CPB Discussion Paper

No 149

May 2010

School Responsiveness to Quality Rankings

An Empirical Analysis of Secondary Education in the Netherlands

Pierre Koning and Karen van der Wiel

The responsibility for the contents of this CPB Discussion Paper remains with the author(s)

CPB Netherlands Bureau for Economic Policy Analysis
Van Stolkweg 14
P.O. Box 80510
2508 GM The Hague, the Netherlands

Telephone	+31 70 338 33 80
Telefax	+31 70 338 33 50
Internet	www.cpb.nl

ISBN 978-90-5833-457-2

Abstract in English

This paper analyzes the response of secondary schools to changes in quality ratings. In doing this, we contribute to the literature in two respects. First, the current analysis is the first to address the impact of quality scores that have been published by a newspaper (*Trouw*), rather than public interventions that aim to track and improve failing schools. Second, our research design exploits the substantial lags in the registration and publication of the *Trouw* scores and that takes into account all possible outcomes of the ratings, instead of the lowest category only. Overall, we find evidence that school quality performance does respond to *Trouw* quality scores. Both average grades increase and the number of diplomas go up after receiving a negative score. These responses cannot be attributed to gaming activities of the school board as an improvement is also observed in the gaming-proof quality indicators. For schools that receive the most negative ranking, the short-term effects (one year after a change in the ranking of schools) of quality transparency on final exam grades equal 10% to 30% of a standard deviation compared to the average of this variable. The estimated long run impacts are roughly equal to the short-term effects that are measured.

Key words: school quality, school accountability.
JEL code: H75, I20, D83.

Abstract in Dutch

Sinds 1997 publiceert het dagblad *Trouw* jaarlijks kwaliteitsinformatie over scholen in het voortgezet onderwijs (de 'Schoolprestaties'). *Trouw* baseert zich daarbij op informatie van de Onderwijsinspectie, over o.a. de eindexamenprestaties en de doorstroom van leerlingen in de onderbouw. Het voorliggende onderzoek laat zien dat scholen hun kwaliteit verbeteren na een slechte beoordeling door *Trouw*: zowel het gemiddelde eindexamencijfer als het aantal eindexamendiploma's nemen als gevolg hiervan toe. Dit effect geldt zowel op de korte termijn (één jaar na de *Trouw*- publicatie) als in de daarop volgende jaren.

Steekwoorden: schoolkwaliteit, informatietransparantie

Contents

Summary

Ranking and accountability have become increasingly common in the delivery of public services. One of the most prominent examples 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. By now, there is strong evidence that schools do respond to the NCLB accountability systems by improving their test scores and by changing the allocation of their. Part of the gains in measured performance can however also be attributed to gaming activities, typically by removing low-performing students from participation in exams. The general picture is that schools respond to accountability pressure by increasing their quality, particularly when the threat of sanctions is present. For studies on the overall effects of accountability systems, school response estimates range from 20 to 40% of the standard deviation of test scores.

This paper investigates the response of Dutch secondary schools to ranking scores, measured in terms of their overall test and diploma performance. For this purpose, we use a sample of 3,032 unique school tracks observed from 1996-2006. The quality rating system has been initiated in 1997 by the daily newspaper *Trouw*, so as to inform parents and their children on the quality of secondary schools. *Trouw* scores by school track are based on several objective quality indicators, such as the average grades in final centralized exams of the students, the percentage of students who obtain a diploma without delay, the percentage students who end up in a lower or higher school track than initially expected and some other quality indicators that differed from year to year. In order to obtain an indicator for value added by schools, the 'gross' quality score that follows from combining the three measures is corrected for the percentage of students with low parental income and from immigrant neighbourhoods. The exact control variables *Trouw* used changed from year to year, just as the weights attached to the quality indicators and the boundary values for the quality categories.

The primary interest in our analysis lies in the response of schools to changes in their ratings over time. As quality scores are predominantly driven by (lagged) test results and passing rates, the key question is whether schools that receive a negative quality score tend to improve these quality indicators. Likewise, we analyse the long-term effects on quality indicators for schools that receive a positive quality score. In doing this, we contribute to the literature in two respects. First, to the best of our knowledge, the current analysis is the first to address the impact of quality scores that have been published by a newspaper, rather than public interventions that aim to track and improve failing schools. The second contribution of this paper concerns the research design that is employed. As school quality indicators are the input of overall ranking scores, estimating the impact of ranking scores on future quality indicators raises endogeneity concerns. We therefore follow a research strategy that exploits the substantial lags in the registration and publication of the *Trouw* scores and that takes into account all possible

outcomes of the ratings, instead of the lowest category only. We also address the persistency of responses to ranking shocks. If e.g. the switch to the most negative quality score turns out to result in a permanent increase in performance, this lends credence to the idea that such a score functions as a 'wakeup call'.

The general picture that emerges from our analysis is that schools do respond to quality information by changing their quality outcomes in the short and in the long run. The size of long-term, permanent effects seems well in line with studies that evaluate the *No Child Left Behind* (NCLB) act that was enacted in US states, with values ranging from 10 to 30% of the standard deviation of performance outcomes. Having said this, it should be stressed once more that the ranking score system analyzed in this paper was initiated by the newspaper *Trouw*. So although the inputs of the *Trouw* rating were obtained from public authorities, the ranking system entailed a private initiative, without any threat of sanctions by the Inspectorate. The outcomes of our analysis also broadens 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 negative score are triggered to improve their long-term outcomes, with accountability incentives that can be qualified as 'ex post' — that is, after the occurrence of receiving a low ranking. Finally, our results indicate that the room and for use of gaming activities after the introduction of the ranking system was only small. Within the context of the current analysis, schools could only 'game' the diploma quality indicator (and therefore the *Trouw* score.) by increasing the interim grade scores. More generally, it seems that the quality indicators that the Inspectorate of Education gathers information on are relatively gaming-proof.

