I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13450

Should School-Level Results of National Assessments Be Made Public?

Atsuyoshi Morozumi
Ryuichi Tanaka

JULY 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13450

Should School-Level Results of National Assessments Be Made Public?

Atsuyoshi Morozumi
University of Nottingham

Ryuichi Tanaka
University of Tokyo and IZA

JULY 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Should School-Level Results of National Assessments Be Made Public?*

Many countries conduct national standardized assessments of educational performance, the results of which may be published at the school level or at a higher level of aggregation. Publication at the school level potentially improves student achievements by holding schools accountable, whereas such accountability pressure may have distributional consequences and/or compromise outcomes beyond education achievements (labeled as non-cognitive skills). Using a Japanese policy reform that created variation in the disclosure system of national assessment results across municipalities, we show that publishing school-level results increases students' test scores across the entire score distribution, with no evidence of adverse impacts on noncognitive skills.

JEL Classification: D80, I20, I28

Keywords: national standardized assessments, information disclosure, school-level results, school accountability, student outcomes

Corresponding author:
Ryuichi Tanaka
Institute of Social Science
University of Tokyo
7-3-1 Hongo Bunkyo-ku
Tokyo 113-0033
Japan
E-mail: ryuichi.tanaka@iss.u-tokyo.ac.jp

* We thank Hideo Akabayashi, Michael Bleaney, David Figlio, Sourafel Girma, Richard Upward, and Miguel Urquiola, and participants at several seminars and conferences for their comments. We also thank Yumiko Mitarai for her assistance in data collection. Any remaining errors are our own responsibility.

1 Introduction

National standardized assessments of educational performance are common across the world. OECD (2015) report that 32 out of 38 OECD (and partner) countries conducted national assessments at the primary education level in 2015. National assessments are a type of standardized student achievements tests which do *not* affect students' progression through school or certification.[1] The main purposes of these assessments at the primary level are to evaluate school performance, provide teachers with student diagnostic information, and provide formative feedback to parents (OECD (2015)). While conducted in many countries, the design of assessments varies substantially across them in various aspects. The particular issue that we examine here is whether the results of national assessments are published at the *school-level* or at a higher level of aggregation.[2]

While national assessment results could be published at the level of country/state, province, municipality, school, or even student, whether school-level results should be made public has been debated extensively.[3] On the one hand, schools held accountable for certain published performance measures which are comparable across schools (e.g., school-mean scores, or proportion of students who achieve threshold scores) may be incentivized to be more efficient to improve those measures. On the other hand, such accountability pressure may induce schools to allocate resources unequally to students to improve the published measures, potentially leaving the disadvantaged students behind. Further, schools under such pressure may be tempted to disregard outcomes other than assessment scores, and even to cheat by manipulating the testing pool through student exclusions.

The present paper investigates these issues in the context of a Japanese policy reform in 2014. In Japan, a national assessment of 6th- and 9th-grade students known as the National Assessment of Academic Ability (NAAA) has been implemented every April since 2007.[4] In 2014, the central government implemented a *reform* on the information disclosure system of the assessment results,

[1] As another type of a standardized test, OECD (2015) define national *examinations* as the ones which have a formal consequence for students, such as an impact on a student's eligibility to progress to a higher education level or to complete an officially-recognized degree.

[2] The OECD (2015) shows that, at the primary education level, 14 of the OECD countries (e.g., Australia, Chile, Netherlands) published school-level results of national assessments.

[3] See Rosenkvist (2010) for a useful survey.

[4] The NAAA did not take place in 2011, the year in which the Great East Japan Earthquake occurred.

whereby education policymakers in all the municipalities were granted discretion to publish the results of schools under their jurisdiction. This reform led to the situation in which some municipalities made use of this discretion and some did not in the same round of assessments. Notably, the performance measure published after the reform was *school-mean test scores*. This paper examines the possible trade-offs of school accountability shaped by the disclosure of school-mean test scores, by using the variation in the disclosure system across municipalities, before and after the reform via a differences-in-differences (DD) approach.

The key identifying assumption to estimate the effect of disclosure of school-mean test scores on a student outcome is that the outcome in treatment and control group would follow the same trend in the absence of treatment. Our basic justification for this assumption is that a decision of disclosing the school-level results is made primarily by education policymakers of a municipality who are given the discretion, exogenous to students and schools. However, identification is threatened when changes to other policies and institutions within municipalities that affect education outcomes are correlated with treatment status. Indeed, this is a concern to the extent that the test-based accountability system was introduced as part of *broader institutional reform* effort.[5] To address this, we isolate the effects of information disclosure by controlling for education expenditures as well as institutional features at the municipality level, such as parental school choice, school inspections, and school autonomy. We further control for school fixed effects which reflect the effect of time-invariant municipality-specific unobserved factors.

We show that the disclosure of school-mean test scores of the NAAA *increased* the test scores. This result holds regardless of the subjects: Japanese language and Mathematics. Inspecting how the information disclosure changed schools' behavior, we find that schools in treated municipalities explained their assessment performance more to students' parents and local residents, and made a better use of diagnostic information from previous assessments to improve their teaching quality. There is no evidence that those schools removed low-performing students from the testing pool to manipulate test results. Further, we show that the information disclosure increased test scores

[5] Figlio and Loeb (2011) emphasize that the possible introduction of a multi-faceted system reform all at once is a particular challenge to measure the effects of the accountability system on student achievement.

of students along the *entire ability distribution*, including the lower end. This suggests that the increase in school-mean scores was *not* accompanied by greater inequity in education. Last, there appears to be *no* evidence of an adverse impact of the information disclosure on student outcomes beyond test scores (labeled below as non-cognitive skills).

The rest of the paper is organized as follows. Section 2 reviews related literature on the topic. Section 3 explains the background of the paper, detailing the educational policy reform in Japan in 2014. Section 4 explains the empirical methodology, and Section 5 describes the data. After Section 6 presents results, Section 7 offers concluding remarks.

2 Related literature

A number of previous works analyzed the effect of large-scale (not necessarily national) standardized tests as an accountability intervention tool. While many focus on accountability system that links schools' test performance with explicit consequences such as monetary awards or takeover threats, others highlight accountability based on the *mere information disclosure* of schools' performance. This paper adds to the latter strand of the literature.[6] For example, Burgess et al. (2013), acknowledging that education systems were identical in England and Wales until 2001, show that the abolition of school performance tables (thus shutdown of the information disclosure) in Wales in that year reduced school effectiveness relative to England, where school performance tables were kept being published. Koning and van der Wiel (2012) show that in Netherlands the publication of relative quality ratings of secondary schools in a national newspapers elicited a positive reaction from schools to improve their performance, particularly from ones that received the negative ranking. Further, focusing on randomly sampled villages in Pakistan, Andrabi et al. (2017) show that an exogenous increase in information provision of child- and school-level test scores increases subsequent test scores, particularly in private schools than in public schools. Camargo et al. (2018) also find that in Brazil the disclosure of school-level test results has a positive impact on student

[6]Examples of works highlighting the role of explicit consequence include Hanushek and Raymond (2005), Chiang (2009), Rockoff and Turner (2010), and Rouse et al. (2013). See Figlio and Loeb (2011) for a review.

performance only in private schools, suggesting that this is because only school managers of private schools are subject to market incentives. The basic message from the prior literature is thus that school accountability system that merely discloses information of school-level performance of standardized tests has potential to increase school effectiveness.

Meanwhile, it has been studied whether school accountability could cause *unintended effects*. Some papers reported that it could have a distributional impact since schools may have an incentive to allocate resources to students that are most critical to the accountability rating. For example, Reback (2008) show that, in the accountability system in Texas where explicit consequences are linked to school performance measured by students' pass rates, schools managed to improve the performance of students who are on the margin of passing. Also, Neal and Schanzenbach (2010) investigate the distributional effects of the accountability system in Chicago which links consequences to the number of students scoring above given proficiency thresholds. They demonstrate that such a system prompted teachers to pay little attention to students who had no realistic chance of becoming proficient or ones who were already proficient. Further, works such as Cullen and Reback (2006) and Figlio and Getzler (2006) find that accountability system could induce gaming/strategic behavior of schools in terms of manipulating the composition of students in the test-taking pool to maximize ratings and/or systematically placing certain selected students (including low-performing students) into education categories outside the accountability system. Reback et al. (2014) examine the effect of accountability on education outcomes beyond education achievements. Specifically, they show that while accountability pressure in the US states from the No Child Left Behind Act brought about short-term gains in test scores, this did not happen at the expense of students' enjoyment of learning or their anxiety over testing.[7]

Further, it is worth highlighting the strand of the literature which has examined the role of the disclosure of school performance information as a basis of *parental school choice*. For instance, Hastings and Weinstein (2008) show that in North Carolina more easily accessible and compara-

[7]Outside the literature on school accountability, Blazar and Kraft (2017), for example, examine the possible tension between improving test scores and outcomes beyond test scores. They show that teachers who are effective at improving students' math achievement tests often are not equally effective at improving their non-cognitive skills, called attitudes and behaviors in their paper.

ble performance information of different schools prompted parents to choose higher-performing schools. Likewise, Koning and van der Wiel (2013) find that in Netherlands the publication of school ranking in a national newspaper affected school choice, whereby more students enrolled in schools which had been ranked high in school quality. However, Mizala and Urquiola (2013) examine whether information on schools' value added rather than absolute outcomes affects parents' school choice and schools' market outcomes in Chile, and find that it does not.

