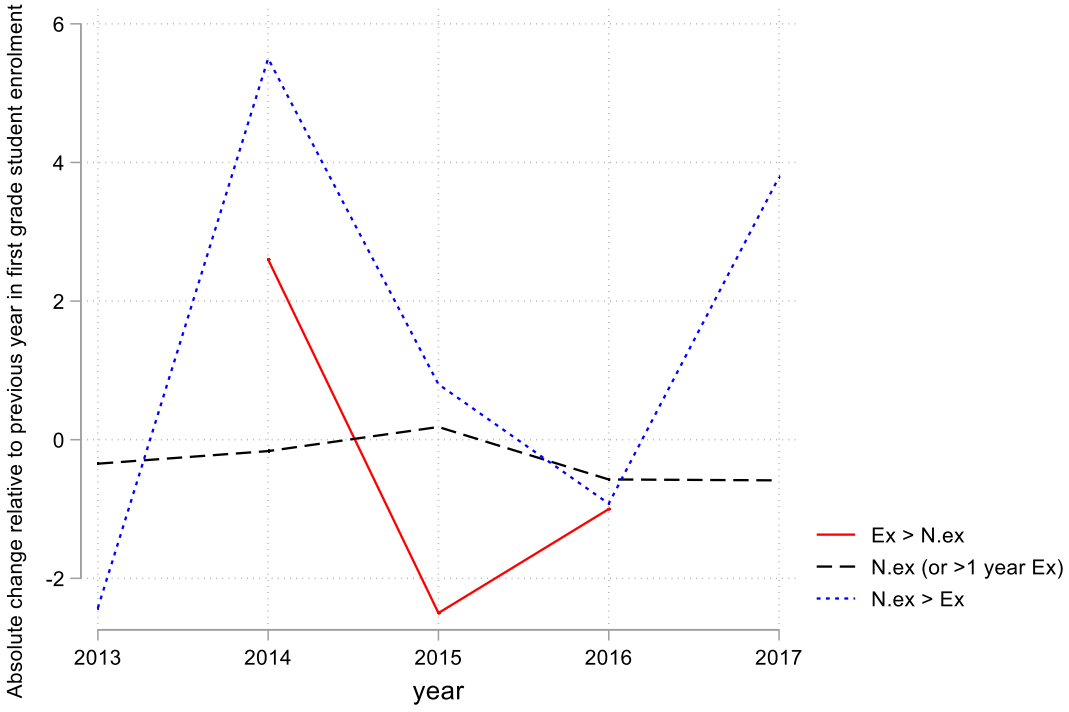
R-squared	0.006	0.017
Time FE	YES	YES
Time*Education FE	YES	YES
Controls	NO	NO

Table 5.2b

DV: Change in inflow into PE by year 1

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
D.Excellent	0.566			2.574*	2.572		
	(1.048)			(1.345)	(1.681)		
1-year lag of D.Excellent		3.786***		4.283***	2.073		
		(1.075)		(1.099)	(1.324)		
2-year lag of D.Excellent			0.750		1.647		
			(1.107)		(1.213)		
N.ex > Ex						0.798	
						(1.304)	
Ex > N.ex						0.0852	
						(1.584)	
1-year lag of N.ex > Ex							5.233***
							(1.292)
1-year lag of Ex > N.ex							-0.793
							(1.806)
Constant	-0.359***	-0.170	0.188	-0.171	0.186	-0.360***	-0.178
	(0.119)	(0.124)	(0.128)	(0.124)	(0.128)	(0.119)	(0.124)
Observations	29,624	22,770	16,159	22,770	16,159	29,624	22,770
R-squared	0.001	0.002	0.002	0.002	0.002	0.001	0.002
Time FE	YES	YES	YES	YES	YES	YES	YES
Controls	NO	NO	NO	NO	NO	NO	NO

Graph 4



Furthermore, I have also regressed the exam scores and year 1 inflow upon a dummy for excellence *and* a dummy for schools which are situated within the same 3 or 4-digit ZIP-code area of an excellent school. If certain trends in exam scores or annual inflow were region-specific (for instance due to changes in regional demographics and/or socio-economic background), these trends will emerge in the fixed effects regressions in Table 5.3a and b (for SE) and Appendix XIV (for PE). I have reproduced the main regressions from the *results*-section in column 1 to facilitate comparison.

The coefficients for the 'Excellent within ZIP-dummies' are nearly identical to the main regression, because the control group is composed of exactly the same non-excellent schools with the exception of those schools (only a fraction of the total population of non-excellent schools) which share a 3 or 4-digit ZIP code with an excellent school. The number of observations within both groups can be found in Appendix X. Whereas the 3-digit control group is larger, which makes the estimates more consistent, it is less comparable to the group of excellent schools than the 4-digit control group because the displayed trends are unlikely to be fully similar.

The results are noteworthy. Despite the insignificance of the 'same 4-digit ZIP, non-excellence dummy', it seems safe to assume that secondary schools located within the same region as their excellent counterparts also experience (at least) a (minor) decline in exam scores. If one takes up the estimate in column 2, approximately 50% of the post-treatment decline in exam scores of excellent schools can be explained for by a regional trend, thereby reducing the overall effect of the policy on final nationwide exam scores to a little over -0.03 points[65]. Taking into account the 'same 4-digit ZIP, non-excellence dummy' and the lack of a clear theoretical explanation for regional trends in exam scores implies that a chance finding may not be ruled out, however, it is highly probable to say the least that the observed effect of the policy on exam scores is at least slightly offset by correlated region-specific unobservables which vary over time.

Amongst primary schools (Appendix XIV) such a regional (downward) trend in exam grades is absent. Furthermore, a regional trend in annual inflow of students into both PE and SE can neither be discerned. The estimates lie far apart and the confidence intervals span 0 by a large margin, which makes drawing strong conclusions regarding a regional trend impossible. Therefore, the estimates from the main regression are robust to different configurations of the control group, although the impact of the label of excellence on exam scores in secondary education is possibly an upper bound.

In order to test a hypothesis I formulated in the very beginning of this thesis, I subsequently interact the dummy for the label of excellence with the degree of competition[66] a school faces within the geographic bounds of its 3 and 4 digit ZIP-code areas. I reasoned that the change in exam scores or annual school inflow will potentially be moderated by the area's degree of competition, if a school obtains the label of excellence in order to signal its outstanding quality and distinctiveness vis-à-vis non-excellent schools in the same region. In scarcely populated areas, the label of excellence is unlikely to induce a large increase the influx of new students, simply because students and their parents trade off travel time with the school's performance. In general, the stakes (and therefore the benefits of the label) are lower for schools which are hardly affected by competition. Therefore, I argued that the

[65] The conditional mean difference (Table 5.3a, column 2) between excellent and non-excellent secondary schools within the same 3-digit ZIP area is -0.0314 final exam points, which is significant at the 1% level. Within the 4-digit ZIP area (Table 5.3a, column 3), the coefficient is -0.0308, which is also significant at the 1% level.

[66] I have recoded competition as a continuous variable into three categories (i.e. two dummies): low, medium and a high degree of competition. The cut-off points for the categories were set at 1/3rd (for medium) and 2/3rd (for high) of the maximum level of competition a school in the dataset faced. Schools are therefore not evenly distributed across the categories, as most schools face relatively little competition. The level of competition faced from 0 to 12 schools and 0 to 35 schools for respectively 4 and 3-digit ZIP codes.

presumed impact (either positive or negative) on exam grades was likely to be lower too for 'monopoly' schools. In Appendix XV, you can find that the empirical evidence does not unambiguously support my conjectures. Within 3-digit ZIP areas, schools in SE which become excellent in a particular year and face medium or high[67] levels of school competition, tend to perform worse than their excellent counterparts which operate in a more 'monopolistic' environment. However, this effect disappears if the analysis is restricted to the degree of competition within 4-digit areas. Regarding primary schools, no significant interaction effect is found within 3-digit areas of competition. However, a massive effect (-5.23 points on the Cito) is found if a school's receipt of the label of excellence is interacted with a high degree of competition in common 4-digit areas. Since the interaction between medium competition and excellence is non-significant and there is no reason to believe that the effect of the interaction is non-linear, I ascribe the significance of the finding above to sampling bias[68]. All in all, it cannot be said with any certainty that the presence of other, competing schools moderate the relationship between excellence and exam grades. Neither is the interaction of excellence with the degree of competition in PE a significant predictor of annual school inflow into year 1. This is quite surprising, but could be explained by three factors: (1) the real effect is lagged (and therefore the time window analysed here is too small), (2) there is no clear trade-off between the attractiveness of an excellent school and the distance towards it or (3) the sample size is simply too small to identify an effect (if there is actually a causal relationship). Either way, it does not seem to be the case that the degree of competition a high school or primary school faces is substantially altering the real effect of the label of excellence on outcomes related to student performance, whereas the disproportionately high concentration of labels in densely populated regions in the Netherlands does suggest this at first glance.

As a final robustness check, I also regressed the school exam scores on the label of excellence. I expected the effect to be much smaller, as the average score on the school exam is an average over either 2 or 3 year (depending on one's level of education). Therefore, only if the label of excellence would have strong leading effects on student performance, an effect was to be expected. In the *results*-section, we already saw that the leading effects are negligible, which is confirmed by the output from the school exam regressions (Appendix XVI): the label of excellence is not associated with school exam grades[69]. Although this finding does not rule out manipulation of school exams by teachers (i.e. 'gaming the system' as a kind of window-dressing), the impact has at least not been strong enough to offset any negative effects the label of excellence might have had on school exam grades, otherwise we should have observed positive coefficients.

To put it differently, if we assume that the prospect of a label of excellence motivates the school staff, improves didactic efficiency and/or encourages teacher cheating, we would expect to find at least a mildly significant positive effect on school exam scores. However, instead I find insignificant, but negative point estimates. The negative effect is hence more than offsetting the alleged positive impact of any or all of the aforementioned positive mechanisms. This finding reinforces the most plausible presumption that the application procedure and associated paperwork in order to obtain the

[67] The coefficient for the interaction between 'high' and 'excellent' is insignificant at the 10% level (p=0.291), but this is most likely due to the small sample size (N=9).

