De Economist (2012) 160:339–355
DOI 10.1007/s10645-012-9194-9

School Responsiveness to Quality Rankings: An Empirical Analysis of Secondary Education in the Netherlands

Pierre Koning · Karen van der Wiel

Published online: 4 September 2012

Abstract This paper assesses the response of Dutch secondary schools to the publication of relative quality ratings in a national newspaper (*Trouw*). Our research design exploits the discontinuities in the ranking formula that was used to generate five consecutive levels for the overall quality of schools. We find previous *Trouw* quality scores to have an offsetting effect on school quality performance, i.e. both average grades and the number of diplomas go up after receiving a negative score. These effects are confined to the lower support of the performance distribution.

Keywords School quality · School accountability · Regression discontinuity designs

JEL Classification H75 · I20 · D83

We thank the editor, two anonymous referees, Lex Borghans, Elbert Dijkgraaf, Jaap Dronkers, Michael Lindahl, Marc van der Steeg, Sietzke Waslander and Dinand Webbink for useful comments to earlier versions of the paper. The Ministry of Economic Affairs is gratefully acknowledged for sponsoring the research. We also would like to thank seminar participants at the IFAU seminar in Uppsala on November 10, 2010; the Labour/Health seminar at Tilburg University on February 8, 2011; and the Day for Belgian Labour Economists in Ghent on April 29, 2011, for useful comments.

P. Koning (✉)
VU University Amsterdam and IZA,
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
e-mail: p.w.c.koning@vu.nl

K. van der Wiel
CPB Netherlands Bureau for Economic Policy Analysis and IZA, P.O. BOX 80510,
2508 GM The Hague, The Netherlands
e-mail: kmvdw@cpb.nl

1 Introduction

Publicly evaluating the providers of public services has become increasingly common over the last decade. One of the most prominent examples in education is the state-level accountability system in the US education system which has been introduced by the *No Child Left Behind Act* (NCLB) in 2001. There is evidence that schools respond to these accountability measures by improving their test scores (Carnoy and Loeb 2003; Hanushek and Raymond 2004; Jacob 2005; Dee and Jacob 2011; Rockoff and Turner 2010) and by changing the allocation of their resources (Rouse et al. 2007; Craig et al. 2009; Chiang 2009; Bacolod et al. 2009; Rockoff and Turner 2010).[1] Some of the resulting gains in performance can be attributed to gaming, such as when schools decide to remove low-performing students from participation in exams (Figlio and Getzler 2002; Jacob 2005). The estimates of overall school responses range from 20 to 40 % of the standard deviation of test scores (e.g. Hanushek and Raymond 2004). Estimates are generally smaller when authors focus on the specific impact of sanctions on failing schools (Rockoff and Turner 2010; Figlio and Rouse 2006; Chiang 2009). While most papers in this literature focus on individual schools, there is some evidence about the overall impact of introducing accountability systems on academic performance. In this respect, Hanushek and Raymond (2005) find that student achievement growth in US states with accountability systems has been higher than in those without.

This paper investigates the response of Dutch secondary schools to ranking scores, measured in terms of their overall test and diploma performance. The Netherlands is an interesting research environment for school accountability issues, as there is a long history of free school choice resulting in the (virtual) absence of school catchment areas. Our analysis is the first to address the impact on schools of quality scores that have been published by a newspaper, rather than public interventions that aim to track, sanction and improve failing schools. The quality rating system that is the focus of this paper has been initiated in 1997 by the daily newspaper *Trouw*, so as to inform parents and their children on the quality of secondary schools. The literature on private initiatives by newspapers or magazines typically focuses on the hospital industry, like in Pope (2009) who analyses the effects of the "America's Best Hospitals" publication of the US News and World Report (Pope 2009). Moreover, in contrast to most of the literature on ranking and accountability effects, the construction of the *Trouw* ratings was not transparent ex-ante. *Trouw* has adapted the specific weights attached to each indicator and each control variable each year, and in certain years variables were added or deleted from the formula. We thus argue that, although schools were aware of their individual continuous quality score, they did not know their exact position in the ranking before *Trouw* published it. Therefore school boards could not respond to the categorical dimension of the *Trouw* score before its publication.

[1] In contrast, there is a limited literature on the effects of school quality information on school choice behavior. In a field experiment, Hastings et al. (2008) find parents of low-income families to respond to simplified information on academic achievements and admission odds if they had never received any explicit information before. Koning and Van der Wiel (2012) find school choice for secondary education in the Netherlands to respond to quality information, particularly for schools that offer the highest school track in secondary education.

In our analysis, we use information on all school tracks observed in The Netherlands from 1996 to 2008. *Trouw* scores by school track are mainly based on three quality indicators: the average grades in final centralised exams, the percentage of students who obtain a diploma without delay and the percentage of students who end up in a lower or higher school track than initially expected (by primary school tests). To obtain an indicator for the value added by schools, the 'gross' (absolute) quality score that follows from clustering the performance measures is corrected for the percentage of students from immigrant neighbourhoods and for the percentage of students with low parental income.