The contribution of the present paper is to integrate the aforementioned strands of the literature on school accountability in the context of national standardized assessments, which have been carried out in many countries. Specifically, our results suggest that the mere disclosure of schools' performance in national assessments potentially increases school effectiveness, *without* producing unintended outcomes of leaving disadvantaged students behind, and having an adverse impact on non-cognitive skills. Further, by focusing on municipalities in Japan where parental school choice is *not* allowed, our results primarily reflect the response of schools to accountability pressure, *independent of* possible selection effects associated with school choice. On the policy front, we add to the active debate on the design of national assessments, by offering one evidence that accountability intervention via the disclosure of school-level results can be a desirable design feature.

3 Background

This section provides the background information for the analysis below. First, we explain general features of national assessments conducted in Japan since 2007. Second, we highlight a reform implemented in 2014 to the disclosure system of national assessment results.

3.1 National assessments in Japan

A national standardized assessment, known as National Assessment of Academic Ability (NAAA), has been conducted every April since 2007 in Japan, except for 2011, when it was canceled due to the impact of Great East Japan Earthquake. The aim of the assessment is to improve students'

learning environments by providing student diagnostic information to schools and parents of students. It does not have any bearing on students' progression through school or certification. Test takers are 6th- and 9th-grade students, last years of primary and junior high schools, respectively, and all students in those grades in public schools throughout the country have participated, except for 2010 and 2012 when randomly sampled students (about 30%) took part. While the subjects assessed changed somewhat year by year, Japanese language (Japanese hereafter) and Mathematics have been assessed every year.[8] To assess separately students' basic knowledge of the subjects and their ability to apply those basics to real-world problems, each subject is further divided into "Basics" and "Applications" components.[9] The NAAA has been accompanied by school and student questionnaires, which provide various information on the school environment and operational practices, and students' home environment and characteristics.

3.2 Information disclosure system of assessment results

To describe the disclosure system of the NAAA results, we first explain the *decentralized* nature of education system in Japan. While the Ministry of Education is the central education authority, local governments, both at the prefecture and municipality levels, are given a high degree of discretion, which creates a substantial variation in the local education system.[10] Across Japan, there are 47 prefectures and 1,718 municipalities (that are subordinate to prefectures), and education policies in each of prefectures and municipalities are designed and implemented by an *education board*, an executive body of usually 5 members.[11] For example, a prefectural education board appoints teachers and allocates them to public schools across municipalities under its jurisdiction (i.e., within the prefecture), whereas a municipal board establishes and abolishes public schools under its juris-

[8] In some years, natural science was also assessed.

[9] To illustrate in the context of mathematics, some of the "Basics" questions can be solved if students are simply familiar with basic algebra, whereas some of the "Applications" questions ask students to utilize basic algebra to solve types of logistic problems which can arise in a daily life.

[10] The Ministry of Education is officially known as the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

[11] The budgetary responsibility of education policies, together with other public policies, are borne by governors of prefectures and mayors of municipalities.

Table 1: A 2014 reform to the information disclosure system

	Level at which results are aggregated		
Disclosing entity	**Prefecture-level**	**Municipality-level**	**School-level**
Before the reform			
Prefectural board	Allowed	Not allowed	Not allowed
Municipal board		Allowed	Not allowed
School			Allowed
After the reform			
Prefectural Board	Allowed	Allowed	Allowed
Municipal board		Allowed	Allowed
School			Allowed

Notes: Differences caused by the reform are underlined. The Ministry of Education has published national-level and prefecture-level results every year, both before and after the reform.

diction (within the municipality), and manages key aspects of school operations, such as textbook selections from candidates certified by the Ministry of Education.

Turning to the information disclosure system of NAAA results, the Ministry of Education announced in November 2013 that a new system would come into force in April 2014.[12] Table 1 summarizes how the system has been changed, specifying the levels of aggregation of NAAA results disclosed by prefectural education boards, municipal boards, and individual schools, before and after the reform. Three levels of aggregation are considered: prefecture-level results (i.e., results aggregated at the level of the prefecture); municipality-level results; and school-level results. Detailed explanations of the table are given as follows:

- Before the reform, a prefectural board was allowed to report the prefecture-level result, but not the municipality- and school-level results, whereas the reform has allowed it to publish test results at the level of municipality and at the level of school, upon the agreement of the municipality board (which manages the schools).

[12]The basic notion behind implementation of this reform was that it is important for education boards and schools to fulfill accountability to parents and local residents. At the same time, the Ministry of Education emphasized that sufficient care has to be taken to avoid the reform having unintended effects. For example, to prevent excessive competition, a municipal board, when disclosing school-level results, was prohibited from ranking explicitly the schools based on the results.

- Before the reform, a municipal board was allowed to publish the municipality-level result, but not the school-level results, whereas the reform has allowed it to publish results of the schools under its jurisdiction. Specifically, the board can publish results of the schools on its own, or instruct them to publish their results on their own.
- Both before and after the reform, a school has been allowed to publish its own result.

The analysis below investigates the effects of publishing *school-level* results on student outcomes. To this aim, we utilize variations of disclosure system *across municipalities*, created by the discretion given to municipal boards after the reform. As clarified below in the context of our dataset, a prefectural board has *not* published results of schools managed by the municipality boards, whereas schools *hardly* publish their own results voluntarily. Taken together, the implication is that when school-level results are available to the public after the reform, this is predominantly because municipal boards have used the discretion to publish school-level results. Given this, we turn to a differences-in-differences approach below, comparing how student outcomes changed between municipalities that made use of the discretion with those that did not.

4 Identification and empirical methods

Our objective is to estimate the effects on student outcomes of accountability pressure put on schools when the disclosure of school-level performances in national standardized assessments enables the public to compare the performances across schools. Similar to Burgess et al. (2013), a relevant conceptual framework is a principal-agent model, where the public, parents in particular, as principals delegate the teaching of their children to schools/teachers as agents. The disclosure of school performances in standardized tests then helps the public and parents to monitor the output of the schools, that can be generalized as students' achievement of knowledge and skills. In this context, accountability pressure from the information disclosure may exert an effect on school behavior, possibly through the implicit sanctions of reputation effects.

Having clarified the objective, two classifications are in order. First, a treatment municipality is defined as the one for which performance of *all the schools* are published, so that the public/parents can *compare* school-by-school results properly, among the schools that share the same municipal (as well as prefectural) education policies and families of similar social backgrounds.[13] This classification is relevant, because, while the reform has allowed municipal boards to "instruct" the schools to publish their results on their own, it is *not* a legal mandate that leads to an explicit penalty in case of neglect. It is possible that some schools may not comply. Likewise, even when a municipal board publishes results of the schools on its own, it needs to "consult" the schools in advance. By focusing on municipalities where all the schools' performances are available, we ensure that accountability pressure through external comparison is considered to be strong.

Second, the analyses that follow only use municipalities where parental school choice is *prohibited*. That is, while each municipality is divided into several school districts with one school in each district, our focus is municipalities where parents can *only* send their child to the school located within the district of their residence. The rationale for this is to exclude possible selection effects through parental school choice. That is, to the extent that school-level test results are published and parental school choice is allowed, schools that have performed well may attract students with higher ability, such that estimates of the disclosure effect would capture a change in school behavior due to the information disclosure, *as well as* a change in the composition of students' ability due to the selection. By excluding school choice, we can focus on the former effect.