1 Introduction[1]

Ranking and accountability have become increasingly common in the delivery of public services. One of the most prominent examples 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 strong evidence that schools do respond to the NCLB accountability systems by improving their test scores (Carnoy and Loeb 2003; Hanushek and Raymond 2004; Jacob 2005; Dee and Jacob 2009) and by changing the allocation of their resources (Rouse et al. 2007; Craig et al. 2009; Chiang 2009; Bacolod et al. 2009).[2] Part of the gains in measured performance can however also be attributed to gaming activities, typically by removing low-performing students from participation in exams (Figlio and Getzler 2002; Jacob 2005). The general picture is that schools respond to accountability pressure by increasing average test scores of their students, particularly when the threat of sanctions is present. For studies on the overall effects of accountability systems, school response estimates range from 20 to 40% of the standard deviation of test scores (Hanushek and Raymond 2004; Dee and Jacob 2005). Estimates are generally smaller when authors focus on the specific impact of sanctions on failing schools (Figlio and Rouse 2006; Chiang 2009).

This paper investigates the response of Dutch secondary schools to ranking scores, measured in terms of their overall test and diploma performance. For this purpose, we use a sample of 3,032 unique school tracks observed from 1996-2006. The quality rating system has been initiated in 1997 by the daily newspaper *Trouw*, so as to inform parents and their children on the quality of secondary schools. *Trouw* scores by school track are based on several objective quality indicators, such as the average grades in final centralized exams of the students, the percentage of students who obtain a diploma without delay, the percentage students who end up in a lower or higher school track than initially expected and some other quality indicators that differed from year to year. In order to obtain an indicator for value added by schools, the 'gross' quality score that follows from combining the three measures is corrected for the percentage of students with low parental income and from immigrant neighbourhoods. The exact control variables *Trouw* used changed from year to year, just as the weights attached to the quality indicators and the boundary values for the quality categories.

[1] We would like to thank Joost Baeten, Lex Borghans, Klaas Bos, Elbert Dijkgraaf, Jaap Dronkers, Inge Groot, Myrthe de Jong, Cissy Pater, Sietzke Waslander and Floor Westendorp for useful comments and suggestions to earlier versions of the paper. The Ministry of Economic Affair is gratefully acknowledged for (partially) sponsoring the research project.

[2] In contrast, there is a limited literature on the effects of school quality information on school choice behaviour. 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 (2010) 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.

The primary interest in our analysis lies in the response of schools to changes in their ratings over time. As quality scores are predominantly driven by (lagged) test results and passing rates, the key question is whether schools that receive a negative quality score tend to improve these quality indicators. Likewise, we analyse the long-term effects on quality indicators for schools that receive a positive quality score. In doing this, we contribute to the literature in two respects.

First, to the best of our knowledge, the current analysis is the first to address the impact of quality scores that have been published by a newspaper, rather than public interventions that aim to track and improve failing schools. Until now, the literature on private initiatives by newspapers or magazines has predominantly focused on the hospital industry, like in Pope (2009) who analyzes the effects of the "America's Best Hospitals" publication of the US News and World Report (Pope 2009). The *Trouw* score addresses a broad range of performance outcomes, including schools that are confronted with the rare event of receiving the lowest and most negative ranking ('--') and schools that are awarded with the highest and most positive ranking category ('++'). Our outcomes are thus informative on the effectiveness of private initiatives to increase school quality transparency.

The second contribution of this paper concerns the research design that is employed. As school quality indicators are the input of overall ranking scores, estimating the impact of ranking scores on future quality indicators raises endogeneity concerns. The recent literature therefore usually employs regression discontinuity designs on the rating boundaries for school performance to estimate the impact of 'rating shocks' (Figlio and Rouse 2006; Craig et al. 2009; Chiang 2009). Discontinuity regressions are useful if the ratings follow from a sharp design, with full information on the relevant underlying quality indicators, their weights and the rating boundaries. Notable disadvantages however are that local average treatment estimates are based on limited supports and are usually confined to a small group of failing schools (Blundell and Dias 2009). Moreover, as the construction of quality scores is fully transparent, schools may try to avoid getting below threshold values, which in turn confounds the impact estimates. Within the context of the current analysis, however, the construction of the ratings is not fully transparent, with some quality variables and their corresponding weights being unobserved. We therefore follow an alternative research strategy that exploits the substantial lags in the registration and publication of the *Trouw* scores and that takes into account all possible outcomes of the ratings, instead of the lowest category only. More specifically, the registration of (all) the underlying quality indicators by the Dutch Inspectorate takes two years, and the subsequent processing by *Trouw* another six months. As a result, the size of endogeneity biases in impact estimates of *Trouw*-scores due to serial correlation in quality indicators is limited. Moreover, given the long time period that is under consideration, we can both estimate the short-term and long-term effects of changes in *Trouw*-scores. In doing this, we extend the recent analysis of Chiang (2009), who studies the medium-run effects of accountability pressure, that is, in the second and third year after the occurrence of sanctions.