The primary interest in our analysis lies in the short and medium term response of schools to the release of *Trouw* ratings (see also Chiang 2009). With quality scores that are predominantly driven by (lagged) test results and passing rates, an important question is whether schools that receive a negative quality score tend to improve these quality indicators. We exploit the fact that the *Trouw* rating consisted of five discrete, consecutive categories ("most negative", "negative", "average", "positive" and "most positive"), enabling us to use Regression Discontinuity (RD) designs with four cut-off points. This method is roughly similar to Rockoff and Turner (2010), who use the discontinuous relationship between accountability grades and the numeric inputs that determine the grades, comparing the subsequent outcomes in schools that received different grades but where otherwise similar (with six grades, ranging from "A" to "F"). The four performance indicators that are used as dependent variables in our analysis include the three main components of the overall *Trouw* score—diplomas without delay, the grades of final exams and the junior years performances—and the grades of the interim exams.

Overall, we find school quality performance to respond to *Trouw* quality scores on two of the examined performance indicators. Given that student numbers directly determine school funds and that school choice of children in the most academic track ('vwo') is influenced by the ranking (Koning and Van der Wiel 2012), it is to be expected that schools respond to this ranking. In particular, both average grades increase and the number of diplomas goes up after receiving a negative score. This is striking, as it may be costly to improve the scores on these quality measures. For schools that receive the most negative ranking, the effects of quality transparency on final exam grades and obtained diplomas can amount to 12 % of a standard deviation, compared to the average of this variable. This suggests that the most negative ranking works like a 'wake-up call' to schools. Using data on worker turnover rates for schools, we find some supporting evidence that schools with low rankings respond by hiring additional managers. Unfortunately, the overall impact of the publication of quality indicators in *Trouw* on the quality of Dutch education in general cannot be measured. *Trouw*'s first publication appeared in 1997, whereas The Netherlands has only participated in the PISA database—a reliable international comparison of student achievement in secondary education—from 2003 onwards (see e.g. Van der Steeg and Vermeer 2011).

This papers proceeds as follows. Section 2 explains the Dutch institutional context, the derivation of the *Trouw* ranking scores and presents some characteristics of the data at hand. Section 3 presents our research design and Sect. 4 the estimation results. Section 5 concludes.

2 Institutions and Data

For our analysis, two datasets are merged at the level of individual school track locations, resulting in a total sample of 24,614 observations. First, we have extracted information from the administrative records of the Inspectorate of Education (*Inspectie van het Onderwijs* or *Onderwijsinspectie*). These data include the number of establishments per school group, school denomination,[2] student numbers, and performance indicators per school track, like the average grade scores and average fractions of diplomas that were obtained.[3] Second, from the Executive Education Office (*Dienst Uitvoering Onderwijs*) we received information about the stock, inflow and outflow rates of employees for three worker levels (management, teachers and support staff) for each school.

In 2002 major reforms were implemented for lower secondary (vocational) education in the Netherlands, causing the school track classification to change. We therefore restrict the sample to the three general education tracks that existed throughout the whole sample period. School tracks include the academically oriented school track that lasts six years, of which a diploma guarantees admission to university (in Dutch: 'vwo'); a less difficult track that lasts five years, of which a diploma guarantees admission to a 'hogeschool' (comparable to community colleges; in Dutch: 'havo'); and a track that provides a general, basic education that lasts four years (in Dutch: 'vmbo-gt').

Second, we have copied all quality scores that *Trouw* has published since 1998. *Trouw* was the first media outlet to publish rankings of secondary schools and by now it is commonly acknowledged as a major source of information on secondary schools, along with the *Elsevier* weekly magazine.[4] The *Trouw* publication, which was distributed in December in most years in the time period under investigation, is based on the performance of schools in the second most recent school year. Thus, the time delay between the realization of performance outcomes (which are realized in July as the academic year ends) and the *Trouw* publication is at least 18 months. Each year *Trouw* receives quality information from the Dutch Inspectorate of Education, and subsequently determines the ranking categories of school tracks. *Trouw* ranking scores are observed for 17,229 combinations of school tracks and years. Missing observations mostly stem from the fact that schools were considered too small to obtain a reliable overall quality score.[5] Although the ratings were based on information of the Inspectorate of Education, these could not be inferred straightaway. We return to this issue later on.

[2] Possible denominations include protestant schools, catholic schools, public schools and various other smaller ones.

[3] We enriched these data with the number of inhabitants in the municipalities the school tracks were located.

[4] Since 2000, the quality information that serves as the input of the *Trouw* scores is made publicly available on the internet by the Dutch Inspectorate of Education. The way this information is presented however—with relatively many details and without a summary score—hampers a direct comparison between schools.

[5] When estimating a two step Heckman model (that allow for missing observations in the first step), we do not find evidence of selection effects on missing observations.

Table 1 presents summary statistics for the selected sample of secondary schools in 1996–2008, both for the full sample and for the sample of schools that are observed over the full time period (i.e. the balanced panel). Notice that the full sample in the table exceeds the sample that can be used to estimate the actual impact of the quality scores which starts in 1998, as this requires a lag of two years. In the full sample we have on average 9.7 yearly observations per combination of school and school track, with 13 yearly observations at maximum. Generally, differences between the means of both samples are only modest. A substantial fraction of schools offer all (three) school tracks and there is no dominant type of denomination. Furthermore, the Inspectorate has defined a variable on the presence of ethnic minorities (in Dutch: 'cumi'-students) as students living at in a zip code area with a relatively high fraction of ethnic minorities.[6] Additional funding has also been available to schools with a high percentage of minority students.