The preceding classifications are incorporated in the following differences-in-differences (DD) model, where we utilize variations in the information disclosure system created by the reform across municipalities:

$$\begin{aligned} Y_{ismt} = {} & \beta_1 + \beta_2 T_m + \beta_3 (T_m * After_t) + X_{ismt}\,\delta + Z_{smt}\,\gamma \\ & + EI_{mt}\,\theta + \upsilon_s + \zeta_t + \varepsilon_{ismt}. \end{aligned} \tag{1}$$

[13] About the relevance of the comparability of standardized test results in education achievements, Bergbauer et al. (2018), for example, indicate that "[t]he comparability of the generated achievement information suggests the possibility of using the tests to support incentives to students, but also to administrators and teachers by making external information available to parents, policy makers, and the general public" (p. 8).

Y_{ismt} represents an education outcome of 6th-grade student i in school s located in municipality m in year t.[14] To consider the fact that education policies are heterogeneous across prefectures, the analysis focuses on municipalities within one prefecture (elaborated below). As mentioned, only municipalities where parental school choice is prohibited are included. The main outcome variable is a test score itself (though we also consider measures of non-cognitive skills). Subjects are Japanese and Mathematics, both of which are further divided into two components of Basics and Applications. This means that we have four estimates of Eq.1. In the right hand side, T_m is a treatment dummy which takes the value of one if an education board in municipality m publishes results of all the schools under its jurisdiction after 2014, the year when the reform was implemented. $After_t$ is a year dummy variable, taking the value of one in years after 2014. X_{ismt} is a vector containing variables reflecting the student's home environment, covering whether she eats breakfast regularly, wakes up/sleeps at a regular time, and attends a tutoring school (private supplementary school).[15] Z_{smt} contains variables related to school environment, specifically, class size in the previous year (when student i was in the 5th-grade), average years of teaching experience of teachers, and the proportion of students who receive financial support.[16]

EI_{mt} contains municipality-level time varying variables, categorized into expenditure (E) and institutional (I) variables. The former include current and capital components of education expenditure per student, averaged over past 3 years (corresponding to the time when students were between the third to fifth grade). Current expenditure is further divided into personnel expenditure and the remaining. Institutional features of education systems include the frequency of school inspections (by municipal education board members) and the degrees of school autonomy in terms of budget allocation and curriculum development. υ_s is school fixed effects, capturing time-invariant, unobserved school-specific factors. ζ_t is time fixed effects, reflecting, for example, effects of prefectural education policies applicable to students in all the municipalities considered. Our main

[14] We focus on 6th-grade (rather than 9th-grade) students, because, at the level of junior high school, school choice is more widely granted to parents.

[15] We regard the student's attendance to a tutoring school as a proxy for the household's socio-economic status. With high socio-economic status, students may tend to be sent to a tutoring school for extra education.

[16] Municipalities provide financial support for students from low-income households. Hence, the proportion of students with school financial support is a proxy for poverty rate of the school district.

interest is the coefficient on the interaction between the treatment and after-reform dummies, β_3: if the disclosure of school-level information improves average test scores, $\beta_3 > 0$.

Our identification strategy builds on the DD estimator with Eq.1. The identifying assumption for the effect of information disclosure on test scores is that the outcome in treatment and control group would follow the same trend in the absence of treatment. This exogeneity assumption can be supported in various ways. First, the decision to disclose school-level results is made primarily by a municipal education board which is given the discretion, exogenous to students and schools. Second, acknowledging the possibility that the identification assumption is threatened when the reform on information disclosure is implemented as part of broader reform effort, Eq.1 controls for differences in municipality-level education expenditures and institutional features. As mentioned, we also consider parental school choice as another institutional feature. Third, school fixed effects also help, because they capture the effects of time-invariant municipality-specific unobserved heterogeneity. Fourth, although we calculate standard errors clustered at school level in the benchmark case, we also consider standard errors clustered at municipality level, accounting for the potential correlation of time-varying shocks across years and schools within a municipality. Last, as a test of the identifying assumption with a DD estimator, we test a parallel trend assumption by conducting an event study analysis.

We also investigate the distributional effect of the information disclosure on individual students, to consider the possibility that accountability pressure may affect equity in education. To this end, we use quantile regression to estimate a DD coefficient, testing whether the effect on test scores differs depending on the initial level of scores. Further, we estimate Eq.1 using measures of students' non-cognitive skills as an outcome variable. This is to examine effects of the information disclosure on student outcomes beyond test scores.

5 Data

This section first describes the information disclosure system of the NAAA in a prefecture called *Saitama*, which is highlighted in the analyses that follow. Second, we explain how the dataset is assembled, and provide descriptive statistics.

5.1 Information disclosure system in Saitama prefecture

As explained in Section 3.2, Japan's education policies are heterogeneous across prefectures. Thus, to isolate the effect of the reform on information disclosure implemented at the municipality level, the analyses below highlight municipalities within one specific prefecture, called *Saitama*. This is a prefecture with the population of 7.34 million (5th largest in 47 prefectures), located north of Tokyo prefecture, the capital of Japan. The main reason why we focus on Saitama, among the 47 prefectures, is that the availability of student-level NAAA data is particularly high.[17] Saitama prefecture consists of 63 municipalities, within which 40 are cities (a municipality with population of at least 50 thousands), 22 are towns (at least 5 thousand), and 1 is a village. We focus on *cities*, since, in towns and a village, the number of primary schools is relatively low (typically only a few), and the disclosure of school-level results may affect student outcomes differently.

First, we confirm that, in Saitama prefecture, (1) the prefectural education board itself publishes *none* of results of the schools managed by the municipalities, and (2) primary schools *rarely* disclose their own results voluntarily.[18] This means that, after the reform year of 2014, school-level

[17] To explain, the data is an administrative data, only available upon the official request to (and agreement by) the Ministry of Education on a project-by-project basis. Before our application for the use of the data was approved in January 2019, the Ministry had contacted all the municipal education boards in Japan to ask if they agreed to let data from their jurisdiction be used for research. Across Japan, about 70 percent of municipal boards did *not* agree as long as the municipality names were attached to the data. This means that for those municipalities, it was not possible to match the data with the treatment status of the disclosure of NAAA results. However, the majority of municipal education boards in Saitama prefecture agreed with us using the student-level data. Specifically, focusing on cities, but excluding Saitama city due to its special status in education policy making as a government designated city (which is not influenced by policies implemented by the prefectural education board), 32 out of the 39 city education boards (i.e., 82 percent) agreed, facilitating this research.

[18] We checked every single website of public primary schools in the control cities considered in the reference analysis below (cf. Table 4). Being in a control group, education boards of these cities neither published school-level results on their websites, nor instructed the schools under their jurisdictions to publish the results on their school websites. We confirmed (as of May 2020) that, among the total of 291 primary schools (in the 21 control cities), only 7 schools

Table 2: Information disclosure system in 4 potential treatment cities in Saitama

Features of disclosure system	**City A**	**City B**	**City C**	**City D**
Where are results disclosed?	Educ board website	School websites	School websites	School websites
Results disclosed (by subjects)	School-means	School-means	School-means	School-means
Results found for all schools?[a]	Yes (13/13)	Yes (7/7)	No (13/15)[b]	No (5/19)
Further breakdown of scores?[c]	Yes	Yes	No	No
Results available in uniform format?	Yes	Yes	No	No
Detailed reflection/improvement plan?	Yes	Yes	No	No
Since when are results disclosed?	2014	2014	After 2014[d]	2014

Notes: (a) Based on our own website check in April 2019. (b) This means that results were available only for 13 out of 15 schools. (c) The breakdown of mean-scores into sections of each subject. (d) The education board could not confirm when it started disclosing school-mean results.

NAAA results in Saitama prefecture are disclosed predominantly by municipal boards (including city boards) which have decided to use the discretion to disclose them. Next, to detail the information disclosure system in cities in Saitama, we acknowledge here that we have access to student-level data of the NAAA for 32 cities (see footnote 17 for the background information). Since no public data exists on the information disclosure system, we have requested an interview with the 32 city education boards to learn their disclosure system in detail, 31 of which have accepted our request. Then, interviews with the 31 city boards have revealed that after the reform in 2014, *4 boards* either have disclosed school-level test results on its board website, or instructed schools under their jurisdiction to disclose their results on their school websites. In either case, the published performance measure is *school-mean scores*. There has been *no* such practice to disclose the proportion of students who scored above certain proficiency thresholds.