Overall, we find evidence that school quality performance does respond to *Trouw* quality scores. Both average grades increase and the number of diplomas go up after receiving a negative score. These responses cannot be attributed to gaming activities of the school board as an improvement is also observed in the gaming-proof quality indicators. For schools that receive the most negative ranking, the short-term effects of quality transparency on final exam grades equal 10% to 30% of a standard deviation compared to the average of this variable. The estimated long run impacts are roughly equal to the short-term effects that are measured one year after a change in the ranking of schools. Moreover, it seems the strongest (positive) long-term effects occur at schools that receive the most negative ranking. This suggests that the most negative ranking works like a wake up call to schools. Reversely, schools with the most positive ranking feel less urgency to maintain high levels of quality.

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 Section 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 20,696 observations.[3] First, we have extracted information from the administrative records of the Inspectorate. These data include the number of plants per school group, school denomination,[4] student numbers and performance indicators per school track, like the average grade scores and average fractions of diplomas that were obtained.[5] It should be noted that there were major reforms for lower secondary (vocational) education in the Netherlands in 2002, causing the school track classification here to change and thus restricting the observed time period per stratum. We therefore restrict the sample to the three general education tracks that existed throughout the 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 there is the track that provides for 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 the major source of information on secondary schools.[6] As we will argue in the next section, the delays in the reporting system of *Trouw* enable us to identify the specific effect of this quality information, next to other sources of information. 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 school tracks in our (full) sample. Missing observations mostly stem from the fact that schools were considered too small to obtain a reliable overall quality score. There is no evidence that selection effects determine which observations are missing.[7] Although the ratings were based on information of the Inspectorate, these could not be inferred straightaway. We return to this issue later on.

[3] Thus, multiple observations per school originate from schools offering different school tracks.

[4] Within the Dutch school system, denominations include protestant schools, catholic schools, public schools and others (see Table 2.1).

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

[6] 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. Next to this, in 2001 the weekly magazine *Elsevier* started publishing similar rankings as *Trouw*, using quality levels that are averaged over three years and without using controls to obtain measures for value added. See Dijkstra et al. (2001) for more information on the Trouw outlet.

[7] We tested for selection effects by estimating a two step Heckman model. The first stage entailed a Probit regression on the occurrence of observing the *Trouw* score, and in the second stage we included the first stage Mills ratio to estimate the Trouw ranking scores. This did not yield significant parameter estimates for the Mills ratio.

Table 2.1 Summary statistics of school and school track data: full sample and balanced panel (1996-2006)

	Full sample (N=20,696)		Balanced panel (N=14,641)	
	Mean	Standard deviation	Mean	Standard Deviation
Share of School tracks				
Lowest general track (VMBO-gt)	0.418	(0.493)	0.379	(0.485)
Middle track (HAVO)	0.291	(0.454)	0.302	(0.459)
'Academic' track (VWO)	0.291	(0.454)	0.319	(0.466)
School tracks per school[a]				
1 school track	0.415	(0.492)	0.347	(0.476)
2 school tracks	0.146	(0.353)	0.153	(0.360)
3 school tracks	0.439	(0.496)	0.501	(0.500)
Market characteristics[a]				
Municipality population	126,646	(178,868)	127,127	(181,260)
Municipality population, aged 10-20	13,866	(18,268)	13,836	(18,448)
Number of schools in municipality	9.936	(12.904)	9.971	(13.061)
School characteristics[a]				
Denomination: protestant	0.225	(0.418)	0.234	(0.424)
Denomination: Catholic	0.269	(0.443)	0.276	(0.447)
Public schools	0.266	(0.442)	0.284	(0.451)
Denomination: other	0.240	(0.427)	0.206	(0.404)
Number of students per school track	202.7	(132.3)	228.8	(128.3)
Number of students per school[a]	2,014.4	(1272.8)	1,864.4	(1,119.4)
Inflow new students per school[a]	181.3	(102.8)	193.8	(101.8)
School performance				
Diploma without delay (%)	70.929	(16.572)	70.087	(16.118)
Grade final exams	6.346	(0.289)	6.360	(0.271)
Grade interim exams	6.602	(0.282)	6.609	(0.275)
Junior years performance (first to third class)	100.071	(9.721)	99.972	(8.903)
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 the first two years in our sample (1996 and 1997). Moreover, in subsequent years on average about 14% of the yearly observations per school track is missing.

Table 2.1 presents summary statistics for the selected sample of secondary schools in 1996-2006, both for the full sample and the sample of schools that are observed over the full time period (i.e. the balanced panel). It should be noted here that the full sample that is presented in the table exceeds the sample that can be used to estimate the actual impact of the quality scores which starts in 1999, as this requires a lag of three years.[8] In the full sample we have on average 6.5 yearly observations per combination of school and school track, with 11 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 underprivileged (in Dutch: 'cumi'-students) as students living at zip codes with a relatively high fraction of ethnic minorities.[9] Schools also receive additional funding for each underprivileged student.