[68] The estimate is based on only 3 excellent schools which face high levels of competition.

[69] With the exception of HAVO, for which a significantly negative effect is found on school exam grades. This implies that students in HAVO-schools that become excellent in their final exam year perform significantly worse on their school exam in both (on average) HAVO 4 and 5. The cause of this HAVO-specific plummet in school exam scores is unclear.

label of excellence is the predominant factor which causes the decline in final exam grades, although it has to be stressed that this remains a speculative hunch.

Table 5.3a

DV: change in exam scores in year by excellence and ZIP

VARIABLES	(1) Main regression	(2) 3 digit ZIP	(3) 4 digit ZIP	(4) 3 digit ZIP by education	(5) 4 digit ZIP by education
Excellent	-0.0615***				
	(0.0130)				
N.ex, same 3-digit ZIP as ex		-0.0297*		-0.0449	
		(0.0153)		(0.0722)	
Excellent (with 3-digit ZIP)		-0.0630***		-0.0243	
		(0.0130)		(0.0377)	
N.ex_3digits#vmbo k				0.0146	
				(0.0907)	
N.ex_3digits#vmbo g(t)				-0.00771	
				(0.0758)	
N.ex_3digits#havo				0.0605	
				(0.0782)	
N.ex_3digits#vwo				0.00643	
				(0.0795)	
Ex_3digits#vmbo k				-0.0441	
				(0.0501)	
Ex_3digits#vmbo g(t)				-0.0385	
				(0.0450)	
Ex_3digits#havo				-0.0683	
				(0.0428)	
Ex_3digits#vwo				-0.0285	
				(0.0490)	
N.ex, same 4-digit ZIP as ex			-0.0282		0.00796
			(0.0303)		(0.0866)
Excellent (with 4-digit ZIP)			-0.0618***		-0.0241
			(0.0130)		(0.0377)
N.ex_4digits#vmbo k					0.00880
					(0.124)
N.ex_4digits#vmbo g(t)					-0.118
					(0.0996)
N.ex_4digits#havo					0.0252
					(0.0990)
N.ex_4digits#vwo					-0.0421
					(0.116)
Ex_4digits#vmbo k					-0.0440
					(0.0501)
Ex_4digits#vmbo g(t)					-0.0352
					(0.0450)
Ex_4digits#havo					-0.0696
					(0.0428)
Ex_4digits#vwo					-0.0278
					(0.0492)
Constant	5.543***	5.541***	5.543***	5.543***	5.543***
	(0.0630)	(0.0628)	(0.0629)	(0.0628)	(0.0629)
Observations	13,196	13,196	13,196	13,196	13,196
R-squared	0.241	0.241	0.241	0.242	0.242
Number of PanelID	2,867	2,867	2,867	2,867	2,867
Panel FE	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES
Time*Education FE	YES	YES	YES	YES	YES

Controls	YES	YES	YES	YES	YES

Table 5.3b

DV: change in yearly inflow in year 1 into SE by BRINVEST and ZIP

	(1)	(2)	(3)
VARIABLES			
Excellent	6.541**		
	(0.942 - 12.14)		
N.ex, same 4-digit ZIP as ex		-1.719	
		(-15.07 - 11.64)	
Ex, 4 digits		6.520**	
		(0.920 - 12.12)	
N.ex, same 3-digit ZIP as ex			4.136
			(-3.056 - 11.33)
N.ex, 3 digits			6.733**
			(1.142 - 12.32)
Constant	200.6***	200.6***	200.6***
	(198.9 - 202.4)	(198.9 - 202.4)	(198.9 - 202.4)
Observations	16,633	16,633	16,633
R-squared	0.034	0.034	0.035
Number of PanelID	3,000	3,000	3,000
Time FE	YES	YES	YES
Time*Education FE	YES	YES	YES
Controls	NO	NO	NO

Section 6: Discussion

In the beginning of this paper, I have raised the question in what way student outcomes are affected by the label of excellence. As we have seen, conferral of the label is clearly negatively associated with objective and verifiable student performance as measured by test scores. In particular students in high-schools seem to perform worse in the year after their school has become excellent. The effect of the label of excellence on exam grades is -0.0615*** grade points, which is equivalent to -0.24SDs. Even the smallest point estimate I found (i.e. controlling for regional time trends) is still well below zero at -0.0314 grade points (or -0.12SDs). Although these numbers may seem small, you should take into account that these values indicate average changes across as many as up to nine different subjects (for VWO) across an entire panel, which on average consists of 87 students who take the final exam. The passing grade at the final exam is 5.5. Therefore, it is quite plausible that a few students would have successfully graduated if the policy had not been implemented, because the decline in exam grades associated with the label of excellence will lower their grades below the required threshold value to pass the final exam. These students (usually) will have to redo the entire year. This comes at (an economic) cost. Throughout the five years in my sample, 136 uniquely identifiable secondary schools have received the label of excellence. If we assume that for every school at least 1 student fails the final exam due to the policy[70], this implies that 136 students will have had to stay in school for another year. The CPB Netherlands Bureau for Economic Policy Analysis has calculated that the direct educational costs of repeating a year in high-school amount to approximately €7500 (van Vuuren & van der Wiel, 2015). Hence, the total educational costs (up until May 2017) of additional student retention due to the policy will lie in the range of 1 million euros. This is a lower-bound: non-educational costs have not been included in the estimation, such as a delayed career, a lower salary and less tax revenue, additional costs of studying payable by the student and the socio-psychological harm associated with retention. Moreover, for the core courses (Dutch, English & Math) a student may

[70] Equivalent to an increase in the retention rate of 15%.

not score more than one five. If the policy would somehow interact with these courses (i.e. a direction for further research), it can be expected that more students will have to retake the final exam the year after, which raises the total educational costs of the policy far beyond the 1 million euros estimated here.

The increased educational costs of the policy will be much lower for primary schools, because the Cito test is unrelated to the successful completion of primary education. Furthermore, the teacher's advice on the level of education which will fit the student best nowadays also occurs *ex ante* the test is made (i.e. since 2015), therefore the poorer performance on the Cito test due to the policy's implementation will not raise costs by much. Moreover, the impact of the policy on exam scores in PE is not so unambiguous. Whereas the effect of the policy as estimated by the main model is almost equal to -1 point on the Cito test (-0.914**, -0.22SDs), the inclusion of various combinations of leads and lags make the coefficients jump around from -1.344 (-0.33SDs) to -0.389 (-0.09SDs) grade points. Furthermore, once a 1-year lag of the dependent variable is added to the model, the coefficient (-0.614, -0.15SDs) loses its significance (p=0.158). Nevertheless, due to the consistently negative (and mainly significant) point estimates a small effect of the label of excellence on exam scores seems plausible. One explanation as to why the effect of the label may be different across primary and secondary schools is that the stakes are much lower for students in PE. Therefore, it is likely that relatively less instruction time is spent on preparation for the Cito test. If the label of excellence shifts the instruction time away from verifiable performance indicators, the negative impact will be smaller for indicators which already receive little or no attention.

There is no univocal trend in student inflow following the receipt of the label of excellence either. Whereas Regioplan (2015; 2016) and Sikkes (2016) report significant counter-trend stabilizations or even increases in student inflow, my findings are typically positive but insignificant. This can partly be attributed to the crude, imprecise measures of inflow which I have adopted for secondary schools. The yearly inflow by BRINVEST (i.e. not at the level of excellence) has on average increased by 6.5 students in the wake of excellence, although the effect fully disappears if first differences are taken. Moreover, it seems to strongly interact with the level of education; especially VWO-schools expand much more. Due to data limitations, I could not control for many, possibly confounding factors, however. Regarding PE, most model specifications yield minor effects of the policy on year 1 inflow, but all point estimates are insignificant at the usual 5%-level. Taking first differences barely changes the results. Interestingly, in the specifications without controls strong lagged effects of the label are exposed. These effects disappear if controls are included, but at the same time the sample size is restricted (due to missing data). As it is difficult to disentangle their disjoint impacts, it could still very well be that the policy has a lagged effect on inflow. This is not wishful thinking: in the studies by Regioplan and Sikkes the effect also only materialized with the passing of time. Presumably, the improved reputation of an excellent school (as capitalized and signalled by its label of excellence) takes time to spread, which is why an immediate effect is not observed.

The negative impact of the label of excellence on test scores is remarkable, to say the least. Taking into account that the policy was established back in 2012 with the objective of counteracting the decline in final exam and PISA scores, you would have expected that the label of excellence would have had anything but a negative impact on test scores. Furthermore, the Inspectorate lists plenty of (side-)effects of the label of excellence which theoretically would improve the test scores of students. Amongst others, for instance, an improved atmosphere to excel, motivated staff members, closer attention for detail and additional feedback from the Inspectorate on perceived weaknesses. These effects could still be present, but the results indicate that they are more than offset by (negative)

countervailing forces. One may argue that with the shift in objectives – from 'Excellent Schools 2015' onwards developing a profile of excellence has gained in importance in assigning a label relative to the school's capacity to maximize verifiable student performance – such a drop in test scores was to be expected, because schools will disproportionately direct their means and efforts to non-tested aspects of education. However, the estimation of a regression equation (not reported) comparing the two time periods (2012-2015 and 2015-2017) reveals that the negative impact of the policy is approximately similar in both periods.