As Table 2 shows, an education board of 1 city (called City A) disclosed school-by-school mean test scores on its website, whereas education boards of 3 cities (Cities B, C, and D) instructed schools under their jurisdiction to disclose school-mean scores on their school websites.

from 4 cities published some indication on their NAAA performance relative to the national average. Thus, it appears reasonable to assume that, for control cities, school-level results are rarely available to the public, unlike for treatment cities where results of all the schools within the cities are published.

However, as emphasized, since the "instruction" by a city board is not a legal mandate, we have checked if the mean scores were *actually* available for all the schools within each city, and found out that City A's education board website showed school-mean scores of all the 13 primary schools (under its jurisdiction) and school websites of *all* the 7 schools in City B provided school-mean scores, whereas for City C (D), only for 13 out of 15 (5 out of 19) school websites showed them.[19] Admittedly, since we have not monitored the school websites *continuously* after 2014, we cannot categorically say that the results were not disclosed properly in Cities C and D. Still, there is a concern for possible incomplete disclosure, hindering the public from comparing school performances within these cities, which supports our choice of using only Cities A and B as treated cities.

Further, *even the issue of possible incomplete disclosure aside*, Cities A and B appear to be more appropriate treatments, in that the information provided by the education boards facilitates the public comparison of school performances. For example, these cities provide not only the school-level mean scores for each subject, but also the breakdown (i.e., averages for different sections within each subject), while Cities C and D provide only the former at most. Next, the results of schools within Cities A and B are displayed using *exactly* the same format, making the comparison of performances straightforward. Further, schools in these cities complement the disclosure of the scores with detailed reflection and improvement plan.[20] Last but not least, in our interview with the education board in City C, they could not confirm the year when they started instructing the schools to disclose the results (albeit it is after the reform year of 2014), which raises a further concern for a measurement error, if the starting year is assumed to be 2014.

[19] Our interviews also revealed that the education boards of Cities B-D instructed the schools to disclose assessment results on their websites within the 6 months of the implementation of assessments. Thus, given that the 2018 assessment was implemented on 17 April, and we checked the school websites of those 3 cities in April 2019, if the schools had followed the instruction and had not withdrawn the results before the time of our survey, the 2018 results should have been still there.

[20] All the school performance information from those 4 cities, which the authors collected in April 2019 from the respective education board/school websites, is available from the authors upon request.

5.2 The dataset

The data source for test scores is the National Assessment of Academic Ability (NAAA). The associated questionnaires for students and schools provide information on student- and school-specific information. The NAAA data is first merged with the preceding information on the information disclosure system in the 31 cities in Saitama prefecture. We then merge city-level data on education expenditures and institutional features. (All the data sources are found in Table 11 in Appendix A.) The expenditure data for primary schools is decomposed into personnel, the rest of current, and capital expenditures, all of which are divided by the total number of primary school students (1st to 6th-grades) in the city, converted to expenditures per student. Regarding institutions, we consider parental school choice, school inspection, and school autonomy in budget allocations and curriculum developments. Parents' choice of school for their children is allowed only in 4 cities within the 31 cities.[21] School inspection is measured by the average number of times city education board member(s) visit each school in the city throughout the year. As for school autonomy in budget allocations, we make a dummy variable which takes the value of one if schools are granted any degree of autonomy by the city education board in the allocation of school budget, otherwise takes zero. A similar dummy variable is created for autonomy in curriculum development.

To reach the final dataset, we consider the adequacy of a control group, which is used to determine what would have happened to student outcomes in the absence of the disclosure of school-level results. Starting with the 27 possible control cities (i.e., the 31 cities excluding Cities A-D in Table 2), we first drop the 4 cities where parental school choice was allowed at the primary level, to abstract from the selection effects (see Section 4).[22] Second, since two education boards published school-by-school results on their websites *without revealing the school names* and this may still affect behavior of the schools, we exclude the two cities from the control group.[23] Last, since one

[21] Those 4 cities include not only the ones where parents are allowed to send their children to any school within the city, but also the ones where parents are allowed to choose a school in the districts adjacent to the school district of their residence. This is because, typically, the latter case still offers parents a few schools to choose from.

[22] In none of our possible treatment cities was school choice allowed at the primary level.

[23] To elaborate, schools in these cities, knowing their own results, can still find out their performance relative to other schools within the city albeit anonymous, which, in turn, may affect their behavior. We do not include these

city merged with other municipalities in 2010 and underwent major structural changes during the sample period, we exclude this city too. All in all, our sample consists of 23 cities, where 2 cities (Cities A and B) form an treatment group and 21 cities form a control group.[24]

Table 3 reports descriptive statistics of the sample of 23 cities, covering the total number of 220,345 6th-grade students who attended pubic primary schools between 2007 and 2018, inclusive.[25] We have repeated observations at the school level rather than the student level. All test scores are standardized within subject and year at the national level with the mean of 50 and the standard deviation of 10. The averages of all the scores are slightly lower than 50, meaning that the average test score in our sample is lower than the national average. Three of the student's home environment variables are ordinal variables, where the value of 3 (maximum) means that her answer to the question is "yes, certainly", and the value of 0 (minimum) means "no, not at all". Thus, the average of 2.86 for the question asking whether she eats break every morning indicates that a majority of students eat breakfast every morning. Averaged responses to the questions on whether she wakes up/go to bed about the same time everyday are relatively low, and the larger standard deviations indicate that there are larger variations in their responses. The use of a tutoring school is a dummy variable, which takes the value of one if she uses it, and zero otherwise. The mean of 0.46 indicates that 46 percent of students use a tutoring school service out of school. Turning to school environment variables, 10.9 percent of students received school financial support, though there is a substantial heterogeneity across schools with the minimum (maximum) of 0 (52.5) percent of students receiving the support. The average class size of the fifth grade in the previous school year is 32.9.[26] Here, the use of the previous-year figure is relevant, because the NAAA takes place in the first month (i.e., April) of the school year. The proportion of teachers with teaching experience less than 5 years is 31 percent on average.

cities in the treatment group either, because our interest is to estimate the accountability effects of the disclosure of school performances through a comparison of the performances by the general public.

[24] Since one city adopted school choice *and* disclosed school-level performances without revealing the school names, we exclude 6 cities from the potential 27 control cities, yielding 21 control cities.

[25] Within the 23 cities under consideration, there is only *one* private primary school (in one city).

[26] We drop schools with the average class size of less than 10 as possible outliers.

Table 3: Descriptive statistics: 23 cities in Saitama

Variable	Mean	Std. Dev.	Min.	Max.
Standardized test scores[a]				
Japanese (Basics)	49.72	9.97	-3.9	68.09
Japanese (Applications)	49.7	10.01	17.39	71.02
Math (Basics)	49.22	10.09	4.58	65.81
Math (Applications)	49.56	9.85	20.45	74.61
Student's home environment				
Eat breakfast every day[b]	2.86	0.47	0	3
Wake up at a regular time every day[b]	2.53	0.69	0	3
Sleep at a regular time every day[b]	2.21	0.82	0	3
Use a tutoring school[c]	0.46	0.5	0	1
School environment				
Prop of students receiving fin support[d]	10.85	7.32	0	52.5
Class size in previous school year	32.91	4.72	10	46
Prop of teachers with experience under 5 yrs[d]	31	13.44	0	100
City-level education expenditures and institutions				
Personnel exp per student (1,000 yen)[e]	34.33	17.02	2.89	70.29
Other current exp per student (1,000 yen)[e]	94.23	27.94	55.16	171.38
Capital exp per student (1,000 yen)[e]	89.73	69.90	8.5	439.9
School visit per school per year[f]	1.94	1.99	0	16.9
Autonomy in budget allocation[g]	0.17	0.34	0	1
Autonomy in curriculum development[h]	1.87	0.44	0	2
Number of observations		220,345		

Notes: Statistics over the 2007-18 period. (a) Test scores are standardized within subject and year at the national level with the mean of 50 and the standard deviation of 10. "Basics" ("Applications") tests the acquisition of basic knowledge of the subject (the ability to apply the basics to real-world problems). (b) 3 (max) means yes, certainly; 0 (min) means no, not at all. (c) 1 (0) means she attends (does not attend) a tutoring school. (d) Measured in percent. (e) Average expenditure over the previous 3 school years. (f) Number of school visits by board member(s) per school in the previous school year. (g) If schools were given autonomy in budget allocations in a given school year, the value of 1 is given, 0 otherwise. Numbers are based on the average over the previous three school years (to be in line with expenditure variables). (h) The sum of autonomy in curricula plus autonomy in auxiliary teaching materials. Each autonomy sub-component takes the value of 1 if school was granted autonomy in the previous school year, 0 otherwise.