The Dutch Inspectorate of Education monitors school quality with a set of three indicators that also are inputs for the *Trouw* rating. The first indicator is the average percentage of students that leaves the school with a diploma without any delay, measured from the third year onwards (with an average per school equal to 70%). This means that there is no room for schools to game their results by excluding low-performing students from final exams in the last year. Second, the Inspectorate 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, centralized 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 organized and the correction is carried out by teachers at other schools. The average test score at the final exams equals 6.4 (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, thus increasing the odds of passing the final exams at the end of the school year. Third, the Inspectorate measures the net percentage of students in third year that are in a school track that is either below what the child's primary school had advised, or above. This 'junior-years performance' is documented as schools could otherwise game their results by forcing students into lower school tracks. A score of 100% indicated that on average students are in their predicted school track.

We stated earlier that the *Trouw* ranking scores cannot be recovered from the performance indicators that are provided to us by the Inspectorate. This is partly because the Inspectorate provides more detailed information to *Trouw* than to us and partly because *Trouw* has adapted their scoring method from year to year. For both ourselves and *Trouw* it was impossible to

[8] This causes the sample size to reduce to 15,201 observations. Moreover, as we do not observe all quality scores in these years, the effective sample size in our regressions is 12,451.

[9] It should be noted that the definition of cumi-students has changed in 2003 and in 2005. The average value of this variable is therefore not presented in Table 2.1. In the estimation of our models, we therefore control for this variable by allowing its impact to vary from year to year.

reconstruct this method, particularly as journalist turnover rates were high over the years. We only know that the three objective quality indicators were recurrent inputs for the rankings score that followed from clustering analysis. Moreover, in an attempt to control for the 'quality' of students *Trouw* corrects the overall score in all years for the fraction of students from predominantly immigrant neighbourhoods. For some years these variables were supplemented with additional control variables, particularly on parental income, so as to obtain more accurate measurements for value-added by schools. We also know that for some years the percentage of students retaking classes was also taken into account as an additional quality indicator. As a result of these seemingly random changes in the calculation procedures, school boards were thus uncertain about how their quality performance, which they did observe in advance, would affect their overall *Trouw* quality rankings.[10]

Table 2.2 Quality indicators per ranking score (full sample on school-track level)

	Most negative ('--')	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

Table 2.1 makes apparent 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 2.2 mirrors the relation between the quality indicators and the resulting quality scores. The spread between the diploma rates is substantial, with 49% for schools in the most negative category and 81% for schools in the most positive category. The relation 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 2.3 shows the dynamics of the *Trouw* ratings per school, measured as year-to-year transition probabilities. 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, whereas 49% moves to the middle category or

[10] Ordered Probit estimation of the *Trouw* ranking scores with the observed quality indicators and the fraction of cumi-students (with time varying coefficients), explains about 70% of the observed variance. With constant weights, the explained variance is 61%. Thus, it seems ranking scores are to a large extent driven by variables other than the observed ones. Moreover, the weights that were attached to the observed quality and controls variables vary over the years.

higher than that. The extreme event of receiving the most negative or most positive rankings is thus largely transitory (see also Dijkstra et al. 2001).

Table 2.3 Transition probabilities between ranking scores (1998-2006); rows = origins, columns = destinations.

	Unknown	Most negative	Negative	Average	Positive	Most positive
Unknown	65.7	0.5	7.1	19.7	6.5	0.6
Most negative ('--')	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

3 Empirical strategy

The baseline model

A common assumption that underlies most studies on the response to school rankings is that schools are ex-ante incompletely informed on their overall ranking position. Within the context of the Dutch ranking system, this means that the computation process by *Trouw* differs from one year to the other and is not known by school boards. Schools also are not aware of their relative position vis-à-vis the other schools with the same school track, with small differences in the overall relative latent performance outcome having potentially strong consequences for their rankings. The event of receiving a low quality score may therefore increase the awareness of schools of their relative quality level and trigger them to change their policies. In our analysis, we argue that such changes in ratings are unanticipated ('rating shocks'), which enables us to obtain consistent estimates of the effects of *Trouw* scores on quality measures. The ranking scores of *Trouw* are also reported with a lag of three years, rendering it likely that endogeneity effects due to serial correlation in quality measures are only small.

The baseline specification we use for quality indicator Q^k (k = 1,..K) measured for school i (i = 1 ,.., I) with track j (j=1,...,J) at time t (t = 1 ,.., T) is:

$$(1) \qquad Q^k{}_{ijt} \quad = \quad \alpha^k R_{ij,t-3} + X_{it}\beta^k + v^k{}_{ij} + \varepsilon^k{}_{ijt} ,$$

with the diploma received percentages, the final and interim exam scores and the junior years performance percentages as the four quality outcome measures under investigation (K = 4). Matrix $\boldsymbol{X}$ includes the time varying municipality and school characteristics that are, amongst others, presented in Table 2.1, together with yearly time dummies. R indicates the ranking category the school track receives (R = 1,..5), with an impact coefficient of α^k for quality indicator k. As we have shown earlier, *Trouw* labels these five ordinal measures as '--', '-', '0', '+' and '++', respectively. In the baseline specification, we furthermore start by assuming that the impact of one higher ranking category is equal for all categories.[11] Vector v^k indicates school track fixed effects per quality indicator k. The relevant stratum we use here is that of school tracks, which are indexed as combinations of i and j. Finally, ε represents residuals that are assumed to be identically and independently distributed with mean zero and variance $\sigma^2{}_k$ for each quality variable. The baseline equation (1) is estimated with school track fixed effects, where standard errors are corrected for clustering effects at the level of school tracks.