What has most likely driven the decline in test scores? A disproportionate focus on the school's profile of excellence has not been the major cause, otherwise the effect on test scores should have been more profound in the latter period. If anything, exam scores in excellent primary schools decline slightly less in 2016 and 2017[71]. Although the reduced-form fixed-effects model I have adopted identifies the effect of the policy, it fails to directly pinpoint down the underlying mechanism(-s) which has (have) mediated the relationship between the provision of a label of excellence and the drop in objective student performance. Nevertheless, quite a bit can be said about the presumed channels and pathways that have contributed to the observed decline. Two clear mechanisms stand out, which are likely to put downward pressure on exam scores in both SE and PE:

1. *The application procedure is bureaucratic and costs a lot of time and energy (i.e. an ex-ante effect)*
2. *Out of complacency and/or the crowding-out of intrinsic motivation schools slack off (i.e. an ex-post effect)*

From 'Excellent Schools 2015' onwards, labels of excellence were not only assigned conditional on a well-defined profile of excellence, but their validity was also extended from 1 to 3 years. As a matter of fact, schools which have received the label in either January 2016 or 2017 have a stronger incentive to slack off than schools which were to reapply the immediate year after the label of excellence was obtained (otherwise they would not be eligible for the label the next year). Even though both before and after the change in policy crowding-out of intrinsic motivation could occur, the effect will be less strong if you run the risk of losing the label the subsequent year. Therefore, I treat the *complacency*-effect and *motivational*-effect as one and the same (since it is statistically impossible to disentangle them).

As already stated before, the comparison of the effect of the label across the two time periods did not yield significantly different estimates. Hence, schools who have received the label either in 2016 or 2017[72] do not display a stronger negative dip in exam scores than excellent schools in the years before. This is probably a consequence of the timeline of events. Schools receive the label only in January, whereas the tests are already made either in April (PE) or May (SE). If it takes time for these (*ex-post*) motivational effects to affect student performance, it makes sense that these effects will not yet be reflected in the test scores (and may only emerge with a 1 or 2-year lag). Hence, it seems reasonable to assume that complacency or the crowding-out of intrinsic motivation are not significantly associated with the sudden drop in test scores and may not even be a consequence of the label of excellence at all. Moreover, other evidence points in the direction of red tape as the primary cause of the declining test scores.

[71] Although not significantly less. Furthermore, the later time period consists of only two years of variation in excellence status relative to three years the time period before. Sampling bias is therefore also larger in the later period.

[72] Even though these schools were also perversely incentivized to spend a disproportionate amount of time and effort on their profile of excellence, which should have reinforced the negative impact on test scores.

In Tables 5.1a and b, we have seen that the effect of losing the label of excellence is nearly inversely proportional to the effect of receipt of the label of excellence on verifiable student performance. Although I do not know exactly how many of these formerly excellent schools have tried to reapply but failed to meet the requirements for excellence the year after or simply decided not to reapply (i.e. this information is confidential), Regioplan (2016) does reveal that the preponderance of excellent schools tends to successfully reapply the year after. Therefore, conditional on an excellent school reapplying, the probability that the label will be kept is quite high. Hence, I assume that most schools which lose their label of excellence have simply not reapplied and therefore avoided the bureaucratic paperwork that comes with it. Since the exam scores of students in the year after the label is dropped suddenly jump up, this can rightly be attributed to the freed-up time and energy that can be spent on students again rather than the application procedure.

The (lack of an) effect of the label of excellence on school exams also corroborates this finding. Under the assumption that school exams should have benefitted slightly from the application procedure for the label of excellence (either through a motivated workforce, addressed didactic weaknesses, outright teacher cheating etc.), we would expect mildly positive point estimates for the label of excellence on school exams[73]. However, what we observe in Appendix XVI is that the point estimates are negative (but insignificant). Any negative effects of the label that occur *ex-post* do not affect the school exam, because the last grade which contributes to the average school exam grade is already obtained very shortly after the label is conferred. This indicates that these plausibly positive effects are more than offset by another mechanism. As all other possible mechanisms have been logically ruled out, the negative effect on exam scores can only be due to the heavy burden of paperwork which must be filled in in order to apply for the label. To sum up, there is plenty of convincing evidence that the opportunity cost of applying for the label of excellence – time spent on paperwork cannot be spent on students – will significantly hurt exam grades once the label of excellence has been obtained. Although there may be other mechanisms involved, this is not substantiated by the empirical data.

It is important to add some caveats to my conclusions. Even though the effect of the policy on exam scores is negative for excellent schools, this does not have to equally apply for non-excellent schools. If the presence of an excellent school in the neighbourhood encourages non-excellent schools to work harder (i.e. the *competition*-argument), this may still raise verifiable student performance in those schools. As long as this effect is sufficiently strong, the costs of having to redo an entire year in school may be easily outweighed by the benefits of better school results at non-excellent schools. Moreover, it should be realized that the identification strategy I have adopted cannot capture fixed effects. Hence, I cannot estimate the cumulative effect of the label a year (or more) after its conferral (if any effect exists), because time-invariant variables are eliminated in a fixed effects model. It is conceivable that the effect changes as a function of the time the label of excellence is kept, because successive cohorts of students which will take the final exam/Cito will have studied at the 'excellent' school for different lengths of time. As of now, it is impossible to say if the exam scores will return to their original levels, stay low or drop even lower in the years after a school has obtained the label of excellence. Similarly, it is important to know what happens to inflow in the long-run, because if excellent schools keep expanding relative to non-excellent schools, verifiable student performance will also be *mechanically* affected. Under the assumption that excellent schools will stay relatively better than the average school (see Table 3 and Appendix VII), relatively larger excellent schools will raise the

[73] See Footnote 57 for an elaborate explanation of what the school exam exactly entails.

average test scores simply by the fact that a larger share of the student population studies at an excellent school.

Further research should also monitor the socio-economic composition of the inflow. If higher-educated parents are better informed and send their children to the better-performing school, this would also contribute to inequality of opportunity along socio-economic lines. Currently, this does not seem to be an issue: excellent schools consist of a disproportionate number of students from a disadvantaged background. Furthermore, the concern that I initially had that excellent schools would become (even) better at the expense of non-excellent schools also seems ungrounded. If anything, excellent schools have lost out relative to non-excellent schools on verifiable indicators of student performance (especially if the positive effect of competition on non-excellent student outcomes would also be empirically observed).

All in all, there is a paradox to the label of excellence which accentuates the ambiguity towards its objectives. In an attempt to maximize student performance and encourage ambition-sensitive education through acknowledging a school's quality beyond a baseline level, it inadvertently impairs the same quality which it tries to reward (at least in the short-run). Although the Inspectorate has repetitively denied that the label is a way to induce inter-school competition, it is the only way through which the label could possibly still have a positive rather than negative effect on verifiable student performance. If non-excellent schools improve themselves as a result of their rivalry with excellent schools, it is plausible that the policy will contribute to the attainment of some of the initial policy goals – better PISA test scores and improved performance of highly able students – as formulated as part of the educational 'plan of action' in 2011. However, it is debatable whether a targeted policy as the label of excellence will have large impacts on non-excellent test scores. Approximately 3% of all Dutch schools bear a label of excellence, which implies that only a handful of non-excellent schools face 'excellent' competition. Moreover, the literature has shown that typically the impact of inter-school competition is small.

It could of course be the case that with the explicit attention for a school's profile of excellence, the Inspectorate is currently willingly trading off objective test scores for 'inspiring, innovative or motivating curricula' or 'a distinctive approach to education for a specific group of students'. If the Inspectorate believes that students will ultimately benefit from these new approaches to education, it is understandable that you want to reward schools which take such initiatives. However, most initiatives seem irreconcilable with raising exam grades, especially if schools target their means toward only a subset of all students. Furthermore, as of now, there is no empirical proof that indicates excellent schools with a particular 'profile of excellence' score better on certain non-tested aspects of education either. In addition, the consistent labelling of these schools as excellent sustains the idea that these schools aim to maximize objective student performance, because this used to be the policy's objective. Public misinformation of this kind wrongly suggests that some schools are better than others, whereas that is not necessarily the case. Future students could even be verifiably better off at a non-excellent school than an excellent school, because the latter is more concerned with its non-tested niche.

As a matter of fact, it would be better if a spade would be called a spade. In order to avoid confusion and unrealistic expectations, the misnomer 'excellent school' can better be replaced by a more neutral term, such as 'innovative school'. This would also clarify the policy's objective. If policy-makers would still want to acknowledge verifiable quality as well, it would be better to rank all schools along a certain scale conditional on the background characteristics of its student population[74], because

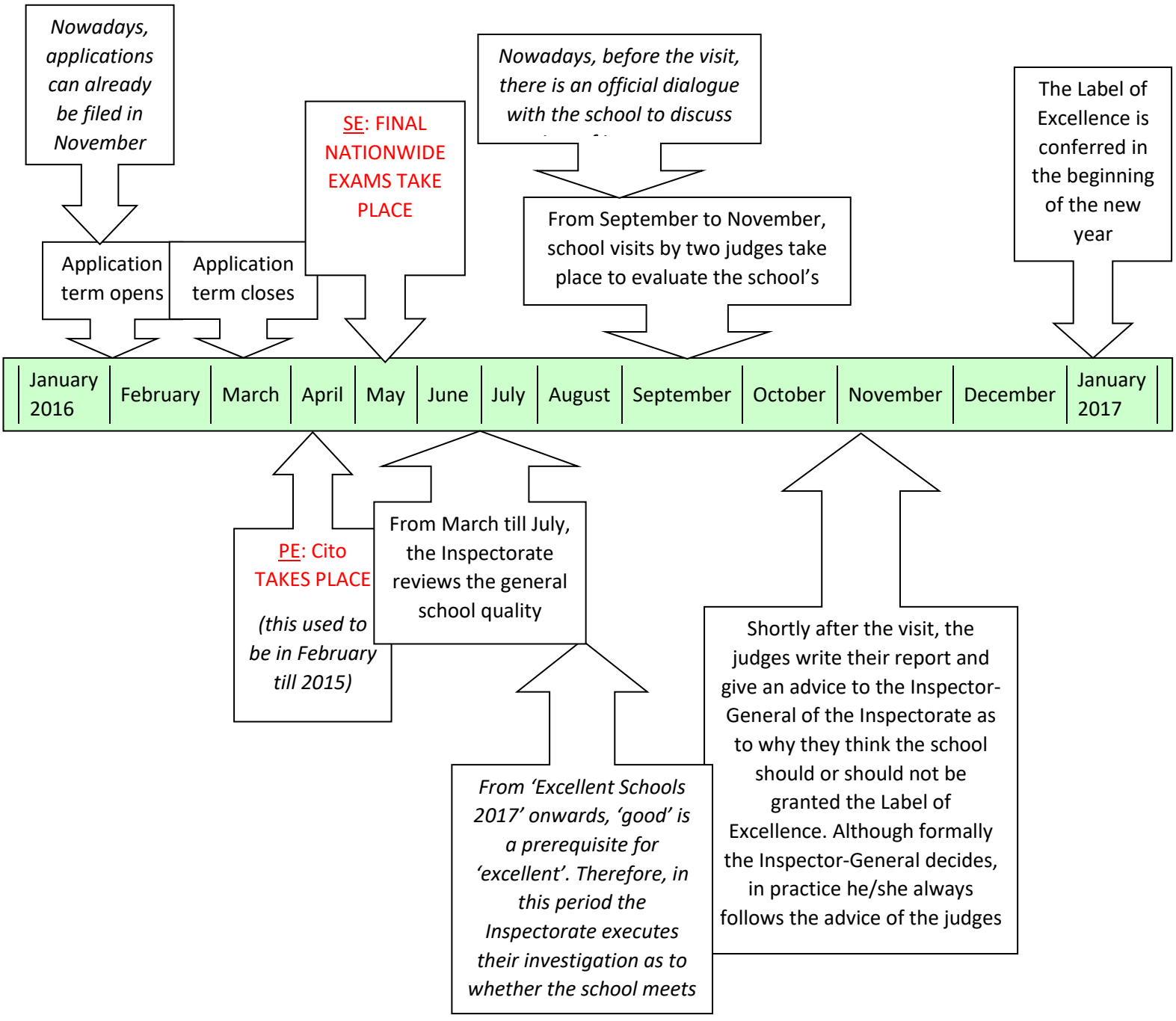
[74] The winners in their respective categories could still be awarded a 'label of excellence'.