The Dutch Inspectorate of Education monitors school quality with a set of three indicators that also serve as inputs for the *Trouw* rating. The first indicator is the average percentage of students that leave the school with a diploma without any delay, measured from the third year onwards (with an average of 70 % per school). Among other things, this indicator checks whether schools game their results by excluding low-performing students from final exams in the last year.

Second, the Inspectorate of Education monitors the average final exam grades at each school track. The grade that determines whether one receives a diploma is the average grade that is obtained in the final, centralised exams and in the interim school-level exams. Interim exams are carried out halfway through the final school year, with individual teachers having the discretion to construct and correct the exams. In contrast, final exams are nationally organised and the correction is carried out by teachers at other schools. The average test score at the final exams equals 6.3 (out of 10 points) for the full sample and 6.6 for the interim exams. This suggests that teachers use their discretion in the interim exams to raise grade scores to some extent, increasing the odds of passing the final exams at the end of the school year.

Third, the Inspectorate of Education measures the net percentage of students in third year that are in a school track that is either below the advice of a child's primary school, or above. The 'junior-years performance' is documented as schools could otherwise game their results by forcing students into lower school tracks. A score of 100 % indicates that on average students are in their predicted school track.

We stated earlier that the *Trouw* ranking scores cannot fully be recovered from the performance indicators that are provided to us by the Inspectorate. This is partly because the Inspectorate provides more detailed variables to *Trouw* than to us and partly because *Trouw* has adapted their scoring method from year to year. For both us and *Trouw* it was impossible to reconstruct the exact procedure used each year, particularly as journalist turnover rates have been high. Still, we know that the three objective quality indicators were recurrent inputs for the 'gross' quality score that followed from clustering analysis. We also know that sometimes the percentage of students

[6] The definition of cumi-students has changed in 2003 and in 2005. The average value of this variable is therefore not presented in Table 1. In the estimation of our models, we therefore control for this variable by allowing its impact to vary from year to year.

Table 1 Summary statistics school/school track data: full sample and balanced panel (1996–2008)

	Full sample ($N = 24,614$)		Balanced panel ($N = 16,679$)	
	Mean	SD	Mean	SD
School track types (fractions)				
Lowest general track (Vmbo-gt)	0.633	(0.482)	0.594	(0.491)
Middle track (Havo)	0.220	(0.414)	0.238	(0.426)
'Academic' track (Vwo)	0.147	(0.354)	0.168	(0.374)
School tracks per school[a]				
1 School track	0.417	(0.493)	0.346	(0.476)
2 School tracks	0.150	(0.357)	0.157	(0.364)
3 School tracks	0.433	(0.496)	0.497	(0.500)
Market characteristics				
Municipality population, aged 10–20	14,087	(18,248)	14,163	(18,447)
Number of schools in municipality	10.013	(13.002)	10.128	(13.185)
School characteristics[a]				
Number of students per school track	986.4	(525.7)	1,081.0	(502.4)
Inflow new students per school track	207.1	(106.4)	221.3	(102.3)
Number of students per school[a]	2,032.1	(1168.8)	1,916.8	(1066.9)
Management staff per school[a]	9.5	(6.6)	8.9	(6.4)
Inflow rate	0.110	(0.191)	0.107	(0.189)
Outflow rate	0.148	(0.210)	0.143	(0.209)
Teacher staff per school[a]	142.3	(78.5)	132.2	(77.2)
Inflow rate	0.100	(0.126)	0.098	(0.124)
Outflow rate	0.104	(0.136)	0.102	(0.134)
Support staff per school[a]	28.6	(28.1)	25.8	(26.6)
Inflow rate	0.269	(0.314)	0.271	(0.319)
Outflow rate	0.144	(0.194)	0.143	(0.195)
School performance				
Diploma without delay (%)	71.470	(16.360)	70.633	(15.821)
Grade final exams	6.335	(0.288)	6.354	(0.265)
Grade interim exam	6.610	(0.285)	6.616	(0.279)
Junior years performance (first to third class)	100.800	(9.611)	100.586	(8.746)
Quality scores[b]				
Most negative ranking: '−−'	0.014	(0.115)	0.012	(0.111)
Negative ranking: '−'	0.182	(0.386)	0.179	(0.383)
Neutral ranking: '0'	0.605	(0.489)	0.615	(0.487)
Positive ranking: '+'	0.191	(0.393)	0.187	(0.389)
Most positive ranking: '++'	0.009	(0.092)	0.007	(0.085)

[a] Average and standard deviation is computed per school (not per school track)
[b] Note that the *Trouw* quality scores are unobserved for 1996 and 1997. In subsequent years, on average about 14 % of the yearly observations per school track is missing

Table 2 Quality indicators per ranking score; full sample on school-track level

	Most neg. ('−−')	Negative ('−')	Average ('0')	Positive ('+')	Most positive ('++')	Average increase per category
Diploma without delay (%)	49.015 (16.684)	60.655 (16.768)	71.888 (15.143)	78.067 (14.208)	81.215 (16.814)	8.050
Grade final exams	5.926 (0.264)	6.140 (0.272)	6.362 (0.228)	6.546 (0.254)	6.614 (0.332)	0.172
Junior years Performance (%)	86.595 (7.338)	95.034 (9.723)	100.095 (8.304)	104.764 (8.974)	112.592 (8.740)	6.499
Grade interim Exams	6.446 (0.264)	6.545 (0.270)	6.587 (0.267)	6.658 (0.285)	6.699 (0.306)	0.063

retaking classes was taken into account as an additional quality indicator. Moreover, in an attempt to control for the 'quality' of students *Trouw* corrected the overall 'gross' score in all years for the fraction of students from predominantly immigrant neighbourhoods. For most years these variables were supplemented with additional controls, particularly the percentage of children from low-income families, so as to obtain more accurate measurements for the value-added of schools. An additional advantage of these controls is that we can filter out most of the (limited) indirect effects of school quality indicators on school performance in later years through neighbourhood characteristics.