We lag city-level variables considering the fact that the NAAA takes place in the first month of the school year. First, expenditure items per student are calculated as the average over the previous three school years, to capture possibly delayed effects on student outcomes. The mean of personnel expenditure (excluding salaries for the legally-stipulated number of full-time teachers) was 34,300 Japanese yen (JPY), which is equivalent to 312 US dollar (USD) using the exchange rate such that 1 USD equals 110 JPY, and the remaining current expenditure was averaged at 94,200 JPY

(856 USD).[27] The mean of capital expenditure, which covers such costs as building and computer equipment costs, was 89,700 (815 USD), while it shows a large variation. Regarding city-level institutions, the average number of school visits by city education board members during the previous year was 1.94 times per school. Autonomy in budget allocation is a dummy variable, which takes the value of 1 if any degree of autonomy in budget allocation is given to schools in the city in the previous year. Autonomy in curriculum development is the sum of two sub-components: autonomy in curricula and auxiliary teaching materials. Each sub-component takes the value of one when any degree of autonomy is granted in the sub-component in the previous year. The average of 1.87 indicates that compared to autonomy in budget allocation, autonomy in this aspect is much more commonly given to schools.

6 Results

First, we present results on the effects of disclosing school-mean scores on students' test scores. Then, we show results on the effects on equity among students, and also on outcomes beyond test scores, labeled as non-cognitive skills.

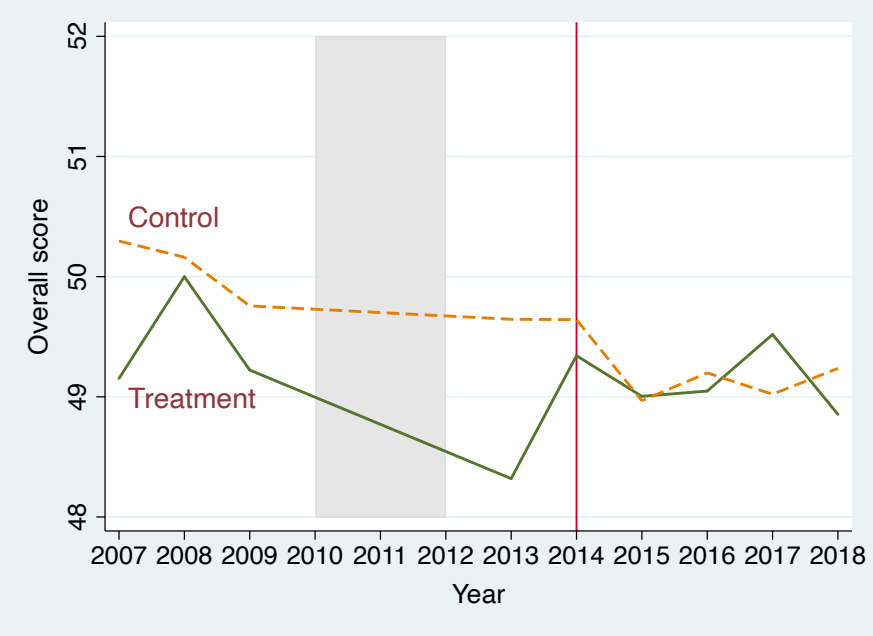
6.1 Effect of information disclosure on test scores

6.1.1 Reference results on test scores

Before estimating the full difference-in-differences (DD) model of Eq.1, we start with a simple chart to examine how our dependent variable of NAAA standardized test scores changed before and after the reform, in both treatment and control groups. An exposition is based on the sample of 23 cities, where two cities (Cities A and B of Table 2) are in treatment, and 21 cities are in control

[27] To explain, the reason why the personnel expenditure is small is that the majority of salaries for full-time teachers at the primary schools are not included in the personnel expenditure at the city level. This is because, in Japan, the baseline number of full-time teachers in the city is determined by the law (e.g., in proportion to the number of students in each school in the city), and the two-thirds of the total salaries paid to these teachers are covered by the prefectural government, and the one-third by the central government. Thus, the personnel expenditure at the city-level covers salaries only for additional teachers (i.e., ones employed beyond the number set by the law) and the other types of workers at schools (e.g., counselors).

Figure 1: NAAA overall scores in treatment and control groups over time



Notes: Overall score is the average of scores in all the four subjects: Japanese and Math, both of which are divided into Basics and Applications. Solid (dashed) line shows how the average of individual student's overall scores evolved for treatment (control) groups of 2 (21) cities for years between 2007 and 2018. For 2010, 2011, and 2012. corresponding to the shaded area, interpolated values are used.

groups. Figure 1 plots, for control and treatment groups and over the period over 2007 and 2018, the average of individual students' overall scores, which, in turn, are the average scores across all the four subjects (Japanese and Math, both are divided into Basics and Applications). Since the sample is smaller for 2010 and 2012 (sampling years), and does not exist for 2011 altogether (canceled year), interpolated values are used for those three years using the 2009 and 2013 values. As mentioned, the reform on the information disclosure system was announced in November 2013, and the new system has been adopted since the assessment in April 2014. Consistent with the notion that the reform may have increased test scores, the figure shows that until 2013 the overall score had been systematically lower in the treatment cities, whereas the gap between the two groups narrowed substantially in 2014, and the scores have been similar ever since.

Moving to estimation of the model, Table 4 summarizes the estimates of the coefficient on the interaction term between the treatment dummy and the after dummy, i.e., DD estimates for each

Table 4: Effects of information disclosure on test scores

	(1)	(2)	(3)	(4)
Japanese (Basics)				
Treat*After	0.921***	0.899***	0.932***	1.122***
	(2.671)	(2.800)	(2.869)	(3.441)
Japanese (Applications)				
Treat*After	0.996***	1.009***	1.038***	1.165***
	(3.492)	(3.522)	(3.548)	(3.853)
Math (Basics)				
Treat*After	0.485*	0.510*	0.564**	0.813***
	(1.724)	(1.912)	(2.130)	(2.823)
Math (Applications)				
Treat*After	0.585**	0.561**	0.583**	0.886***
	(2.118)	(2.175)	(2.249)	(3.211)
Year fixed effects	No	Yes	Yes	Yes
School fixed effects	No	Yes	Yes	Yes
Student environment	No	No	Yes	Yes
School environment	No	No	Yes	Yes
City exp & institution	No	No	No	Yes
Observations		220,345		

Notes: Estimated model applicable to Column (4) is Eq.1: $Y_{ismt} = \beta_1 + \beta_2 T_m + \beta_3 (T_m * After_t) + X_{ismt}\,\delta + Z_{smt}\,\gamma + EI_{mt}\,\theta + \upsilon_s + \zeta_t + \varepsilon_{ismt}$. Dependent variable, Y_{ismt} is test score standardized within subject and year at the national level with the mean of 50 and and the standard deviation of 10. DD estimates of β_3 from regressions of different specifications (Columns (1) to (4)) are shown for different subjects. T_m is a treatment dummy taking the value of one if in city m school-level results are published after 2014 (the reform year), and $After_t$ is a year dummy variable, taking the value of one in years after 2014. X_{ismt} (Z_{smt}, EI_{mt}) contains student's home environment (school environment, city-level education expenditure and institution) variables. υ_s (ζ_t) is school (time) fixed effects. Sample of 23 cities is used: control and treatment groups consist of 2 (Cities A and B in Table 2) and 21 cities. Robust t-statistics in parentheses. Clustered standard errors are used to adjust for correlation of error terms within school. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

subject. We sequentially add controls to the full model in Column 4. (Estimates of the other coefficients of the full model are presented in Table 12 in Appendix B.) Column 1 shows the estimate of the effect of information disclosure by regressing each test score without control. Coefficients are all positive and statistically significant, though only at 10 percent level when Math (Basics) is a dependent variable. The results are generally consistent with the simple unconditional plot in Figure 1. The coefficient of 0.92 for Japanese (Basics) implies that the test score of 6th-grade students in public primary schools of treated cities increased by 0.92 points (i.e., 0.092 standard deviation)

compared to those in control cities, after education boards in treated cities disclosed school-mean scores in 2014. The effects are larger for Japanese than for Mathematics.