this would make a school's relative performance accountable to the public (and not just the 'exceptional' performance of an excellent school). Moreover, these lists would encourage large-scale competition, as opposed to the small-scale, region-specific competition which may arise from the current policy. The benefits of large-scale competition will presumably also be much larger. It is hard to see how in the past five years the benefits (if any) of the label of excellence have outweighed the substantial (economic) costs. Currently, although the label may be excellent in its acknowledgement of a school's above-average quality and innovative spirit[75], it disappoints in the encouragement of real educational excellence.

[75] In 2015, all (100%) of the participants who obtained a label of excellence reported to be satisfied or largely satisfied by the label of excellence (Regioplan, 2016). One year earlier, this was 96% (Regioplan, 2015).

Appendices

Appendix I: Timeline 'Excellent Schools 2016'



NOTE: Over time, the precise timeline for the application and selection of excellent schools has changed slightly. However, the most important events are included on this timeline. Moreover, this visual representation predominantly serves to illustrate that both ex ante and ex post the receipt of the label of excellence an effect on exam scores is to be expected, as the entire procedure spans approximately a year.

Appendix II

Additional information on the five different levels of Dutch secondary education

Level (in ascending order of difficulty)	Type	Duration	Description
VMBO-B	Pre-vocational secondary education	4 years	Prepares for MBO; aimed at trade-oriented learning
VMBO-K	Pre-vocational secondary education	4 years	Prepares for MBO and combines elements from B and G(T)
VMBO-G(T)	Pre-vocational secondary education	4 years	Prepares for MBO and is a requirement for students who want to continue with HAVO after graduation; predominantly theory-based
HAVO	Senior general secondary education	5 years	Prepares for HBO (higher vocational education) and is a requirement for students who want to continue with VWO after graduation
VWO	Pre-university secondary education	6 years	Prepares for university

Appendix III:

Cross-tabulation of year by status of excellence (PE)

	(1)	(2)	(3)	(4)	(5)	(6)
	2012	2013	2014	2015	2016	2017
	Freq	Freq	Freq	Freq	Freq	Freq
Excellent	(Percent)	(Percent)	(Percent)	(Percent)	(Percent)	(Percent)
0	5,934***	5,962***	6,005***	6,593***	6,522***	6,410***
	(100)	(99.50)	(99.55)	(99.55)	(99.47)	(99.16)
1		30***	27***	30***	35***	54***
		(0.501)	(0.448)	(0.453)	(0.534)	(0.835)
Total	5934	5992	6032	6623	6557	6464

Cross-tabulation of year by status of excellence (SE)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
	Freq	Freq	Freq	Freq	Freq	Freq	Freq	Freq	Freq	Freq
Excellent	(Percent)	(Percent)	(Percent)	(Percent)	(Percent)	(Percent)	(Percent)	(Percent)	(Percent)	(Percent)
0	2,685***	2,686***	2,697***	2,710***	2,725***	3,017***	3,070***	3,049***	2,921***	2,916***
	(100)	(100)	(100)	(100)	(100)	(98.89)	(98.37)	(97.69)	(97.63)	(96.65)
1						34***	51***	72***	71***	101***
						(1.114)	(1.634)	(2.307)	(2.373)	(3.348)
Total	2685	2686	2697	2710	2725	3051	3121	3121	2992	3017

Cross-tabulation of level of education by excellence (SE)

	(1)	(2)	(3)	(4)	(5)
	Vmbo b	Vmbo k	Vmbo g(t)	Havo	Vwo
	Freq	Freq	Freq	Freq	Freq
Excellent	(Percent)	(Percent)	(Percent)	(Percent)	(Percent)
0	4,779***	4,942***	8,009***	5,430***	5,316***
	(99.03)	(99.08)	(98.93)	(98.80)	(98.46)
1	47***	46***	87***	66***	83***
	(0.974)	(0.922)	(1.075)	(1.201)	(1.537)
Total	4826	4988	8096	5496	5399

Cross-tabulation of level of education and year by excellence (SE)

Binary_treatment_variable	Level_of_education and year																			
	vmbo b										vmbo k									
	8	9	10	11	12	13	14	15	16	17	8	9	10	11	12	13	14	15	16	17
Non-excellent	464	458	455	458	459	498	512	507	485	483	484	482	483	485	483	508	519	514	493	491
Excellent						5	7	12	9	14						5	7	12	8	14

Binary_treatment_variable	Level_of_education and year																			
	vmbo (g)t										havo									
	8	9	10	11	12	13	14	15	16	17	8	9	10	11	12	13	14	15	16	17
Non-excellent	766	771	777	769	778	835	854	849	805	805	477	478	484	491	497	608	612	607	588	588
Excellent						11	13	18	17	28						3	10	15	17	21

Binary_treatment_variable	Level_of_education and year									
	vwo									
	8	9	10	11	12	13	14	15	16	17
Non-excellent	494	497	498	507	508	568	573	572	550	549
Excellent						10	14	15	20	24

Appendix IV

Transition probabilities for primary schools

EXCELLENT	*NO*	*YES*	*TOTAL*
NO	99.74	0.26	100.00
YES	22.69	77.31	100.00
TOTAL	99.44	0.56	100.00

Frequency counts of between and within data variance for primary schools

	OVERALL		BETWEEN		WITHIN
EXCELLENT	*Freq.*	*Percent*	*Freq.*	*Percent*	*Percent*
NO	37426	99.53	7053	99.94	99.59
YES	176	0.47	82	1.16	40.55
TOTAL	28805	100.00	7135 (n=7057)	101.11	98.91

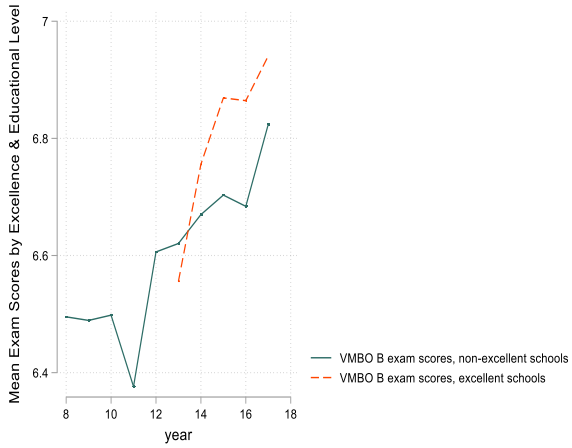
Transition probabilities for secondary schools

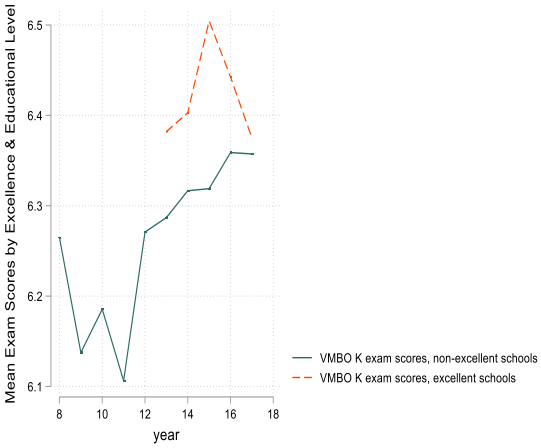
EXCELLENT	*NO*	*YES*	*TOTAL*
NO	99.45	0.55	100.00
YES	14.29	85.71	100.00
TOTAL	98.70	1.30	100.00

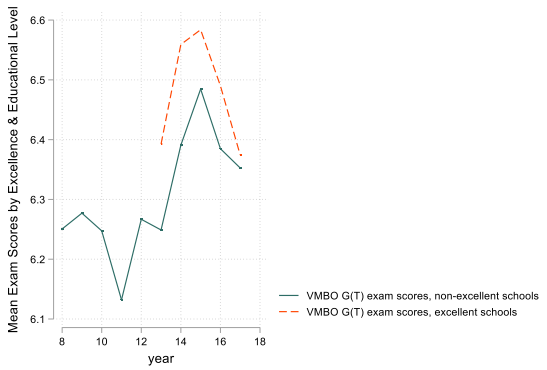
Frequency counts of between and within data variance for secondary schools