Because *Trouw* used factor analysis to determine the overall quality score, the relative weight of each performance indicator differed over the years. Additionally, the cut-off points dividing school in five quality groups were determined each year ex-post, depending on the overall distribution of performances. As a result of these seemingly random changes in the calculation procedures, school boards were ex-ante not aware of how their own continuous quality performance, which they did observe in advance, would result in one of the five relative *Trouw* quality ranks.

Table 1 shows that 1.4 % of the schools received the most negative ranking and 0.9 % the most positive one. The majority of schools were in the average category (60.5 %) and the remaining schools were distributed almost evenly over the other two categories. Table 2 mirrors the relation between the quality indicators and the resulting quality scores. The spread between the diploma rates is substantial, with 49 % (81 %) for schools in the most negative (positive) category. The relationship between the rankings and the interim exam grades is less marked, suggesting that schools with lower ratings use their discretion to compensate their lower performance on this performance measure (De Lange and Dronkers 2007).

Finally, Table 3 shows the dynamics of the *Trouw* ratings per school, measured as year-to-year transition probabilities. Note that there are virtually no schools that close down in our sample. Schools with the lowest and highest quality score are not very likely to receive a similar ranking in the next period. In particular, only about 8 % of schools stay at the most negative ranking. The extreme event of receiving the

Table 3 Transition probabilities between ranking scores (1998–2006); rows = origins, columns = destinations

	Unknown	Most neg.	Negative	Average	Positive	Most positive
Unknown	65.7	0.5	7.1	19.7	6.5	0.6
Most neg. (‘−−’)	9.7	8.3	45.2	32.7	4.2	0.0
Negative (‘−’)	7.5	2.4	32.9	51.0	6.2	0.1
Average (‘0’)	5.5	0.6	14.8	64.2	14.7	0.3
Positive (‘+’)	5.9	0.1	6.3	52.3	33.8	1.7
Most positive (‘++’)	9.9	0.0	3.5	32.4	44.4	9.8

most negative or most positive ranking is thus largely transitory (see also Dijkstra et al. 2001). This confirms earlier findings of Figlio and Lucas (2004), who argue that grading systems can have a large stochastic component. Schools that have large idiosyncratic gains from one cohort to the next in the year preceding the school rating will have more difficulty matching those same idiosyncratic gains in the subsequent year, and may fall back on that basis. Similarly, aggregate test scores (and suspension and absenteeism rates) themselves have large noise components to them.

3 Empirical Strategy

In our empirical analysis, we argue that schools are not aware of the exact quality score that *Trouw* hands out as they are unaware of their exact position vis-à-vis other schools and as the procedure and cut-off points that determine the five quality ranks change every year. Although it takes two years for the *Trouw* publication to get published after the realisation of the underlying variables, we thus assume that the exact rating is unanticipated and exogenous once we take the level of the underlying variables into account. The event of receiving a low quality score in the newspaper *Trouw* may therefore increase the awareness of schools of their relative quality level and trigger them to change policies.

Figure 1 presents an intuitive, graphical explanation of how our identification strategy works. The latent variable R^* is a weighted average of quality measures of a particular school, which is mapped in five discrete categories to obtain the rating scores (with $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$ and $\boldsymbol{\mu}_4$ as cut-off points). Obviously, we may expect the value of a particular quality measure $\boldsymbol{Q}$ to increase with R^*. However, when there is also an impact of the rating system in itself, this will be mirrored by offsetting effects at the cut-off points—with all rating effects in this case being equal to γ. At $\boldsymbol{\mu}_1$, for example schools just below the threshold will use their very low score as a ‘wake-up’ call, and these schools might henceforth improve their quality measure $\boldsymbol{Q}$ more than those schools just above the threshold.

More formally, we assume that, conditional upon the inputs of the *Trouw* score—i.e. the four performance measures (denoted as matrix $\boldsymbol{Q}$) and ethnicity and parental income as the quality controls (denoted as matrix $\boldsymbol{Z}$)—the ranking grade R (with

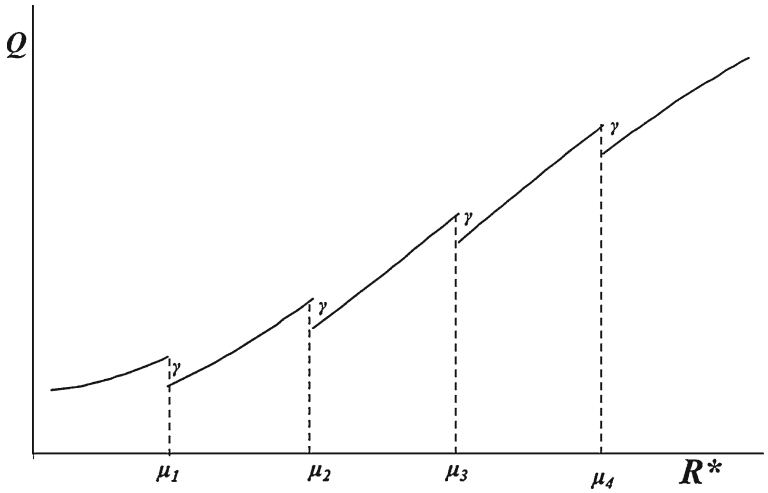

Fig. 1 The regression discontinuity design, applied to the *Trouw* grading system