[11] This corresponds to the linearity assumption of Pope (2009) who studies the effects of rankings on the number of hospital clients.

Identification and robustness

The key challenge in estimating the impact of rankings is that time demeaned values of *Trouw* scores are correlated with time demeaned school quality measures. Time demeaning follows from using school track fixed effects. Given the limited number of time observations per school track, this strategy is likely to yield inconsistent estimates of our parameter of interest α (see e.g. Wooldridge 2002, pp.270).[12] To illustrate this point, suppose we focus on school tracks with high absence rates of teachers in one particular year, which is an omitted variable that causes the residual terms in our equation to be low. This school therefore receives the most negative ranking at time *t*. Given the limited time span school track is observed, the low residuals will be partially misperceived as low school track fixed effects. Conditional on the low score at time *t*, the expected values of the other year observations will thus be higher than the fixed effect estimate. More generally, high (low) quality scores for schools with high (low) ranking scores are partially misperceived as high (low) fixed effects, with the remaining variation unjustly attributed to the ratings. Ranking responses will be thus biased from zero. In the literature, this effect is often referred to as the 'mean reversion bias' (see also Chiang 2009).

With the data at hand, we can easily infer the size of this (negative) time demeaning bias. For this purpose, we first specify the ranking score R_{ijt} for school track *ij* at time *t* as

$$(2) \qquad R_{ijt} \quad = \quad \Sigma^K \ \gamma^k \ Q^k_{\ ijt} \ + \ Z_{it}\, \eta_t^{\ k} \ + \ \upsilon_{ij} + \ \psi_{ijt} \ .$$

with γ^k indicating the approximate weight of quality indicator *k* in the ranking and ***Z*** as a matrix including yearly dummies and the fraction of underprivileged students for school *i* at time *t*. Note that the impact of the fraction of underprivileged is allowed to vary over time. υ indicates school track fixed effects and ψ represents residuals that are assumed to be identically and independently distributed with mean zero and variance $\sigma^2_{\ \psi}$.

When defining $Q^k_{\ ij}$ and $\varepsilon^k_{\ ij}$ as the school track time demeaned values of the quality measure *k* and the residual terms of equation (1), respectively, we can show that the coefficient estimate of α^k will have a bias that is equal to

$$(3) \qquad \mathrm{E}\,(\,R_{ij,t\text{-}3} - R_{ij}\,)\,(\,\varepsilon^k_{\ ijt} - \varepsilon^k_{\ ij}\,) \quad = \quad \gamma^k\ \mathrm{E}\,(\,Q^k_{\ ij,t\text{-}3} - Q^k_{\ ij}\,)\,(\,\varepsilon^k_{\ ijt} - \varepsilon^k_{\ ij}\,) \quad =$$

$$= \quad \gamma^k\ \mathrm{E}\,(\,\varepsilon^k_{\ ij,t\text{-}3} - \varepsilon^k_{\ ij}\,)\,(\,\varepsilon^k_{\ ijt} - \varepsilon^k_{\ ij}\,) \quad = \quad -\gamma^k\,\sigma^2_{\ k}\ /\ T\ ,$$

for $k = 1,..K$, with $.R_{ij}$ indicating the mean values of the ranking scores, $\varepsilon^k_{\ ij}$ representing the residual terms and $Q^k_{\ ij}$ reflecting the quality measures per school track *ij* over time. The

[12] In particular, for fixed effects estimation it is well known that the correlation coefficient of residuals in equation (1) will be equal to –1/(T–1) (Wooldridge 2002).

equation makes apparent that the bias in the coefficient estimate if α^k is determined by the weight of the indicator in the ranking score, the variance of the quality indicator, and the time span covered by the data.

We propose three research strategies to address the time demeaning bias in estimation equation (1). Our first research strategy entails the calculation of the time demeaning bias itself. This means we first perform a fixed effects estimation of both equation (1) of all quality measures k (k = 1,..K) and of equation (2) with school track fixed effects. In doing this, we obtain coefficient estimates of γ^k and σ^2_k (k = 1,..K) that are necessary to calculate the bias presented in equation (3). Next, the bias estimate is compared to and subtracted from the value estimate of α^k that is obtained from direct estimation of (1).

Our second and third strategy draw further upon the idea that the bias in equation (3) originates from the correlation between the time demeaned value of the residual terms of quality measure k, measured at time t and that time t-3. In order to control for this, the second strategy enriches the equation (1) with $Q^k_{ij,t-3}$ as and additional explanatory variable. The coefficient estimate of this variable then accounts for the spurious serial correlation that follows from time demeaning. Obviously, the coefficient estimate will itself be inconsistent, but the biasing effect on γ^k is controlled for.[13] The third robustness strategy also uses a lagged variable approach to remove the time demeaning bias, but now with the 'gross' *Trouw* score as control variable. Recall from the previous section that this ranking score follows from the weighted average of performance outcomes, but without controlling for student characteristics. Including the gross *Trouw* score as an additional control variable thus results in quality response estimates that are identified from variation in the *Trouw* rankings that originate from the conversion from 'gross' measures to 'net' or value added measures. As a consequence, we control for alternative sources of information that may be correlated with the gross *Trouw* score.