Estimates of the DD coefficients reported in the remaining columns are ones with different sets of controls. We sequentially add controls of year and school fixed effects (Column 2), student home environment and school environment variables (Column 3), and city-level expenditure and institution controls (Column 4). Column 4 corresponds to the full model. The estimates are statistically significant regardless of the specifications, indicating that the positive effects of information disclosure are robust to the inclusion of a battery of controls. In Column 4 where the addition of city-level education expenditures and institutions helps address an omitted variable problem which arises in case the information disclosure is implemented as a part of broader institutional reform, the magnitude of the coefficients is largest, and they are all significant at 1 percent level.[28]

6.1.2 Parallel trend and lagged effects

To test the validity of a parallel trend assumption, we conduct an event study analysis. Specifically, we check whether changes in test scores are correlated with treatment status before the reform. Based on Eq.1, we estimate the following regression equation:

$$\begin{aligned} Y_{ismt} = \beta_1 + \beta_2 T_m + \sum_{\tau=2007, \tau \neq 2013}^{2018} \beta_3^{\tau} Year_t^{\tau} + \sum_{\tau=2007, \tau \neq 2013}^{2018} \beta_4^{\tau} \left(T_m * Year_t^{\tau}\right) \\ + X_{ismt}\,\delta + Z_{smt}\,\gamma + EI_{mt}\,\theta + \upsilon_s + \varepsilon_{ismt}, \end{aligned} \tag{2}$$

where $Year_t^{\tau}$ is a dummy variable taking the value of one if $t = \tau$, otherwise zero. We take the year 2013, a year before the reform, as a reference year (hence setting $\beta_3^{2013} = 0$ and $\beta_4^{2013} = 0$). The remaining resembles Eq.1. If the parallel trend assumption is satisfied, we expect that $\beta_4^{\tau} = 0$ for $\tau = 2007, ..., 2012$.

Table 5 summarizes coefficients on the interaction term between year dummies and the treatment dummy (i.e., β_4^{τ}) for all the subjects. The specification considered contains all the controls

[28] As shown in Table 12 in Appendix B, coefficients of the city-level controls are jointly significant except for Math, Applications (even for which the p-value of joint significance test is 0.14).

Table 5: Pre-reform trend and lagged effects

	Japanese		**Mathematics**	
	Basics	Applications	Basics	Applications
Treat*Year (β_4^τ)	(1)	(2)	(3)	(4)
$\tau = 2007$	0.141	-0.610	0.168	0.300
	(0.250)	(-0.901)	(0.229)	(0.484)
$\tau = 2008$	0.804*	0.599	0.968	1.102**
	(1.658)	(1.127)	(1.545)	(2.062)
$\tau = 2009$	0.781	0.215	0.385	0.882
	(1.466)	(0.324)	(0.653)	(1.623)
$\tau = 2010$	0.489	-0.795	2.006	0.719
	(0.284)	(-0.489)	(1.220)	(1.044)
$\tau = 2012$	0.102	0.921	-0.025	1.239
	(0.242)	(0.831)	(-0.030)	(1.188)
$\tau = 2014$	1.976***	0.782*	0.237	1.264**
	(4.922)	(1.940)	(0.396)	(2.356)
$\tau = 2015$	1.430***	1.465***	1.547***	1.406**
	(2.894)	(2.960)	(2.698)	(2.304)
$\tau = 2016$	1.033**	1.320**	1.452**	1.694***
	(1.990)	(2.550)	(2.336)	(3.596)
$\tau = 2017$	2.098***	1.840**	1.852***	1.719***
	(3.052)	(2.542)	(3.046)	(2.767)
$\tau = 2018$	1.100	0.955	1.034	1.333**
	(1.494)	(1.164)	(1.524)	(2.033)
Year fixed effects	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
Student environment	Yes	Yes	Yes	Yes
School environment	Yes	Yes	Yes	Yes
City exp & institution	Yes	Yes	Yes	Yes
Observations	220,345			

Notes: Estimates of β_4^τ (see Eq.2). Dependent variable is standardized test score. Robust t-statistics in parentheses. Clustered standard errors are used to adjust for correlation of error terms within school. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

including city-level expenditures and institutions, corresponding to Column 4 of Table 4. The table is divided into two panels by a horizontal line right above the interaction for 2014. The estimates in the upper panel reveal that the coefficients on the interaction terms are mostly insignificant for years from 2007 to 2012, except 2008 for Japanese (Basics) and Math (Applications). These results strongly validate our identifying assumption of the parallel trend before the reform.

Results from the event study are also useful to investigate how long it took for the effects of information disclosure to emerge. The lower panel of Table 5 reports the estimates of the interaction terms for years from 2014 to 2018. For Japanese (Basics and Applications), and Math (Applications), the coefficients on the interaction term between year 2014 and the treatment dummy are positive and statistically significant. Since the Ministry of Education announced in November 2013 that the reform on the information disclosure would be implemented from the assessments in April 2014, these results suggest that the effects of information disclosure emerged in a short time period for these subjects. Further, as suggested by the fact that all the coefficients on the interaction between year 2017 and the treatment dummy are still positive and significant, the effects of disclosing school-mean scores persisted regardless of the subjects.

6.1.3 Robustness checks

The preceding analyses indicate that the public disclosure of school-mean scores of national assessments increases test scores. Here, we conduct four robustness checks.

Using standard errors clustered at city level First, there may be correlation of unknown form between schools within a city. To address this, we calculate standard errors clustered at city level, the level of intervention. This is an alternative approach to the above where standard errors are calculated at school level, the level at which decisions are taken reacting to the intervention.

Controlling for the disclosure of city-level results Second, remember from Table 1 that the reform has made it possible for a prefectural education board to publish *municipality-level* results upon the agreement of the municipality board. Among the 23 cities covered in the preceding analyses, 16 city boards agreed the city-level results to be published at the prefecutural board website since 2014 onwards, whereas, by 2018, all the remaining 7 boards had agreed. Here, to consider its possible relevance, we control for a dummy variable which takes the value of one for city (municipality) m in the years when results of the city were published and zero in the other years (including all the pre-reform years).

Table 6: Robustness checks

	Japanese		**Mathematics**	
	Basics	Applications	Basics	Applications
	(1)	(2)	(3)	(4)
Panel 1: Use standard errors clustered at city level				
Treat*After	1.122**	1.165***	0.813***	0.886**
	(2.464)	(4.781)	(3.162)	(2.769)
Panel 2: Control for city-level information disclosure				
Treat*After	1.156***	1.126***	0.815***	0.860***
	(3.512)	(3.690)	(2.798)	(3.107)
Panel 3: Add previous city-level average				
Treat*After	0.956***	0.957***	0.627**	0.754***
	(2.975)	(3.171)	(2.360)	(2.823)
Prev city ave	0.052	0.047	0.207***	0.109**
	(1.230)	(1.085)	(3.939)	(2.515)
Panel 4: Use adjacent control cities				
Treat*After	1.406***	0.739*	0.602	0.961**
	(3.460)	(1.669)	(1.388)	(2.149)
Year fixed effects	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
Student environment	Yes	Yes	Yes	Yes
School environment	Yes	Yes	Yes	Yes
City exp & institution	Yes	Yes	Yes	Yes

Notes: DD estimates of β_3 (cf. Eq.1). Dependent variable is standardized test score. Robust t-statistics in parentheses. Panels 1 and 2 (Panel 3, Panel 4) are based on observations of 220,345 (190,169, 50,702). In Panel 1 (Panels 2, 3, and 4), clustered standard errors are used to adjust for correlation of error terms within city (school). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Controlling for previous city-average score Third, the decision of a city education board to disclose school-level results may be correlated with past results of national assessments in the city. For example, if a city board observes that schools under its jurisdiction performed relatively badly in the previous assessments, it may have a stronger incentive to make the schools more accountable for their performance by disclosing the results. Meanwhile, there may be some persistence in education policies implemented at the city level. We test robustness by additionally controlling for the average test scores at the city level in the previous year.[29]

[29]We do not control for this variable in the main specification, because controlling for a lagged value inevitably reduces the sample size. In particular, we not only lose observations in 2007 (the first year of the sample) but also ones in 2012 since there is no data for 2011 (due to the assessment cancellation caused by the Great East Japan Earthquake).

Addressing possible geography-related confounding factors Fourth, the preceding analyses use treatment and control cities regardless of geographical locations within Saitama prefecture. However, comparison of municipalities located distantly may suffer from confounding factors which are correlated to changes both in test scores and municipality's characteristics (which may bring about information disclosure). To mitigate this possible violation of the identifying assumption associated with locations, we repeat analysis by focusing on the two treatment cities and four control cities which are *adjacent to* those treatment cities.