$R = 1, \ldots, 5$) of a school is exogenous. We formalise this by first specifying an Ordered Probit function for the *Trouw* grades, with R^* as the latent variable:

$$R^*_{ijt} = \boldsymbol{Q}_{ijt}\boldsymbol{\delta}_t + \boldsymbol{Z}_{ijt}\boldsymbol{\alpha}_t + \varepsilon_{ijt} \tag{1a}$$

with

$$\begin{aligned} R_{ijt} &= 1 \quad \text{if } R^*_{ijt} \leq \mu_1 \\ R_{ijt} &= 2 \quad \text{if } \mu_1 < {R_{ijt}}^* \leq \mu_2 \\ &\vdots \\ R_{ijt} &= 5 \quad \text{if } \mu_4 < R^*_{ijt} \end{aligned} \tag{1b}$$

for school $i (i = 1, \ldots, I)$ with track $j (j = 1, \ldots, J)$ at time $t (t = 1, \ldots, T)$. In these equations, $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ describe the impact of the performance measures and the quality controls on the latent variable value R^*, for the *Trouw* grade R. Acknowledging the fact that *Trouw* changed its computation over time, we allow $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ to vary across years. We assume the residual value ε to be (standard-) normally distributed with mean zero and variance equal to one. Vector $\boldsymbol{\mu}$ includes four cut-off points that are needed to estimate the probabilities of obtaining the five *Trouw* grades. The parameters $\{\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\mu}\}$ of Eq. (1) can be estimated by Maximum Likelihood.

We next define ê as the difference between the actual *Trouw* grade and the expected value estimate of the *Trouw* grade that follows from Eq. (1):

$$\hat{e}_{ijt} = R_{ijt} - \sum_{s=1}^{5} s\mathrm{P}\left(R_{ijt} = s\right) \tag{2}$$

Our key RD assumption is that the conversion from expected values to categorical *Trouw* grades is unanticipated and exogenous (see also Rockoff and Turner 2010). This

means we can use $\hat{e}$ instead of R to estimate the impact on performance outcomes of school tracks consistently.

The second step of our estimation procedure entails the estimation of the effect of the residual value of Eq. (3) on quality outcomes of school tracks. This means we isolate the (discontinuous) variation in R by replacing it with $\hat{e}$. We thus specify the quality indicator $Q^k (k = 1, \ldots K)$ for school $i (i = 1, \ldots, I)$ with track $j (j = 1, \ldots, J)$ at time $t (t = 1, \ldots, T)$ as

$$Q^k_{ijt} = \gamma^k \hat{e}_{ij,t-2} + X_{it}\beta^k + \nu^k_{ij} + \eta^k_{ij} t + \zeta^k_{ijt} \tag{3}$$

with the percentage of diplomas received, the final and interim exam scores and the junior years performance as the four quality outcome measures under investigation. We normalise the values of $\boldsymbol{Q}$ to variables with mean zero and standard deviation of one, so as to measure coefficient effects in terms of standard deviations that are easier to interpret and comparable to other studies. Note that we delay the ranking grade R of school tracks with two years, as it takes at least one and a half years for the *Trouw* publication to be released after the realisation of its inputs.

For expositional reasons, we start by assuming that the impact of one higher ranking category is equal for all ranking categories (see also Fig. 1). This corresponds to the linearity assumption of Pope (2009) who studies the effects of rankings on the number of hospital clients. We relax this assumption later on, as an extension to the model—in a way that is similar to the setup of Rockoff and Turner (2010).

Matrix $\boldsymbol{X}$ includes the time varying municipality and school characteristics that are, amongst others, presented in Table 1, together with yearly time dummies and the variables that are included in $\boldsymbol{Z}$. By including yearly time dummies we control for any policy changes like the introduction of new educational systems; we assume these changes to affect all schools equally. Vector ν^k indicates school track fixed effects per quality indicator k. The relevant stratum we use here is that of school tracks that are indexed as combinations of i and j. Next to school track fixed effects, we also allow for (linear) time trends to differ between school tracks. These are represented by the vector η^k. Finally, ζ^k represents residuals that are assumed to be identically and independently distributed with mean zero and variance σ^2_k for each quality variable.

To estimate school track fixed constants and time trends in Eq. (3), we first estimate time trend coefficients with standard OLS for the (time varying) variable values $\boldsymbol{Q,X}$, and $\boldsymbol{Z}$ for each school track separately. We thus only include school tracks of which we have at least three yearly observations. Next, we de-trend the values of these variables and estimate the baseline Eq. (3), while allowing for the clustering of standard errors at the level of school tracks.