Highest and lowest rankings: short and long-term effects

It may well be that the quality rankings have a longer lasting impact on quality performance than just in the year after the announcement of new *Trouw* scores (see also Chiang 2009). Changes in policies and investments usually take longer than one year to be completed, which calls for an estimation approach that would incorporate the possibility of more persistent quality effects. To analyze the persistency of all rankings scores jointly is however cumbersome, as school tracks usually receive a sequence of rankings that are either negative, average or positive. It thus would be unclear how such a sequence of ranking scores would drive

[13] Including the lagged value of the quality measure also controls for biases resulting from 'true' serial correlation in the error terms. To illustrate this, suppose the residuals in equation (1) follow an autoregressive process with a parameter value ρ_k for k = 1,..K. As a result, the bias in the fixed effects coefficient estimate of γ_k will consist of a time demeaning bias and a bias due to 'true' serial correlation: $\gamma^k \sigma^2_k\ [\,-1 / \{(1-\rho_k) T\} + \rho_k^3 / (1-\rho_k^2)\,]$. By including the (three period) lagged value of the quality measure k, both the inconsistency due to demeaning and that of 'true' serial correlation are controlled for.

permanent changes in quality. By contrast, the extreme events of receiving the most negative or most positive rankings are much more clear-cut and thus more informative in this respect. More specifically, in the sample about 6% (4%) of the school tracks have received the most negative (positive) ranking in the time period under investigation without any overlap between these two 'extreme' categories. We will therefore analyse the persistency of school responses after they enter into the most negative category ('--') or the most positive category ('++'). We thus extend equation (1) with dummy values that equal one from the moment the event of receiving the school receives the most negative or most positive ranking. The coefficients of these dummies can be scaled (i.e. divided by two[14]) so as to obtain value estimates that can be compared with those for the short-term ranking responses.

[14] Note that both the most negative and most positive ranking score differ from the (average) reference group by two categories. Thus, when comparing this result to the short-term effects, the coefficient estimate of the permanent response should be divided by two.

4 Estimation results

The baseline model

Table 4.1 presents coefficient estimates of our baseline model for the four quality measures, i.e. the diplomas without delay, the grades of final and interim exams and the junior year's performance. Overall, we find that higher (lower) *Trouw* ranking scores lower (increase) the percentage of students receiving a diploma and the average grades of students, with values equal to about 5% of the standard deviations of the respective scores. As the most negative (positive) ranking is two positions below (higher than) the average, this means that school tracks in this category improve (worsen) their quality performance with about 10% of the standard deviation. We thus conclude that school tracks with high rankings tend to lower their efforts and those schools with low rankings are triggered to improve their performance. In contrast, we find no significant effect for the junior years performance score. Presumably this can be explained by the fact that this score measures performance over three consecutive school years, whereas the ranking responses are measured as one-year, transitory effects. It also should be noted that the ranking score estimate in the interim exam regression is lower than that in the final exam regression. As the centralized 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. In particular, 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.

As to the remaining estimation results in Table 4.1, an important finding is that 60 to 75% of the unexplained variance is attributed to school track fixed effects. As these effects are positively correlated with the ranking scores, excluding school track fixed effects would yield response coefficients that are biased upwards.[15] Furthermore, the school quality measures are sometimes lower in municipalities with many school tracks and with smaller schools. Although we cannot qualify these effects as causal in this context, these findings are in line with Dijkgraaf et al. (2009) who also find quality measures to decrease in the scale of schools. Finally, for three of the quality measures the yearly time dummies reveal an upward trend. This is not the case for the final exam scores, which are probably less prone to gaming.

[15] This is confirmed when estimating equation (1) for the quality outcomes without school track fixed effects.

Table 4.1 School track fixed effects estimation of quality measures (1999-2006)[a,b]; standard errors corrected for school track clustering effects; *, and *** denote significance at 10%-5%-1%.**

	Diploma without delay	Grade final exams (x10)	Grade interim exams (x10)	Junior years performance
Ranking response (*t*-3)	– 0.819***	– 0.162***	– 0.070**	0.076
	(0.145)	(0.031)	(0.032)	(0.172)
Municipality population aged 10-20, log value	1.231	0.347**	0.158	0.870
	(0.846)	(0.164)	(0.182)	(0.923)
# Schools in municipality	– 0.257**	– 0.089***	0.014	– 0.063
	(0.103)	(0.022)	(0.028)	(0.135)
# School tracks per school	– 0.274	– 0.037	– 0.050	0.387**
	(0.196)	(0.040)	(0.046)	(0.242)
# Students per school track, log value	– 3.124***	– 0.486**	– 0.229	– 5.193***
	(0.903)	(0.192)	(0.191)	(1.213)
Year = 2000	0.148	0.316***	0.150**	–0.484
	(0.360)	(0.079)	(0.068)	(0.405)
Year = 2001	2.839***	– 0.013	0.446***	–1.117***
	(0.387)	(0.080)	(0.078)	(0.485)
Year = 2002	5.562***	– 0.030	0.635***	0.336
	(0.385)	(0.081)	(0.082)	(0.451)
Year = 2003	7.068***	– 0.049	0.659***	3.610***
	(0.385)	(0.085)	(0.083)	(0.534)
Year = 2004	7.812***	– 0.609***	0.753***	3.954***
	(0.417)	(0.088)	(0.084)	(0.500)
Year = 2005	6.560***	– 0.670***	0.841***	4.757***
	(0.406)	(0.095)	(0.088)	(0.525)
Year = 2006	6.900***	–0.939***	0.652***	5.998***
	(0.421)	(0.090)	(0.092)	(0.533)
Variance (σ_k)	8.043	1.643	1.708	6.903
Fraction variance due to FE	0.741	0.676	0.671	0.605
R-squared	0.089	0.058	0.004	0.028

[a] We also included the fraction of underprivileged students as controls in the regressions. As the definition of this variable changed during the period under investigation, the effect of this variable was allowed to vary per year.