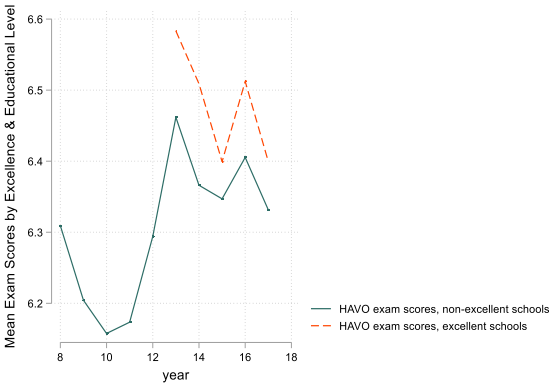
	OVERALL		BETWEEN		WITHIN
EXCELLENT	*Freq.*	*Percent*	*Freq.*	*Percent*	*Percent*
NO	28476	98.86	3528	100.00	99.01
YES	329	1.14	130	3.68	26.86
TOTAL	28805	100.00	3658 (n=3528)	103.68	96.45

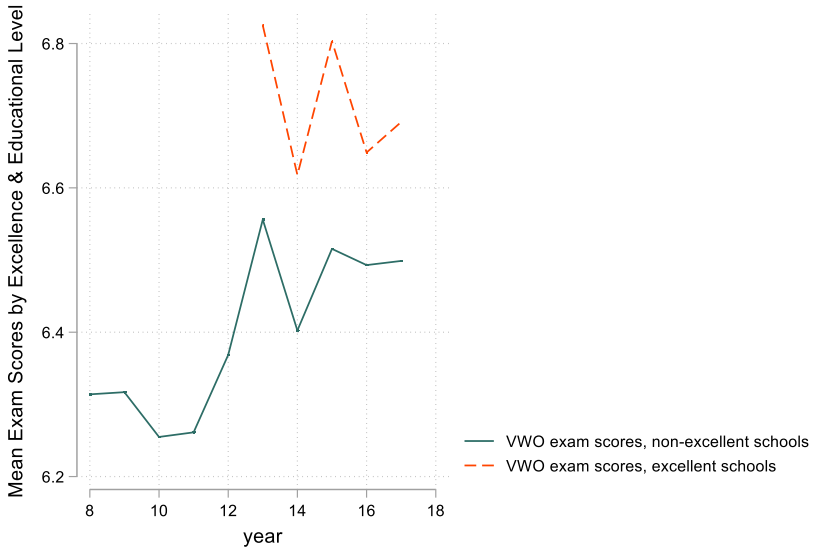
Appendix V: *Mean exam scores by excellence plotted separately by level of education*











Appendix VI

Frequency table of central exams taken by primary schools

Type of central exam	(1) 2012 Freq (Percent)	(2) 2013 Freq (Percent)	(3) 2014 Freq (Percent)	(4) 2015 Freq (Percent)	(5) 2016 Freq (Percent)	(6) 2017 Freq (Percent)
CET	5,522***	5,576***	5,573***	5,395***	4,864***	4,061***
	(100)	(100)	(100)	(95.47)	(75.87)	(64.83)
IEP				172***	1,083***	1,588***
				(3.044)	(16.89)	(25.35)
ROUTE8				82***	442***	584***
				(1.451)	(6.894)	(9.323)
CET+IEP				1***	16***	22***
				(0.0177)	(0.250)	(0.351)
CET+ROUTE8				1***	4***	6***
				(0.0177)	(0.0624)	(0.0958)
IEP+ROUTE8					2***	2***
					(0.0312)	(0.0319)
CET+IEP+ROUTE8						1***
						(0.0160)
Number of PanelID	656	656	656	656	656	656
Total	5522	5576	5573	5651	6411	6264

Frequency table of central exams taken by excellent primary schools

Type of central exam	(1) 2013 Freq (Percent)	(2) 2014 Freq (Percent)	(3) 2015 Freq (Percent)	(4) 2016 Freq (Percent)	(5) 2017 Freq (Percent)
CET	30	27	24	27	36
	(100)	(100)	(92.31)	(77.14)	(66.67)
IEP			2	6	12
			(7.692)	(17.14)	(22.22)
ROUTE8				2	6
				(5.714)	(11.11)
Number of PanelID	656	656	656	656	656
Total	30	27	26	35	54

Appendix VII

Comparison of background characteristics of excellent and non-excellent schools in PE

Variable	**Mean difference (N.ex – Ex)**	**T-statistic**	**Description of the variable**
Delta mean exam scores (DV)	0.630*[76]	(2.15)	Exam scores at *t-1* minus exam scores at *t*
Cito N students	-5.285***	(-3.50)	Number of students that have taken the Cito test
Total number of students	-83.15***	(-6.57)	Total number of students per BRINVEST
Supervisory arrangement	-0.0239***	(-24.91)	Ordinal variable; a high score implies you require little supervision as educational results are decent
Impact area	-0.0216	(-0.57)	Area with high levels of unemployment and poverty
(Socio-economic) weight of school	-22.19***	(-4.04)	Weighted indicator per school of the educational background of its students' parents. A high value implies that few parents have completed a higher level of education than VMBO. Partly corrected for school size.
Influx into year 1 (DV)	-11.47***	(-6.89)	Total number of students which enrol into school in year 1 by BRINVEST
IEP N students	-11.80*	(-2.30)	Number of students which have taken the IEP exam
IEP mean score (DV)	1.114	(1.16)	Mean difference on the IEP exam
ROUTE8 N students	-7.713	(-1.47)	Number of students which have taken the ROUTE8 exam
ROUTE8 mean score (DV)	-7.230	(-1.41)	Mean difference on the ROUTE8 exam

[76] * p<0.05, ** p<0.01 & *** p<0.001

Appendix VIII

Frequency distribution by PE's supervisory arrangement and excellence

Supervisory arrangement	(1) Non-excellent Freq (Percent)	(2) Excellent Freq (Percent)
Very weak	96***	
	(0.260)	
Weak	747***	
	(2.020)	
Baseline	36,145***	175*
	(97.72)	(100)
Total	36988	175

Frequency distribution by SE's supervisory arrangement and excellence

Excellent	(1) Very weak Freq (Percent)	(2) Weak Freq (Percent)	(3) Average/Basic Freq (Percent)	(4) Missing data Freq (Percent)
0	59	1,003***	14,486***	12,928***
	(100)	(99.80)	(97.84)	(99.95)
1		2***	320***	7***
		(0.199)	(2.161)	(0.0541)
Total	59	1005	14806	12935

Cross-tabulation of 'calculated judgement' on the 5 main quantitative indicators for SE by excellence; sub-judgements 1-5 below

Excellent	(1) Insufficient Freq (Percent)	(2) Insufficient, unless Freq (Percent)	(3) Sufficient Freq (Percent)	(4) No verdict Freq (Percent)	(5) Missing data Freq (Percent)
0	352***	96	7,417***	953***	19,658***
	(99.72)	(100)	(97.02)	(98.65)	(99.56)
1	1***		228***	13***	87***
	(0.283)		(2.982)	(1.346)	(0.441)
Total	353	96	7645	966	19745

Appendix IX

(1) Cross tabulation of the upward flow of students' level of education in year 3 relative to primary school advice

	(1) Below target Freq	(2) Above target Freq	(3) Missing data Freq
Excellent	(Percent)	(Percent)	(Percent)
0	786***	7,190***	20,500***
	(98.00)	(97.15)	(99.50)
1	16***	211***	102***
	(1.995)	(2.851)	(0.495)
Total	802	7401	20602

(2) % of students that complete the first three years of high school (or first two for VMBO) within the therefore set time

	(1) Below target Freq	(2) Above target Freq	(3) Missing data Freq
Excellent_	(Percent)	(Percent)	(Percent)
0	726***	7,552***	20,198***
	(98.78)	(97.12)	(99.53)
1	9***	224***	96***
	(1.224)	(2.881)	(0.473)
Total	735	7776	20294

(3) % of students that fulfil the upper half of their high school classes within the therefore set time

	(1) Below target Freq	(2) Above target Freq	(3) Missing data Freq
Excellent	(Percent)	(Percent)	(Percent)
0	752***	7,390***	20,334***
	(99.73)	(96.91)	(99.55)
1	2***	236***	91***
	(0.265)	(3.095)	(0.446)
Total	754	7626	20425

(4) Indicative as to whether a PanelID has reached its exam score target

	(1) Below target Freq	(2) Above target Freq	(3) Missing data Freq

Excellent	(Percent)	(Percent)	(Percent)
0	620***	7,240***	20,616***
	(99.36)	(96.87)	(99.56)
1	4***	234***	91***
	(0.641)	(3.131)	(0.439)
Total	624	7474	20707

(5) Assessment of the mean difference in final (nationwide) exam scores and school (i.e. by panel) exam scores

	(1)	(2)	(3)	(4)
	Very large difference	Large difference	Minor difference	Missing data
	Freq	Freq	Freq	Freq
Excellent	(Percent)	(Percent)	(Percent)	(Percent)
0	1	114	7,744***	20,617***
	(100)	(99.13)	(97.03)	(99.56)
1		1	237***	91***
		(0.870)	(2.970)	(0.439)
Total	1	115	7981	20708

Appendix X

Frequency table of the number of non-excellent schools sharing a ZIP with an excellent school

Dummy	SE-3digits	SE-4digits	PE-3digits	PE-4digits
Different ZIP as ex	28,061***	28,354***	35,721***	37,018***
	(97.42)	(98.43)	(95.00)	(98.45)
N.ex, same 3 or 4-digit ZIP as ex	422***	129***	1,705***	408***
	(1.465)	(0.448)	(4.534)	(1.085)
Ex	322***	322***	176***	176***
	(1.118)	(1.118)	(0.468)	(0.468)
Number of PanelID	3,000	3,000	5,883	5,883
Total	28805	28805	37602	37602