4 Estimation Results

4.1 The Baseline Model

Table 4 presents coefficient estimates of our baseline model (i.e. Eq. (3)) for the four quality outcomes, i.e. the diplomas without delay, the grades of final and interim exams

Table 4 School track FE estimation of standardised and de-trended quality performance measures (1998–2006)[a]

	Diploma without delay	Grade final exams	Grade interim exams	Junior years perform
Ranking response ($t-2$)	−0.041***	−0.037**	−0.002	0.012
	(0.012)	(0.015)	(0.015)	(0.016)
Municipality population 10–20, log value	−0.031	0.002	−0.002	0.010
	(0.031)	(0.058)	(0.057)	(0.042)
# Schools in municipality	−0.007	−0.012	0.006	−0.003
	(0.005)	(0.008)	(0.009)	(0.007)
# School tracks per school	0.014	0.009	−0.005	0.031*
	(0.012)	(0.014)	(0.016)	(0.016)
# Students per school track, log value	−0.197***	−0.240***	−0.127**	−0.366***
	(0.044)	(0.058)	(0.055)	(0.056)
Fraction variance due to FE	0.184	0.224	0.206	0.188
R-squared (overall)	0.009	0.003	0.001	0.009

[a] We also have estimated the baseline model for sub samples of school level types. This yields coefficient estimates of the coefficient estimates of the ranking responses that do not differ significantly than those obtained for the full sample. The results of these regressions are available upon request

Note that we also year dummies as additional explanatory variables. Moreover, we included the variables of matrix $\boldsymbol{Z}$ (i.e. ethnicity and parental income). As the definition of these variables changed in the time period under investigation, the impact of these variables was allowed to vary across years

St. errors corrected for school track clustering; */**/*** denote significance at 10, 5 and 1 %

and the junior year's performance.[7] Overall, we find that higher *Trouw* ranking scores lower the percentage of students receiving a diploma and also lower the final grades of students, with values around 4 % of the standard deviation of the respective scores for a one-level change in the *Trouw* rating. As the final exam scores are organised nationally, these outcomes could not result from increased gaming efforts. The size estimates of these outcomes are comparable to the results of Rockoff and Turner (2010) on math scores for "D" and "F" grades, measured in the first year after publication of accountability systems.

Interestingly, we do not find any effects for the interim exam scores. As the centralised score leaves no room for gaming, this suggests that school tracks that received lower rankings did not engage in (additional) gaming activities to improve their quality performance in future periods. School tracks could have increased the interim exam scores to improve the percentage of students receiving a diploma which is one of the inputs of the *Trouw* formula. Such effects are however seemingly small and insignificant. At the same time, note that schools only have a limited time to change interim

[7] As to the Order Probit regression (i.e. Eq. (1)), we found that 67 % of the variance in *Trouw* grades could be explained by the performance outcomes and quality control. As we will show later on, our tests in the next subsection reveal that the resulting residuals are not correlated with the grading score. Moreover, as the coefficient estimates of the Ordered Probit regression were allowed to differ across (all) years in our sample, we decided not to include them in the paper. The results are available on request.

exam requirements that are examined halfway through the year. This also holds for the junior years performance as an outcome measure, which is an indictor that is realized over a period of two years.[8] We return to this issue when discussing the persistency of effects.

As to the remaining estimation results in Table 4, we find most coefficients not to be substantial or significant. There are only strong and negative scale effects of school size on all performance outcomes. Although we cannot qualify these effects as causal, these findings are in line with Dijkgraaf et al. (2009) who also find quality measures to decrease in the scale of schools.

4.2 Robustness Tests and Model Extensions

Similar to earlier work in this field of research, the validity of our RD design essentially hinges upon two assumptions: (i) exogeneity of the *Trouw* grades given the levels of the continuous performance indicators and the control variables; and (ii) the timing of the response to the *Trouw* publication, which is assumed to be unanticipated by school boards. In this subsection, we test for the robustness of these assumptions. As we stated earlier, the linearity assumption of the rating effects may be restrictive. Therefore we also examine a model version where the effects are specified as a step function of the Trouw ratings.

To start with, the exogeneity assumption states that discontinuities of R should be well identified and not correlated with other school track characteristics. One concern may be that test scores are noisy measures of school performance, causing the impacts of e.g. poor grades to be biased by regression to the mean effects. In principle, the RD avoids such effects if the spline function is sufficiently flexible and residuals are thus exogenous (Rockoff and Turner 2010). To test for this exogeneity assumption, we re-estimate Eq. (3) while using a Random Effects (RE) specification instead of Fixed Effects for school tracks. As such, we assume the unobserved, time constant school track characteristics to be uncorrelated to the rating residuals. If this assumption holds, the Random Effects specification would yield coefficient outcomes of the *Trouw* rating that are sufficiently similar to those from the Fixed Effects model. This is equivalent to a Hausman test on panel data models, but then applied to only one parameter. Table 5 shows coefficient estimates of the ranking grades that are not significantly different from the baseline specification, suggesting that the residual estimate is exogenous. This means that regression to the mean effects are adequately dealt with. Notice that only for the final exam grades we find the coefficient to become less significant.

The second assumption of our model is that school boards are not aware of their relative position in the first year after the realisation of the performance measures that serve as inputs into the *Trouw* grades. We assume that it takes at least 18 months for them to know their *Trouw* rating. Particularly school tracks that have performed in the

[8] The diploma received indicator is measured using the passing rates of all students in a particular calendar year. The junior years performance indicator however is measured following one cohort of students from their entry into the school up to their final school track in year three.