[b] We also have estimated the baseline model for subsamples 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.

Robustness checks

We argued earlier that three research strategies can be followed to test for the robustness of the baseline model: (i) calculating the time demeaning bias; (ii) re-estimating equation (1) with the three-year lagged variable of the quality outcome as an additional control variable; of (iii) re-estimation equation (1) with the 'gross' *Trouw* score as an additional control variable. Table 2.2 shows the outcomes that follow from these strategies. Generally, the estimated biases appear small and do not change our result of ranking scores affecting the diploma and interim and final exam scores. Note that the time demeaning bias for the interim exam grades is almost close to zero, as there is no weight attached to this variable in the *Trouw* score. The finding that any inconsistencies due to — true or spurious — serial correlation are small is confirmed in the second and third estimation strategies. That is, the coefficient estimates for diploma and grade scores become somewhat smaller and are close to the corrected coefficient estimates that follow from the first estimation strategy. It is only for the junior years' performance measure that both robustness checks change our findings. When controlling for the time demeaning bias here, weak evidence emerges that higher rankings increase the junior year's performance.

Table 4.2 Coefficient estimates of quality response: robustness checks

	Diploma without delay	Grade final exams (x10)	Grade interim exams (x10)	Junior years performance
Baseline model				
Quality response coefficient	– 0.819***	– 0.162***	– 0.070**	0.076
	(0.145)	(0.031)	(0.032)	(0.172)
Robustness check (i): bias calculation				
FE estimate quality on ranking ('weight')	0.026***	0.101***	– 0.002	0.031***
	(0.001)	(0.003)	(0.003)	(0.001)
Implied demeaning bias[a]	– 0.290	– 0.047	0.001	– 0.255
	(0.000)	(0.000)	(0.000)	(0.000)
Corrected quality response coefficient	– 0.529***	– 0.115***	– 0.069**	0.311*
	(0.145)	(0.031)	(0.032)	(0.172)
Robustness check (ii): controlling for lagged values				
Coefficient quality, *t*-3	– 0.062***	– 0.027**	– 0.046***	– 0.122***
	(0.013)	(0.012)	(0.011)	(0.028)
Corrected quality response coefficient	– 0.405**	– 0.130***	– 0.052	0.478*
	(0.165)	(0.034)	(0.032)	(0.262)
Robustness check (iii): controlling for gross Trouw score				
'Gross' *Trouw* score, *t*-3	– 0.738***	– 0.142***	0.002	– 0.033
	(0.182)	(0.036)	(0.036)	(0.234)
Corrected quality response coefficient	– 0.368*	– 0.117***	– 0.048	– 0.101
	(0.217)	(0.043)	(0.047)	(0.282)

[a] Standard errors for the time demeaning biases were obtained using the Delta method.

Short- and long-term effects

Table 4.3 presents the estimation results of school responses to the most positive or most negative quality rankings, with the distinction between short-term and long-term effects. As argued in the previous section, we add dummy variables capturing persistency effects after the occurrence of receiving the most negative or the most positive ranking. In contrast to the baseline model with linear effects, this also allows us to address the possibility of asymmetry in the effects of these rankings. When conducting the specification, the reference group exists of the pooled sample of school tracks with average, positive and negative rankings; this group gradually decreases to about 90% of the sample in 2006. From the table, we infer the following findings. First, although the (scaled) short-term coefficient values of the responses to extreme rankings tend to be stronger than in the baseline model, differences are small and insignificant. Aggregating the positive and negative ranking categories to the (average) reference group thus does not affect our short-term response estimates substantially. Although the smaller treatment group of schools with the most negative or positive ranking decreases the efficiency of our estimates, the estimates are remain significant in most cases.

Our second general finding is that the size of (negative) quality response estimates increases when distinguishing between short and long-term impacts. This suggests that in the model with short-term effects only, permanent changes in quality outcomes are partially absorbed in the school track fixed effects, causing the short-term impact to be underestimated. Thus, estimates for the short-term impact in the baseline model are biased towards zero. In the model with long-term effects, we find coefficient estimates of both the short and long-term impacts to be negative for all performance measures and significant in most cases as well. The long-term impacts are somewhat stronger for the diploma and final exam grade scores. This is in line with Chiang (2009), who finds short and medium-term effects to be roughly of equal size. For the diploma and final exam grade scores the estimated short and long-term impacts vary between 25% and 35% of their respective standard deviations, whereas the interim exam score is about 10% of the standard deviation (both in the short and in the long term). For the junior year's performance, we only find the long-term response estimate to be significant and equal to about 20% of its standard deviation.

Finally, the lower part of Table 4.3 shows that both short and long-term quality responses are strongest for school tracks receiving the most negative ranking. For instance, the permanent increase in final exam grades for school tracks receiving the most negative ranking equals 1.14 compared to the reference group (= -/- 2 × -/- 0.57), whereas the decrease in final exam grades for school tracks receiving the most positive ranking equals 0.64 (= 2 × 0.32). We thus conclude that responses to 'rating shocks' are not symmetric.