Appendix XI

DV: average score on final nationwide exams		
VARIABLE	(1)	(2)
Excellent	-0.0737***	-0.0615***
	(-0.101 - -0.0465)	(-0.0869 - -0.0361)
N_students completing upper half of high-school within target time	0.000635***	0.000498***
	(0.000443 - 0.000826)	(0.000316 - 0.000681)
% of students completing upper half of high-school within target time	0.0119***	0.0114***
	(0.0106 - 0.0131)	(0.0101 - 0.0127)
N_CE-takers	0.000719	0.00158*
	(-0.000979 - 0.00242)	(-7.58e-05 - 0.00323)
N_CE-takers in 'assisted learning'	-0.00181***	-0.00254***
	(-0.00279 - -0.000820)	(-0.00349 - -0.00159)
N_CE-takers from 'apcg'	-0.00206***	-0.00208***
	(-0.00312 - -0.00101)	(-0.00306 - -0.00109)
N_courses taken by students at CE	-0.000300**	-0.000397***
	(-0.000550 - -5.04e-05)	(-0.000639 - -0.000154)
Total amount of students by BRINVEST	-3.67e-05	-3.65e-05*
	(-8.28e-05 - 9.44e-06)	(-7.93e-05 - 6.18e-06)
Total amount of students by BRINVEST & level of education	-0.000580***	-0.000407***
	(-0.000738 - -0.000423)	(-0.000552 - -0.000262)
Constant	5.522***	5.543***
	(5.397 - 5.647)	(5.420 - 5.667)
Observations	13,196	13,196
R-squared	0.131	0.241
Number of PanelID	2,867	2,867
Panel FE	YES	YES
Time FE	YES	YES
Time*Education FE	NO	YES

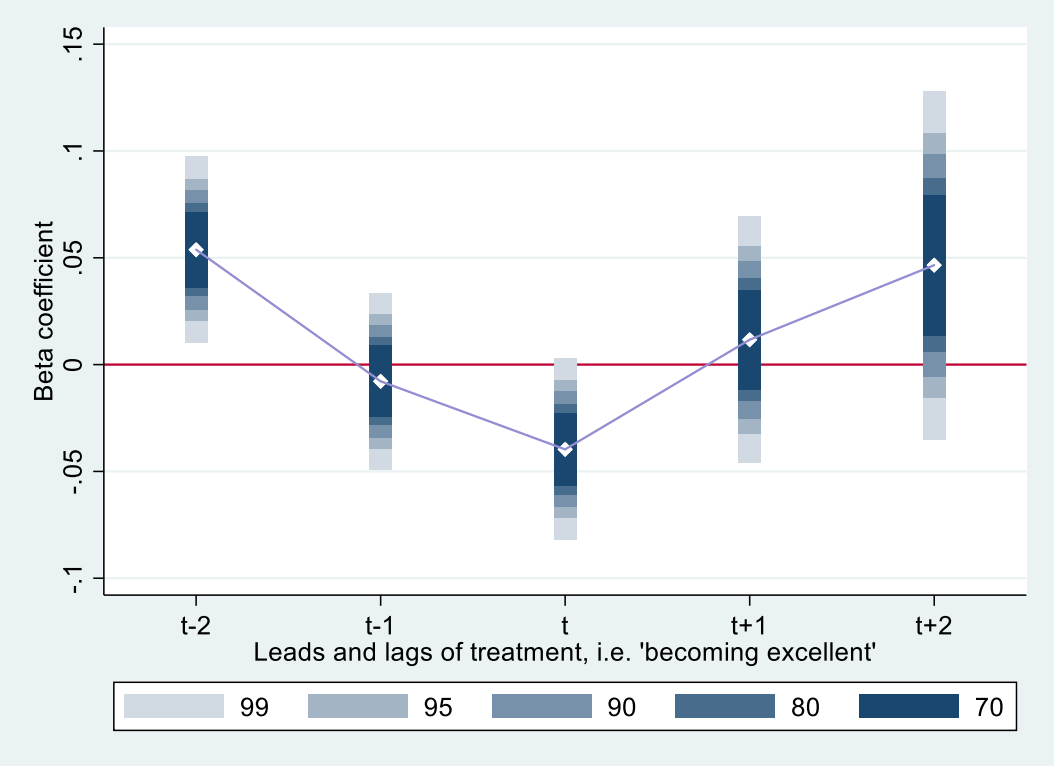
DV: average score on final nationwide exams						
VARIABLES	(1) vmbo b	(2) vmbo k	(3) vmbo g(t)	(4) havo	(5) vwo	(6) vmbo combined
Excellent	-0.0212	-0.0629*	-0.0588**	-0.0959***	-0.0433	-0.0592***
	(0.0377)	(0.0329)	(0.0242)	(0.0201)	(0.0304)	(0.0182)
Constant	6.067***	5.534***	5.106***	5.653***	5.727***	5.448***
	(0.154)	(0.138)	(0.128)	(0.111)	(0.198)	(0.0782)
Observations	2,161	2,308	3,710	2,490	2,527	8,179
R-squared	0.187	0.133	0.367	0.279	0.202	0.187
Number of PanelID	479	505	832	522	529	1,816
Group FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Controls	YES	YES	YES	YES	YES	YES

Appendix XII

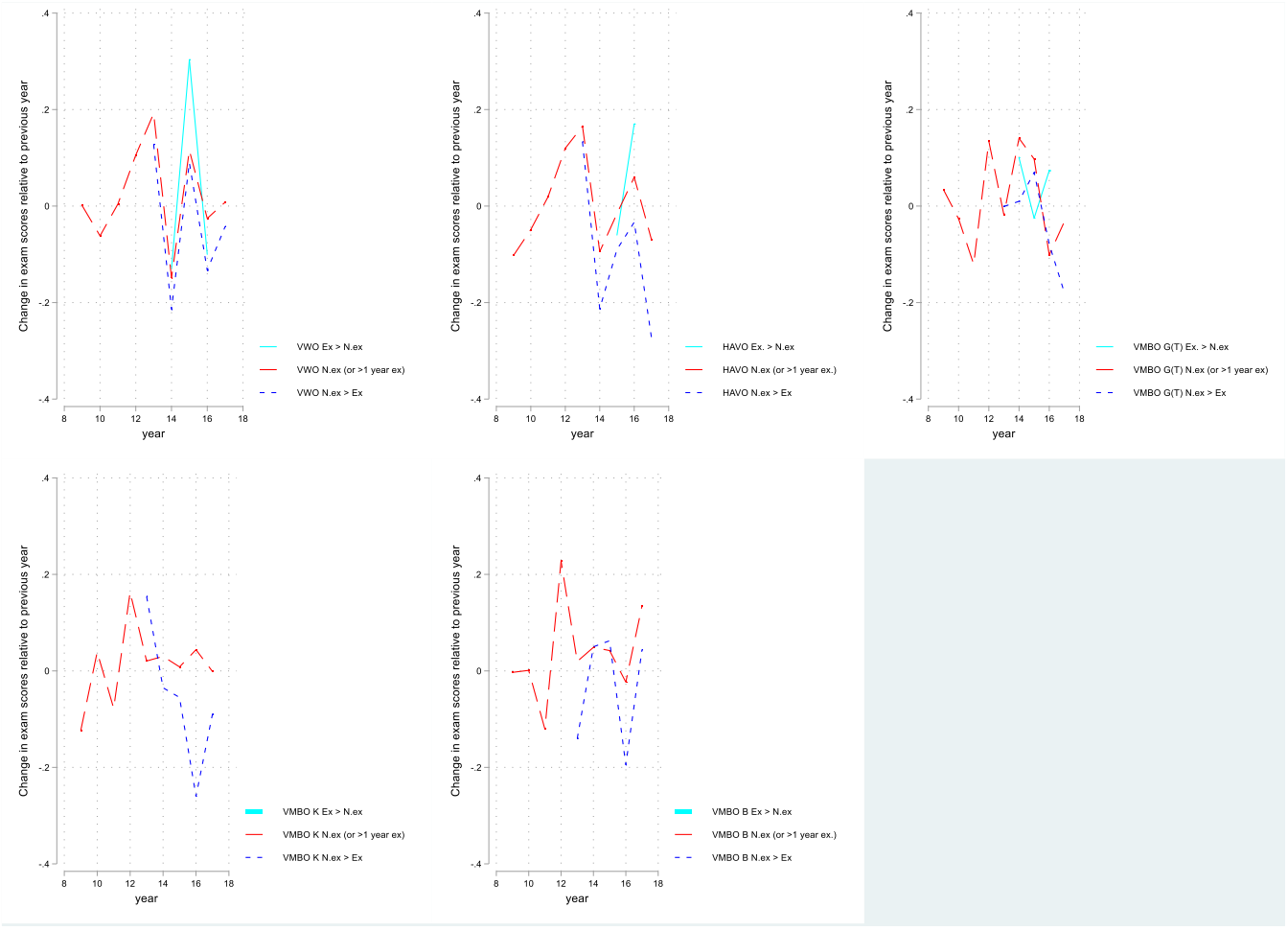
DV: average score on final nationwide exams