Table 5 Coefficient estimates of quality response: robustness checks and model extensions

	Diploma without delay	Grade final exams	Grade interim exams	Junior years perform
Baseline model				
Quality response coefficient $(t-2)$	−0.040***	−0.037**	−0.002	0.012
	(0.012)	(0.015)	(0.015)	(0.016)
Exogeneity test				
Random Effects specification	−0.027***	−0.022*	0.001	0.006
	(0.010)	(0.012)	(0.012)	(0.013)
Robustness check: Placebo test				
Quality response coefficient $(t-1)$	0.006	−0.019	0.001	−0.017
	(0.013)	(0.015)	(0.016)	(0.020)
Quality response coefficient $(t-2)$	−0.032**	−0.014	0.016	0.008
	(0.014)	(0.016)	(0.016)	(0.018)
Quality response coefficient $(t-3)$	−0.022*	−0.009	0.013	0.025
	(0.013)	(0.015)	(0.016)	(0.018)
Quality response coefficient $(t-4)$	−0.042***	−0.033*	0.002	0.012
	(0.012)	(0.015)	(0.015)	(0.018)
Step function				
Most negative ranking	0.121*	0.045	0.020	−0.056
	(0.071)	(0.078)	(0.089)	(0.101)
Negative ranking	0.045***	0.061***	0.007	−0.047**
	(0.016)	(0.020)	(0.021)	(0.023)
Neutral ranking (=reference group)	–	–	–	–
Positive ranking	−0.007	0.024	0.015	−0.005
	(0.015)	(0.018)	(0.019)	(0.021)
Most positive ranking	−0.060	−0.016	0.010	0.025
	(0.060)	(0.065)	(0.068)	(0.076)
Test (i): no effect negative rankings	$P = 0.011$	$P = 0.009$	$P = 0.926$	$P = 0.129$
Test (ii): no effect positive rankings	$P = 0.602$	$P = 0.351$	$P = 0.728$	$P = 0.898$

*, ** and *** denote significance at 10 %-5 %-1 %

lower tail of the quality distribution may be confronted with school inspections before the *Trouw* publication, so there may be 'shocks' that occur earlier. To test for the possibility that (some) schools were capable of anticipating their relative position earlier than that, we re-estimated Eq. (3) with a Placebo effect one year prior to the *Trouw* publication. At the same time, and in order to gain more insight in the persistency of school response effects in a broader sense, we also expanded the number of lags to four years.

Table 5 shows the resulting estimated patterns of responses to the *Trouw* score. Generally, while there is no significant one-year (Placebo) effect for all performance outcomes, we find significant persistency effects for the diploma indicator only. For the latter variable, it seems that responses to the *Trouw* grades lasted three years at least. As to the final exam grades, however, there is weak evidence that the effects that occurred two years after the *Trouw* publication were larger than the direct effects. Still, it seems that the flexible specification of persistency effects results in standard errors that are too large here to obtain efficient estimates.

Finally, we relaxed the (linearity) assumption that the effect of the *Trouw* publication is equal at all cut-off points. This means we re-specify the effect of R as a step function. As such, we cannot perform the two step model of Eqs. (1), (2) and (3), as this setup requires the effect of the *Trouw* rating to be linear, with the coefficient estimate $\hat{e}$ as our (only) variable of interest. We therefore follow Rockoff and Turner (2010), who use a step function for accountability grades to estimate the impact of the grades, while additionally including a spline function in one regression specification for the quality indicators. Within the context of our model, we thus re-estimate Eq. (3) with a step function of the *Trouw* score, as well as the expected value of the *Trouw* grade that follows from estimating Eq. (1) as an additional control variable.

The lower part of Table 1 displays the results that follow from this approach. Clearly, the effects for diplomas and final exam grades are confined to the most negative and negative rankings, suggesting that these schools are able to improve their performance to some extent. This lends credence to the idea that the negative scores can be qualified as 'wakeup calls' to schools. Again, note that these findings are in line with Rockoff and Turner (2010), who find significant effects for "F" and "D" school tracks only (compared to the "C" grade). To test for the robustness of these results, we further zoomed into the lower support of the distribution of *Trouw* grades. In particular, we clustered together the most negative and the negative scores, and only considered one single cut-off point for these (i.e. μ_2), using a first stage Probit model to obtain $\hat{e}$.[9] The results that follow from this approach are very similar to those of the benchmark model, suggesting that the extrapolation to other parts of the distribution of quality grades is not restrictive. In particular, we then find the coefficient estimate of μ_2 of the diploma scores, the grades of the final exams, the grades of the interim exams and junior years performance to be equal to 0.044 (0.016), 0.064 (0.019), 0.010 (0.020) and −0.047 (0.023), respectively.

4.3 Effects on Worker Staff

Within the context of NCLB in the U.S., one important driver of the response of schools quality indicators to quality information is the threat of sanctions, and the subsequent replacement of principals or cutbacks in staff (Chiang 2009). These changes may

[9] We also could have opted for considering the 'most negative' grade only; this also would yield estimates that are not significantly different from those of the benchmark model. In light of the limited number of schools that is classified in this category, however, the distinctive power of such a test is only limited.