Table 4.3 Coefficient estimates of quality responses: short and long-term impacts based on 'most positive' and 'most negative' rankings (with '0', '+' and '-' in reference group).

	Diploma without delay	Grade final exams (x10)	Grade interim exams (x10)	Junior years performance
Baseline model: implied short-term quality response of most positive ranking	–1.638***	– 0.324***	– 0.140**	0.152
	(0.290)	(0.062)	(0.064)	(0.344)
Specification with short and long-term effects (scaled)				
Short-term impact	– 2.060***	– 0.338***	– 0.236***	– 0.496
	(0.289)	(0.072)	(0.058)	(0.378)
Long-term impact	– 2.767***	– 0.481***	– 0.260***	– 0.932*
	(0.278)	(0.071)	(0.068)	(0.520)
Specification with short- and long-term as well as separate effects for 'most negative' and 'most positive' rankings				
Short-term impact of most negative ranking	2.119***	0.450***	0.290***	0.518
	(0.390)	(0.096)	(0.079)	(0.513)
Short-term impact of most positive ranking	–1.852***	– 0.154	– 0.149*	– 0.473
	(0.419)	(0.101)	(0.094)	(0.585)
Long-term impact of most negative ranking	3.366***	0.572***	0.359***	1.003
	(0.359)	(0.092)	(0.084)	(0.718)
Long-term impact of most positive ranking	–1.680***	– 0.318***	– 0.078	– 0.818
	(0.443)	(0.113)	(0.117)	(0.750)

5 Discussion

The general picture that emerges from our analysis is that schools do respond to quality information by changing their quality outcomes in the short and in the long run. The size of long-term effects seems in line with studies that evaluate the *No Child Left Behind* (NCLB) act that was enacted in US states, with values ranging from 10 to 30% of the standard deviation of performance outcomes. Having said this, it should be stressed once more that the ranking score system analyzed in this paper was initiated by the newspaper *Trouw*. So although the inputs of the *Trouw* rating were obtained from public authorities, the ranking system entailed a private initiative, without any threat of sanctions by the Inspectorate. Obviously, one may question the adequacy and transparency of the *Trouw* ranking formula, but it appears that 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, with exit and voice as its driving mechanisms.

The outcomes of our analysis also broadens 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 negative score are triggered to improve their outcomes over longer time periods, with accountability incentives that can be qualified as 'ex post' — that is, after the occurrence of receiving a low ranking. This contrasts to a situation where (all) schools would be fully informed on their relative performance and where the incentives of accountability would be set ex ante. Our results suggest that schools are not fully aware of their relative quality ranking instead, and respond information updates.

Finally, our results indicate that the room and for use of gaming activities after the introduction of the ranking system was only small. Within the context of the current analysis, schools could only 'game' the diploma quality indicator (and therefore the *Trouw* score.) by increasing the interim grade scores. More generally, it seems that the quality indicators that the Inspectorate of Education gathers information on are relatively gaming-proof.

References

Bacolod, M., J. DiNardo and M. Jacobson, 2009, Beyond incentives: Do schools use accountability rewards productively?, NBER Working Paper 14775.

Blundell, R., and M.C. Dias, 2009, Alternative Approaches to Evaluation in Empirical Microeconomics, *The Journal of Human Resources*, 44(3), 565-640.

Canoy, M, and S. Loeb, 2002, 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.

Chorny, V., and D. Webbink, 2010, The effect of accountability policies in primary education in Amsterdam, CPB Discussion Paper 144.

Craig, S.G., S.A. Imberman and A. Perdue, 2009, Does it Pay to Get an A? School Resource Allocations in Response to Accountability Ratings, mimeo.

Cullen, J.B. and R. Rebeck, 2006, Tinkering towards accolades: school gaming under a performance accountability system, NBER Working Paper 12286.

Dee, T.S. and B. Jacob, 2009, The impact of No Child Left Behind on student achievement, NBER Working Paper 15531.

Dijkgraaf, E., R.H.J.M. Gradus and M. de Jong, 2009, Competition and Educational Quality: Evidence from the Netherlands, Tinbergen Institute Discussion Paper 2009-100/3.

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

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

Hanushek, E.A. and M.E. Raymond, 2004, Does School Accountability Lead to Improved Student Performance?, mimeo.

Hanushek, E.A. and M.E. Raymond, 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.

Koning, P., Experience rating and the Inflow into Disability Insurance, *De Economist*, 157(3), 315-335.

Koning, P. and K. van der Wiel, 2010, Ranking the Schools: How Quality Information Affects School Choice in the Netherlands, CPB Discussion Paper 150.

Lange, M. de, and J. Dronkers, 2007, Hoe gelijkwaardig blijft het eindexamen tussen scholen in Nederland? Discrepanties tussen cijfers voor het schoolonderzoek en het centraal examen in het voortgezet onderwijs tussen 1998 en 2005, EUI Working Paper 2007/03.

Luginbuhl, R., D. Webbink and I. de Wolf, 2007, Do School Inspections Improve Primary School Performance?, CPB Discussion Paper 83.

Pope, D.G., 2009, Reacting to rankings: Evidence from “America’s Best Hospitals”, *Journal of Health Economics*, 28(5), 1154-1165.

Reback, R., 2008, Teaching to the rating: School accountability and the distribution of student achievement, *Journal of Public Economics*, 92(5-6), 1394-1415.

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