Table 6 reports DD coefficients for the robustness tests, all of which are based on the full model (cf. Column 4 of Table 4). The first panel shows results using standard errors clustered at the city level. DD coefficients are still all significant and positive for all the subjects. The second panel considers the disclosure of assessment results at the city level. DD coefficients are again always significant and positive. Though not shown, coefficients on the dummy for the disclosure of city-level results are always insignificant. The third panel controls for city-level average scores in the previous year, showing that DD coefficients are significant and positive. Coefficients on previous year's city-level averages are insignificant for Japanese, but significant and positive for Math. The fourth panel is based on students in the treatment and the adjacent control cities alone. Despite the smaller sample, DD coefficients are positive and significant except for Math (Basics).

6.1.4 Effects on school behavior

Having seen evidence that the reform increased assessment scores in treated cities after 2014, we consider *how* it changed schools' behavior. In particular, we explore whether schools in treated cities actually became more accountable for their performances, and how they managed to raise test scores. To this aim, we investigate schools' responses to the following questions (which were asked as a part of the NAAA) on *school accountability* to parents and local residents, and on schools' use of NAAA results as a *diagnostic tool*:

> (On accountability) *To what extent did you explain your school's results on the NAAA from the previous year to parents and local residents?*

Table 7: Reform and school behavior

	Accountability	**Diagnostic tool**	**Take-up rate**
	(1)	(2)	(3)
Treat*After	0.488***	0.212**	0.004
	(3.321)	(2.503)	(0.913)
Year fixed effects	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes
Student environment	Yes	Yes	Yes
School environment	Yes	Yes	Yes
City exp & institution	Yes	Yes	Yes
Observations	1,871	2,171	1,805

Notes: Estimates of β_3 (see Eq.3). Dependent variable in Column (1) (Column (2)) is the measure of school accountability (the school's use of previous NAAA as a diagnostic tool), created from the school's response to the corresponding question in the questionnaire accompanying the NAAA. Dependent variable in Column (3) is a take-up rate (share of students who took assessments to the total number of students in the school). Robust t-statistics in parentheses. Clustered standard errors are used to adjust for correlation of error terms within school. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

(On diagnostic tool) *To what extent did you make use of the analysis of your school's results in the previous year's NAAA to improve the teaching quality of your school?*

Formally, we estimate the following regression model:

$$\begin{aligned} Y_{smt} = {} & \beta_1 + \beta_2 T_m + \beta_3 (T_m * After_t) + \overline{X_{ismt}}\,\delta + Z_{smt}\,\gamma \\ & + EI_{mt}\,\theta + \upsilon_s + \zeta_t + \varepsilon_{smt}, \end{aligned} \tag{3}$$

where Y_{smt} is the school's responses to questions on the preceding questions. Schools' responses are ordinal variables: 2 means "very well", 1 "reasonably well", and 0 "hardly".[30] Since the unit of an observation is school, $\overline{X_{ismt}}$ contains the school averages of student environment variables. Other variables are as defined in the analyses above, i.e., Eq.1. Columns (1) and (2) of Table 7 present the estimates of the effect of information disclosure on school accountability and schools' use of assessment results as a diagnostic tool. All the sets of controls are used. In both columns, the

[30] As a caveat, the question on accountability (diagnostic tool) was not asked in 2007, 08, 12, 13 (2007). Besides, the question on accountability (diagnostic tool) was simpler in 2009 and 10 (2008, 09, 10), asking schools to answer either "yes, they explained (made use of the previous results)" or "no". To make these answers in the early years comparable to the later years when the three possible responses were available, we converted "yes" to 1.5 and "no" to zero for those early years.

estimates are positive and statistically significant, indicating that the public disclosure of school-level results of national assessments enhanced accountability of schools, and also prompted schools to use assessment results better to improve their teaching quality.

We also examine the conjecture that the relative increase of test scores of students in schools of treated cities may be driven by manipulative behavior by schools. One specific concern might be that schools in treated cities influence who sit in assessments, since they may have a stronger incentive to select students who they expect will perform well. To address this concern, we use the model of Eq.3, and estimate the effect of the reform on the school-level *take-up rate* of assessments, calculated as the share of students who took assessments to the total number of 6th-grade students in the school. Column (3) of Table 7 shows that the reform has an insignificant effect on the take-up rate, implying that the schools did *not* resort to cheating through manipulation of the testing pool.[31] (Descriptive statistics of the variables on school behavior are found in Table 13 in Appendix B.)

6.1.5 Discussions

On the sample with four treated cities In our survey on the information disclosure system of the NAAA across cities in Saitama, four education boards of Cities A to D reported that they published school-mean scores at the board website (City A) or instructed the schools to publish them at school websites (Cities B to D). However, the detailed investigation on the system prompted us to focus on Cities A and B as adequate treated cities (see Table 2). Instead, Table 14 in Appendix B shows the estimates of the DD coefficients assuming that all the 4 cities constitute a treatment group, and also assuming each city as a separate treatment group.[32] Throughout, the same 21 cities form the control group (as in Table 4), and the specification follows the one with all the controls. When

[31] The main reason why the number of observations is limited (to 1,805) here is that we cannot match all the schools in the 23 cities in the reference sample with school-level data taken from the *outside* of the school questionnaire of the NAAA. Specifically, for the students in the 5 control cities, we cannot identify the names of the schools they attend (though we know the names of the cities where the schools are located). In the current context, since the total number of 6th-year students in each school, used as a denominator to calculate the take up rate, is taken from the Saitama Prefecture School Handbook, and the number of the assessment takers is counted for each school using the NAAA data without knowing the school names in case of the 5 cities, we cannot obtain a take-up rate for the schools in those cities. This means that all the schools in the 5 cities are missing entirely from the analysis on the take-up rate.

[32] Since, as mentioned, there is ambiguity about the year in which City C started the information disclosure, the exercise assumes that the starting year is 2014, which is subject to a possible error.

all the 4 cities form a treatment group, the DD coefficients retain statistical significance only for Japanese (Basics and Applications). Next, when considering each of the 4 cities separately, for Cities A and B, the DD coefficients are all positive and significant for 3 out of the 4 subjects, whereas, for cities C and D, some of the coefficients are even negative and none are significant. These results are consistent with the apparent inadequacy of Cities C and D as treated cities.

On the possibility of selective migration As emphasized, the preceding result that the disclosure of school-mean scores of the NAAA improved students' test scores is free from the selection effect based on parental school choice. However, the migration of students across cities could still be a potential threat to the identification of the accountability effect. For example, if students with high ability migrate selectively from schools in cities without disclosure of school-level results to schools in cities with disclosure, our estimates could be biased due to selective migration. To examine whether this type of migration can be a serious concern, we test whether the number of 6th-grade students in cities with the disclosure policy increased relative to other cities after the reform in 2014. If parents in other cities without disclosure policy expect that their children benefit from learning in schools with high test scores, they may move into such schools revealed by the disclosure.[33] Thus, the number of students in cities where education boards disclose school-level results may increase if selective migration across cities is a plausible concern.

To test this hypothesis, we use the number of children at age 11 in a city as a proxy for the total number of 6th graders, and regress it on the interaction terms between the treatment dummy and the dummy variable indicating year after 2014, controlling for city fixed effects and year fixed effects. The analysis is based on the reference sample with 2 treated and 21 controlled cities, using standard errors clustered at city level. Column 1 of Table 15 in Appendix B shows the estimate of the coefficient of the interaction term. The sign is negative but statistically insignificant, indicating that there is no relative increase of the number of 6th grade students in treated cities. Further, considering the possibility that the selective migration may occur even before the year of 6th grade,

[33] Similarly, students in a city with the disclosure policy may move within the city, towards a district with a school with high scores. However, this type of within-city migration does not affect the average test scores of the city.

Columns 2 and 3 report estimation results with the number of children whose age is from 7 to 11 (as a proxy of all children in primary school below the grade 6) and those aged from 0 and 11 (all children below the grade 6) as a dependent variable. Both coefficients are again insignificant, suggesting that it is unlikely that selective migration of students across cities biases our estimates of the information disclosure effect.