Table 6 Coefficient estimates of (de-trended) worker staff and turnover responsiveness to quality scores on single school tracks (FE model)

	# Workers	Inflow rate	Outflow rate
Management	0.411**	−0.022	0.016
	(0.165)	(0.015)	(0.014)
Teachers	0.209	−0.005	0.005
	(0.570)	(0.010)	(0.010)
Support staff	−0.460	−0.011	0.007
	(0.440)	(0.011)	(0.011)

*, ** and *** denote significance at 10 %-5 %-1 %

increase the average quality of workers, and improve performance measures over time. With the data at hand, we can investigate whether the effect of the *Trouw* scores on quality indicators is driven by the same mechanism. In this subsection we examine the influence of the *Trouw* scores on the inflow and outflow rates of the management, teachers and the support staff of schools. In this respect, one may argue that bad rankings are followed by the hiring of new personnel, and the firing or quitting of incumbent workers. To estimate these effects, we estimate the baseline regression as in [3] on worker staff variables, with similar controls and school track fixed effects and with the corrected *Trouw* grade (i.e. e obtained from Eq. (2)). Since it is unlikely that changes in worker staff occur in the school year in which the *Trouw* publication is released, we now delay the *Trouw* rating with three years. We de-trend the number of workers per function level, but not the inflow and outflow levels. Furthermore, we restrict the analysis to schools with single school tracks, as worker turnover rates in our sample are measured at the level of (complete) schools.

Table 6 displays the coefficient estimates of the response to the *Trouw* scores that follow from these regressions, stratified according to outcome measure (number of workers, inflow rate and outflow rate) and worker types (management, teachers and supporting staff). As the table shows, we only find (some) evidence of responsiveness in worker staff for the management of single track schools. Particularly the effect on the number of managers is substantial, amounting to 5 % of the average number of managers in single track schools (about 10). It may well be that low performing schools hire additional managers. These results contrast to those that are obtained for teachers and supporting staff, for which all responsiveness coefficients are small and insignificant.

5 Discussion

The general picture that emerges from our analysis is that schools do respond to quality information by changing their quality outcomes on diplomas and final exam grades. With values of around 4 % of the standard deviation of performance for a one-step change in the rankings, the size of effects seems comparable to those in Rockoff and

Turner (2010) and Chiang (2009). Having said this, it should be stressed once more that the rating system analysed in this paper entailed a private initiative, without the threat of formal sanctions. As student numbers directly determine school funds and as student numbers are partly determined by the ranking (Koning and Van der Wiel 2012), it is to be expected that schools do respond, even without a formal sanction in place. Although one may question the adequacy and transparency of the *Trouw* ranking formula, this outlet receives more attention than the website of the Dutch Inspectorate of Education. 'Naming and shaming' can thus be a substitute for public interventions.

The outcomes of our analysis also broaden our knowledge of the functioning of rankings and accountability systems in another aspect, namely by explicitly addressing the persistency of accountability effects. Our results indicate that schools that receive a low score are triggered to improve their outcomes over longer time periods. Schools do not seem to be fully aware of their relative quality ranking and respond to information updates. For the sub-sample of schools that offer one school track, we furthermore find some evidence for schools that perform in the lower part of the quality distribution to respond by hiring additional managers.

References

Bacolod, M., DiNardo, J., & Jacobson, M. (2009). Beyond incentives: Do schools use accountability rewards productively? NBER working paper 14775.

Carnoy, M., & Loeb, S. (2003). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305–331.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9–10), 1045–1057.

Craig, S. G., Imberman, S. A., & Perdue, A. (2009). Does it pay to get an A? School Resource allocations in response to accountability ratings. Mimeo.

De Lange, M., & Dronkers, J. (2007). Hoe gelijkwaardig blijft het eindexamen tussen scholen in Nederland? EUI working paper 2007/03.

Dee, T.S., & Jacob, B. (2011). The impact of no child left behind on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418–446.

Dijkgraaf, E., Gradus, R. H. J. M., & de Jong, M. (2009). Competition and educational quality: Evidence from the Netherlands. Tinbergen institute discussion paper 2009-100/3.

Dijkstra, A. B., Karsten, S., Veenstra, R., & Visscher, A. J. (2001) Het oog der natie: Scholen op rapport. Standaarden voor de publicatie van schoolprestaties. Assen, Koninklijke van Gorcum.

Figlio, D. N., & Getzler, L. S. (2002). Accountability, ability and disability: Gaming the system, NBER Working Paper 9307.

Figlio, D. N., & Lucas, M. E. (2004). What's in a grade? School report cards and the housing market. *American Economic Review, 94*(3), 591–604.

Figlio, D. N., & Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools?. *Journal of Public Economics, 90*(1–2), 239–255.

Hanushek, E. A., & Raymond, M. E. (2004). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association, 2*(2–3), 406–415.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance?. *Journal of Policy Analysis and Management, 24*(2), 297–327.

Hastings, J., van Weelden, R., & Weinstein, J. (2008). Information, school choice and academic achievement: evidence from two experiments. *The Quarterly Journal of Economics, 124*(4), 1373–1414.

Jacob, B. (2005). Accountability, incentives and behavior: Evidence from school reform in Chicago. *Journal of Public Economics, 89*(5–6), 761–796.

Koning, P., & Van der Wiel, K. (2012, forthcoming). Ranking the schools: How quality information affects school choice in the Netherlands. *Journal of the European Economic Association*.

Pope, D. G. (2009). Reacting to rankings: Evidence from "America's best hospitals". *Journal of Health Economics, 28*(5), 1154–1165.

Rockoff, J., & Turner, L. J. (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy, 2*, 119–147.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. NBER working paper 13681.

Van der Steeg, M. & Vermeer, N. (2011). Nederlandse onderwijsprestaties in perspectief. *CPB Achtergronddocument*.