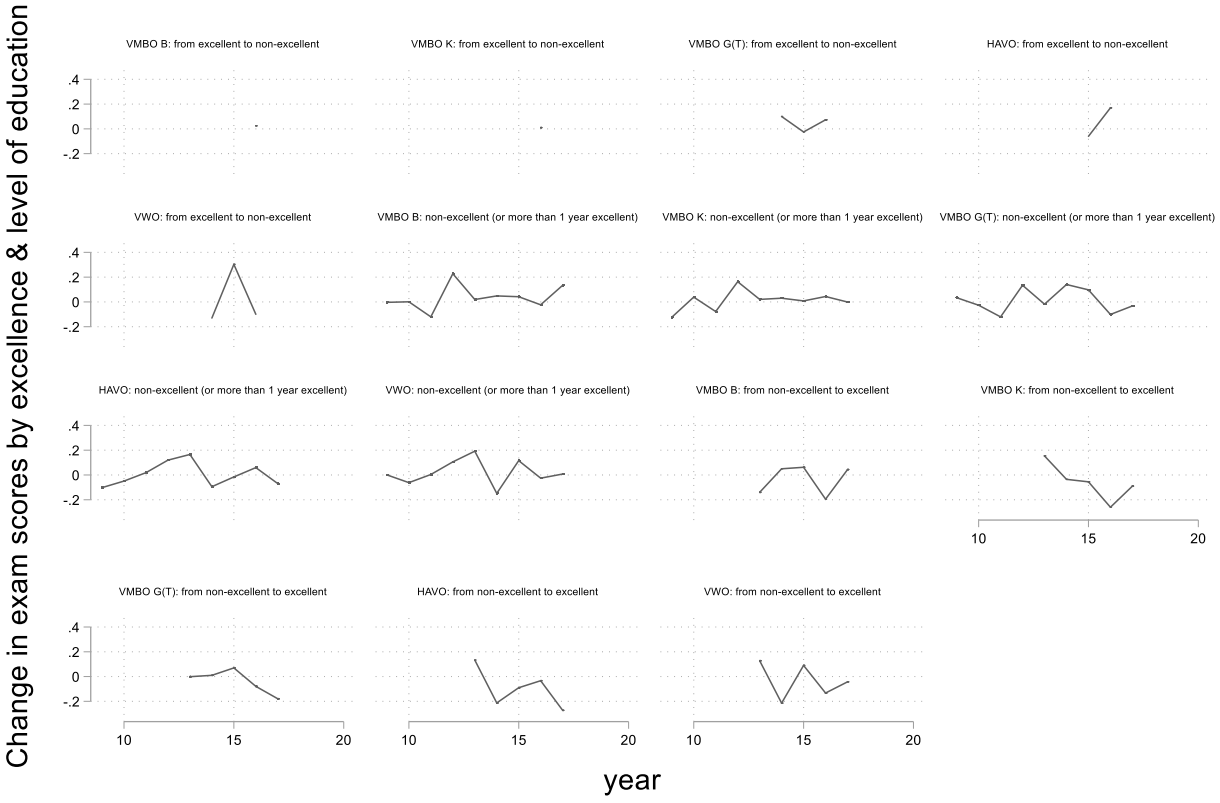
VARIABLES	(1) L1	(2) F1	(3) L1+F1	(4) L1+L2&F1+F2
Excellent	-0.0395***	-0.0445***	-0.0521***	-0.0378**
	(0.0134)	(0.0133)	(0.0132)	(0.0165)
L.Excellent	0.0205		0.0257	0.0104
	(0.0151)		(0.0186)	(0.0224)
L2.Excellent				0.0454
				(0.0317)
F.Excellent		0.0420***	0.0282*	-0.0115
		(0.0151)	(0.0150)	(0.0159)
F2.Excellent				0.0524***
				(0.0171)
Constant	6.289***	6.320***	6.291***	6.270***
	(0.00404)	(0.00408)	(0.00397)	(0.00369)
Observations	24,025	23,956	20,910	14,922
R-squared	0.261	0.236	0.264	0.314
Number of PanelID	3,241	3,225	3,130	2,905
Panel FE	YES	YES	YES	YES
Time FE	YES	YES	YES	YES
Time*Education FE	YES	YES	YES	YES

Visual representation of Model 4, directly above



Appendix XIII




Change in exam scores by excellence & level of education
VMBO B: from excellent to non-excellent
VMBO K: from excellent to non-excellent
VMBO G(T): from excellent to non-excellent
HAVO: from excellent to non-excellent
VWO: from excellent to non-excellent
VMBO B: non-excellent (or more than 1 year excellent)
VMBO K: non-excellent (or more than 1 year excellent)
VMBO G(T): non-excellent (or more than 1 year excellent)
HAVO: non-excellent (or more than 1 year excellent)
VWO: non-excellent (or more than 1 year excellent)
VMBO B: from non-excellent to excellent
VMBO K: from non-excellent to excellent
VMBO G(T): from non-excellent to excellent
HAVO: from non-excellent to excellent
VWO: from non-excellent to excellent
.4
.2
0
-.2
10
15
20
year

Appendix XIV

Cito scores regressed upon excellence within ZIP areas

VARIABLES	(1) Main regression	(2) 3 digit ZIP	(3) 4 digit ZIP
Excellent	-0.914**		
	(0.442)		
N.ex, same 3-digit ZIP as ex		0.213	
		(0.169)	
Ex, 3-digit ZIP		-0.890**	
		(0.442)	
N.ex, same 4-digit ZIP as ex			0.0416
			(0.301)
Ex, 4-digit ZIP			-0.912**
			(0.442)
Constant	535.6***	535.6***	535.6***
	(0.322)	(0.322)	(0.322)
Observations	19,461	19,461	19,461
R-squared	0.012	0.012	0.012
Number of PanelID	5,986	5,986	5,986
Panel FE	YES	YES	YES
Time FE	YES	YES	YES
Controls	YES	YES	YES

DV: Inflow into PE by year 1 and ZIP

VARIABLES	(1) Main regression	(2) 3 digit ZIP	(3) 4 digit ZIP
Excellent	1.608		
	(1.992)		
N.ex, same 3-digit ZIP as ex		-0.101	
		(0.387)	
Ex, 3-digit ZIP		1.597	
		(1.995)	
N.ex, same 4-digit ZIP as ex			0.669
			(0.813)
Ex, 4-digit ZIP			1.630
			(1.991)
Constant	38.44***	38.41***	38.45***
	(9.429)	(9.432)	(9.428)
Observations	19,466	19,466	19,466
R-squared	0.007	0.007	0.008
Number of PanelID	5,986	5,986	5,986
Panel FE	YES	YES	YES
Time FE	YES	YES	YES
Controls	YES	YES	YES

Appendix XV

Grade on final nationwide exams (SE) regressed upon an interaction of the degree of competition a school faces and its status of excellence

VARIABLES	(1) 3 digit ZIP	(2) 4 digit ZIP
Excellent (main effect)	-0.0396**	-0.0707***
	(0.0170)	(0.0176)
Medium competition within 3-digit ZIP (main effect)	-0.0325	
	(0.0296)	
High competition within 3-digit ZIP (main effect)	-0.0194	
	(0.0397)	
Ex#Medium competition within 3-digit ZIP	-0.0503*	
	(0.0263)	
Ex#High competition within 3-digit ZIP	-0.0668	
	(0.0633)	
Medium competition within 4-digit ZIP (main effect)		0.0407
		(0.0406)
High competition within 4-digit ZIP (main effect)		0.0974*
		(0.0508)
Ex#Medium competition within 4-digit ZIP		0.0153
		(0.0280)
Ex#High competition within 4-digit ZIP		0.0258
		(0.0382)
Constant	5.485***	5.444***
	(0.0604)	(0.0638)
Observations	12,972	12,972
R-squared	0.238	0.238
Number of PanelID	2,744	2,744
Panel FE	YES	YES
Time FE	YES	YES
Time*Education FE	YES	YES
Controls	YES	YES

Cito scores and year 1 inflow (PE) regressed upon an interaction of the degree of competition a school faces and its status of excellence

VARIABLES	(1) 3 digit Competition - Grades	(2) 4 digit Competition - Grades	(3) 3 digit Competition - Inflow	(4) 4 digit Competition - Inflow
Excellent (main effect)	-0.411	-0.347	1.840	2.106
	(0.612)	(0.461)	(1.720)	(1.533)
Medium competition within 3-digit ZIP (main effect)	0.526***		0.352	
	(0.182)		(0.345)	
High competition within 3-digit ZIP (main effect)	0.171		1.511	
	(0.453)		(1.078)	
Ex#Medium competition within 3-digit ZIP	-1.079		-0.518	
	(0.799)		(3.447)	
Medium competition within 4-digit ZIP (main effect)		-0.0183		0.824**
		(0.183)		(0.419)
High competition within 4-digit ZIP (main effect)		0.450		0.703
		(0.428)		(1.135)
Ex#Medium competition within 4-digit ZIP		-0.759		-1.655
		(0.716)		(4.556)
Ex#High competition within 4-digit ZIP		-5.234***		0.679

		(1.999)		(4.438)
Constant	535.3***	535.6***	38.43***	38.13***
	(0.336)	(0.330)	(9.425)	(9.422)
Observations	19,461	19,461	19,466	19,466
R-squared	0.013	0.012	0.008	0.008
Number of PanelID	5,986	5,986	5,986	5,986
Panel FE	YES	YES	YES	YES
Time FE	YES	YES	YES	YES
Controls	YES	YES	YES	YES

Appendix XVI

School exams in SE regressed upon excellence (robustness check to main regression)

VARIABLES	(1)	(2)	(3)	(4)	(5)
Excellent	-0.00744	0.00837	-0.00723	0.0138	-0.00372
	(0.00887)	(0.0171)	(0.00893)	(0.0177)	(0.00869)
Excellent#vmbo k		0.0123		0.00596	
		(0.0277)		(0.0287)	
Excellent#vmbo g(t)		-0.0143		-0.0221	
		(0.0256)		(0.0260)	
Excellent#havo		-0.0485**		-0.0502**	
		(0.0228)		(0.0236)	
Excellent#vwo		-0.0164		-0.0241	
		(0.0246)		(0.0254)	
Constant	6.462***	6.462***	6.462***	6.462***	5.937***
	(0.00209)	(0.00209)	(0.00210)	(0.00210)	(0.0379)
Observations	13,735	13,735	13,735	13,735	13,091
R-squared	0.012	0.012	0.015	0.015	0.086
Number of PanelID	3,020	3,020	3,020	3,020	2,811
Group FE	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES
Time*Education FE	NO	NO	YES	YES	YES
Controls	NO	NO	NO	NO	YES

Reference list

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533-575.

Bevan, G. & Wilson, D. (2013). Does 'naming and shaming' work for schools and hospitals? Lessons from natural experiments following devolution in England and Wales. *Public Money & Management*, 33:4, 245-252.

Bouma, K. (2018). Excellente scholen: wat levert zo'n eretitel eigenlijk op? In: de Volkskrant [22-01-18]. Retrieved from: https://www.volkskrant.nl/nieuws-achtergrond/excellente-scholen-wat-levert-zo-n-eretitel-eigenlijk-op-~b2193cf2/

Britton, J., & Propper, C. (2016). Teacher pay and school productivity: Exploiting wage regulation. *Journal of Public Economics*, *133*, 75-89.

Burgess, S. M., Propper, C., Slater, H., & Wilson, D. (2005). Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools.

Chetty, R. et al. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *Quarterly Journal of Economics*, 126(4), pp. 1593-1660.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633-79.

Cito (2018). Cito: maker van Centrale Eindtoets. Retrieved from: http://www.cito.nl/onderwijs/primair%20onderwijs/centrale_eindtoets

College voor Toetsen & Examens (2018). Uitzonderingen. Retrieved from: https://www.centraleeindtoetspo.nl/schoolbesturen-en-scholen/voorbereiding/uitzonderingen

Didactief (2013). Het verdriet van Cito. Retrieved from: https://didactiefonline.nl/artikel/het-verdriet-van-cito

Dijkgraaf, E., Gradus, R. H., & de Jong, J. M. (2013). Competition and educational quality: Evidence from the Netherlands. *Empirica*, 40(4), 607-634.

Dronkers, J. (2014). 'Excellente' en niet-'excellente' basisscholen vergeleken. Retrieved from: http://stukroodvlees.nl/excellente-en-niet-excellente-basisscholen-vergeleken/

Dutch Council of Education (2011). Naar hogere leerprestaties in het voortgezet onderwijs. Retrieved from: https://www.onderwijsraad.nl/upload/documents/publicaties/volledig/nieuw-naar-hogere-leerprestaties-in-het-voortgezet-onderwijs.pdf

Expertgroep PO (2016). Rapportage Vergelijkbaarheid Eindtoetsen. Retrieved from: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2016/12/01/rapportage-vergelijkbaarheid-eindtoetsen/rapportage-vergelijkbaarheid-eindtoetsen.pdf

Figlio, D., & Loeb, S. (2011). School accountability. In *Handbook of the Economics of Education* (Vol. 3, pp. 383-421). Elsevier.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education review*, *30*(3), 466-479.

Hoxby, C. M. (2002). *The cost of accountability* (No. w8855). National Bureau of Economic Research.

Hoy, W. K., Tarter, C. J., & Hoy, A. W. (2006). Academic optimism of schools: A force for student achievement. *American educational research journal*, *43*(3), 425-446.

HP/De Tijd (2014). Cito-toets afschaffen of niet? De argumenten op een rij. Retrieved from: https://www.hpdetijd.nl/2014-01-06/cito-toets-afschaffen-of-niet-de-argumenten-op-een-rij/

Inspectorate of Education (2011). De Staat van het Onderwijs 2009/2010. Utrecht. Retrieved from: https://zoek.officielebekendmakingen.nl/blg-109703.pdf

Inspectorate of Education (2012). Toezichtkader PO/VO 2012. Utrecht. Retrieved from: https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/publicaties/2012/07/01/toezichtskader-po-vo-2012/Toezichtskader+primair+en+voortgezet+onderwijs+2012.pdf

Inspectorate of Education (2014). Maatgevende scholen II: excellente scholen in primair, (voortgezet) speciaal en voortgezet onderwijs. Retrieved from: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2014/01/13/maatgevende-scholen-ii-excellente-scholen-in-primair-voortgezet-speciaal-en-voortgezet-onderwijs/maatgevende-scholen-ii-excellente-scholen-in-primair-voortgezet-speciaal-en-voortgezet-onderwijs.pdf

Inspectorate of Education (2015). Maatgevende scholen III: excellente scholen in primair, (voortgezet) speciaal en voortgezet onderwijs. Retrieved from: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2015/01/26/maatgevende-scholen-iii/maatgevende-scholen-iii.pdf

Inspectorate of Education (2016a). De Staat van het Onderwijs 2014/2015. Utrecht. Retrieved from: https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/publicaties/2016/04/13/staat-van-het-onderwijs-2014-2015/de-staat-van-het-onderwijs-2014-2015.pdf

Inspectorate of Education (2016b). Excellente Scholen 2016: Informatie over het traject Excellente Scholen in 2016 en een terugblik op 2015. Utrecht. Retrieved from: https://www.excellentescholen.nl/binaries/excellentescholen/documenten/brochures/2016/10/24/brochure-excellente-scholen-2016/Webversie+gewijz+versie+brochure++ES+2016.pdf

Inspectorate of Education (2017a). De Staat van het Onderwijs 2015/2016. Utrecht. Retrieved from: https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/rapporten/2017/04/12/staat-van-het-onderwijs-2015-2016/Staat+van+het+Onderwijs+geheel.pdf

Inspectorate of Education (2017b). Onderzoekskader 2017 voor het toezicht op het voortgezet onderwijs. Utrecht. Retrieved from: https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/rapporten/2018/07/13/onderzoekskader-2017-voor-het-toezicht-op-het-voortgezet-onderwijs/Onderzoekskader_vo_1_juli_2018.pdf

Inspectorate of Education (2018a). De Staat van het Onderwijs 2016/2017. Utrecht. Retrieved from: https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/rapporten/2018/04/11/rapport-de-staat-van-het-onderwijs/108126_IvhO_StaatvanhetOnderwijs_TG.pdf

Inspectorate of Education (2018b). Veelgestelde vragen Excellente Scholen. Retrieved from: https://www.excellentescholen.nl/documenten/vragen-en-antwoorden/predicaat

Kelley, C., Heneman III, H., & Milanowski, A. (2002). Teacher motivation and school-based performance awards. *Educational Administration Quarterly*, *38*(3), 372-401.

Koning, P., & Van der Wiel, K. (2012). School responsiveness to quality rankings: An empirical analysis of secondary education in the Netherlands. *De Economist*, 160(4), 339-355.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The quarterly journal of economics*, *114*(2), 497-532.

Ladd, H. F. (2012). School Accountability: To what ends and with what effects. In: *Keynote address for Conference on Improving Education through Accountability and Evaluation: Lessons from Around the World, Rome, Italy.*

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly*, *38*(4), 494-529.

Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of political Economy*, *110*(6), 1286-1317.

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of educational research*, *78*(3), 608-644.

MacNeil, A.J., Prater, D.L. & Busch, S. (2009). The effects of school culture and climate on student achievement. *International Journal of Leadership in Education*, 12:1, 73-84.

Ministry of Education, Culture and Science (2011a). Actieplan Beter Presteren: opbrengstgericht en ambitieus: het beste uit leerlingen halen. Den Haag. Retrieved from: https://zoek.officielebekendmakingen.nl/blg-114435.pdf

Ministry of Education, Culture and Science (2011b). Actieplan 'Basis voor Presteren': naar een ambitieuze leercultuur voor alle leerlingen. Retrieved from: https://www.rijksoverheid.nl/documenten/kamerstukken/2011/05/23/actieplan-po-basis-voor-presteren

Ministry of Finance (2017). IBO: Onderwijsachterstandenbeleid, een duwtje in de rug? Den Haag. Retrieved from: http://www.rijksbegroting.nl/system/files/12/ibo-onderwijsachterstandenbeleid-eindrapport-een-duwtje-de-rug.pdf

Mooij, T., & Fettelaar, D. (2010). Naar excellente scholen, leraren, leerlingen en studenten. *ITS, Radboud Universiteit Nijmegen*.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263-283.

Noailly, J., Vujić, S., & Aouragh, A. (2012). The effects of competition on the quality of primary schools in the Netherlands. *Environment and Planning A*, 44(9), 2153-2170.

Nunes, L. C., Reis, A. B., & Seabra, C. (2015). The publication of school rankings: A step toward increased accountability?. *Economics of Education Review*, *49*, 15-23.

PO Raad (2014). Cito eindtoets laatste keer in februari. Retrieved from: https://www.poraad.nl/nieuws-en-achtergronden/cito-eindtoets-laatste-keer-in-februari

Regioplan (2015). Evaluatieonderzoek naar het predicaat excellente school. Amsterdam. Retrieved from: https://www.excellentescholen.nl/binaries/excellentescholen/documenten/rapporten/2015/01/27/evaluatie-onderzoek-predikaat-excellente-school-kopie/eindrapport-evaluatieonderzoek-naar-het-predicaat-excellente-school.pdf

Regioplan (2016). Extern evaluatieonderzoek traject Excellente Scholen. Amsterdam. Retrieved from: [received through e-mail correspondence]

Rockoff, J., & Turner, L. J. (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy*, *2*(4), 119-47.

Sahlberg, P. (2007). Education policies for raising student learning: the Finnish approach. *Journal of Education Policy*, 22:2, 147-171.

Schanzenbach, D. W. (2014). Does class size matter? Boulder, CO: National Education Policy Center. Retrieved from: http://nepc.colorado.edu/publication/does-class-size-matter

Sikkes, R. (2016). Is goed niet goed genoeg? In: Onderwijsblad [05-11-16]. Retrieved from: https://www.aob.nl/wp-content/uploads/2018/01/Onderwijsblad-november-2016-Is-goed-niet-goed-genoeg.pdf

van Vuuren, D., & van der Wiel, K. (2015). Zittenblijven in het primair en voortgezet onderwijs: Een inventarisatie van de voor-en nadelen. *Centraal Planbureau.*

van Walsum, S. (2018). VO-Raad vindt predicaat 'excellente school' verwarrend en wil ervan af. In: de Volkskrant [17-03-18]. Retrieved from: https://www.volkskrant.nl/nieuws-achtergrond/vo-raad-vindt-predicaat-excellente-school-verwarrend-en-wil-ervan-af~baa7a994/

van Wieringen, A.M.L. (2011). Verwachtingsvol onderwijs. Retrieved from: https://www.onderwijsraad.nl/upload/documenten/afscheidscollege-fons-van-wieringen.pdf

Weibel, A., Rost, K., & Osterloh, M. (2007). Crowding-out of intrinsic motivation - opening the black box. Working paper available at SSRN – http://ssrn/com, ID 957770.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Zijlstra, W. (2017). Normering als basis voor betrouwbare rapportages [PowerPoint slides]. Retrieved from: https://www.centraleeindtoetspo.nl/binaries/centraleeindtoets/documenten/verslagen/2017/06/08/sessie-6--7-normering-als-basis-voor-betrouwbare-rapportages/sessie_6__7_Normering_als_basis_voor_betrouwbare_rapportages.pdf