6.2 Investigation of possible trade-offs of information disclosure

6.2.1 Compromising equity in education?

The preceding results suggest that the information disclosure of school-mean scores of the NAAA increased school effectiveness. To shed light on the desirability of the information disclosure as a design feature of the assessment, it is important to examine possible trade-offs. First, we examine the possibility that the information disclosure may have an adverse distributional impact across individual students. The concern here is that the school held accountable for the mean scores may attempt to improve them by allocating resources unequally to students, such that equity in education may suffer with some students being left behind.

To proceed, we test whether the effect of the information disclosure on test scores differs by students' ability levels. To clarify, as Woessmann (2005) explains, the student's ability is something unmeasured, but once individual characteristics, school resources and institutional effects are controlled for, the conditional test performance should be closely associated with ability. Specifically, what we do here is to run quantile regressions to estimate the disclosure effect on test scores at different points on the ability distribution.

Table 8 summarizes the DD coefficients estimated by quantile regressions using Eq.1 for different subjects. All the control variables are added throughout. Coefficients are reported for 9 quantiles ranging from 0.1 to 0.9, complemented by the OLS estimates obtained above (cf. Column 4, Table 4). Starting with Japanese (Basics) in Column 1, the coefficients are all positive and significant at all the quantiles. This implies that students of all the ability levels benefited from the disclosure of school-mean scores. Still, notice that while the effect at the top end of the distribution

Table 8: Heterogeneous effects by student ability: quantile regressions

	Japanese		**Mathematics**	
	Basics	Applications	Basics	Applications
Treat*After	(1)	(2)	(3)	(4)
Quantile (Q) regressions				
$Q = 0.1$	**1.991*** **	0.933***	**1.055*** **	0.751***
	(6.266)	(3.200)	(3.035)	(3.130)
$Q = 0.2$	1.661***	1.156***	1.035***	0.945***
	(6.474)	(5.269)	(3.020)	(4.019)
$Q = 0.3$	1.200***	1.143***	0.925***	1.041***
	(4.516)	(4.378)	(3.247)	(4.102)
$Q = 0.4$	1.338***	1.153***	0.748***	**1.223*** **
	(4.687)	(4.878)	(2.728)	(4.634)
$Q = 0.5$	1.038***	**1.536*** **	0.824***	1.157***
	(4.657)	(6.213)	(3.502)	(4.485)
$Q = 0.6$	1.178***	1.385***	0.875***	1.003***
	(5.270)	(5.334)	(3.865)	(3.984)
$Q = 0.7$	1.038***	1.013***	0.690***	0.779***
	(5.134)	(4.233)	(3.731)	(3.232)
$Q = 0.8$	0.745***	0.190***	0.457***	0.659***
	(4.944)	(3.633)	(3.236)	(2.825)
$Q = 0.9$	0.344**	0.951***	No	0.235*
	(2.375)	(9.081)	estimate	(1.719)
OLS regressions (Column 4 of Table 4)				
	1.122***	1.165***	0.813***	0.886***
	(3.441)	(3.853)	(2.823)	(3.211)
Year fixed effects	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
Student environment	Yes	Yes	Yes	Yes
School environment	Yes	Yes	Yes	Yes
City exp & institution	Yes	Yes	Yes	Yes
Observations	220,345			

Notes: Estimates of β_3 (see Eq.1). Dependent variable is standardized test score. Robust t-statistics in parentheses. Clustered standard errors are used to adjust for correlation of error terms within school. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

($Q = 0.9$) is relatively small as expected (with little room for improvement), the effect at the lower end tends to be larger than the one in the middle of the distribution. For instance, while the OLS coefficient is 1.12, the largest quantile coefficient at $Q = 0.1$ (expressed in bold) is 1.99, indicating

that disclosing school-means increases the test score by about 0.2 standard deviation at this lowest point of the ability distribution.

Turning to Japanese (Applications) in Column 2, the result still indicates that the scores of students of different ability levels improve significantly. However, unlike Japanese (Basics), an increase tends to be larger in the middle of the distribution, with the largest coefficient observed at $Q = 0.5$. This difference in results between Basics and Applications appears to be consistent with a difference in the contents of the tests. That is, the former primarily tests the acquisition of basic knowledge of the subject, whereas the latter tests the student's ability to apply the basic knowledge to real-world problems. Thus, it may be the case that while it is relatively easy to improve students' knowledge accumulation at the lower end of the distribution, it may be less straightforward to improve the ability to apply basic knowledge of those students. These patterns are somewhat repeated for Mathematics, where test scores generally improve along the entire ability distribution, but the highest improvement in Math (Basics) is observed at $Q = 0.1$, while for Math (Applications), it is observed at $Q = 0.4$.[34]

Here, the main message is that the benefit of a rise in school effectiveness caused by the disclosure of school-level test results does *not* have to be offset by an adverse distributional effect. This may seem to contradict to the results of some previous works in the literature that indicate a possible adverse distributional effect of school accountability through resource reallocation towards specific students that are critical to improve students' pass rates or students' proportion above some proficiency thresholds (e.g., Reback (2008) and Neal and Schanzenbach (2010)). Since those works highlight accountability mechanism created by linking school performance to explicit consequences (rather than by merely disclosing school performances), a simple comparison may be difficult to begin with. However, one plausible explanation appears to be that when school-mean scores are a published performance measure (as in here), schools may have an incentive to put an effort across the entire ability distribution.

[34]No estimate of DD coefficient is available at $Q = 0.9$ for Math (Basics), because there is no room for improvement in scores at this quantile.

Table 9: Within-school variance of test scores

	Japanese		**Mathematics**	
	Basics	Applications	Basics	Applications
	(1)	(2)	(3)	(4)
Treatment*After	-0.625***	-0.171	-0.264	-0.182
	(-2.758)	(-0.687)	(-0.892)	(-0.897)
Year fixed effects	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
Student environment	Yes	Yes	Yes	Yes
School environment	Yes	Yes	Yes	Yes
City exp & institution	Yes	Yes	Yes	Yes
Observations	2,816			

Notes: Estimates of β_3 (cf. Eq.3). Dependent variable is within-school variance of standardized test scores. Robust t-statistics in parentheses. Clustered standard errors are used to adjust for correlation of error terms within school. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

To check robustness of the result that publishing school means does not have an adverse distributional effect, we estimate its effect on the *within-school* variance of test scores. The estimated model, where the unit is school, is similar to Eq.3 (cf. Table 7). Table 9 summarizes estimated DD coefficients for each subject. While all the estimates are negative, the estimate is statistically significant for Japanese (Basics). This means that in the case of Japanese (Basics), the information disclosure reduces the within-school variance of test scores, which, in turn, is in line with the relatively large increase in individual scores at the lower end of the ability distribution observed in Table 8. The estimates for the other subjects are insignificant, suggesting that disclosing school-means does not increase the disparity of test scores among students within a school. These results thus support the lack of adverse distributional effect of the informational disclosure.

6.2.2 Adverse impact on outcomes beyond test scores?

Next, we consider the possibility that the positive impact of the information disclosure of school-level results of the NAAA on test scores may be offset by an adverse impact on outcomes beyond them, referred to as *non-cognitive skills*. Although this trade-off is admittedly rather speculative, possible tension between improving test scores and non-cognitive skills itself is not new.

For instance, Blazar and Kraft (2017) show that just because teachers are effective at improving students' test scores does not mean that they are also effective at improving their non-cognitive skills (they call "attitudes and behaviors"). In our context, schools held accountable for the NAAA performance may put excessive focus on subject teaching, possibly at the expense of nurturing non-cognitive skills of students. To the extent that this is the case, such effects can be seen as a disadvantage of the information disclosure, because it is well known that some non-cognitive skills have a positive effect on later grades at school and/or labor market outcomes in adulthood.[35]

Here, using some of students' responses in student questionnaire accompanying the NAAA, we formulate some measures of non-cognitive skills, and examine how the information disclosure may affect them. At the outset, however, we clarify that the analysis below is not meant to be a comprehensive analysis of the effects on non-cognitive skills. Rather, with the restricted range of relevant questions on students' personal traits available in the questionnaire, the aim is to examine if there is any suggestive evidence of adverse effects. In what follows, we highlight students' responses to questions on *self-confidence*; *grit (perseverance)*; *social interactions*; and *self-control*. On self-confidence and grit, the questionnaire contains the following questions:

(On self-confidence) *Do you think you have strong (good) points?*

(On grit) *Do you have future dreams and goals?*

On social interactions, we consider the questions on compassion for others and respect for rules:

(On compassion) *Do you think that bullying is unacceptable under any circumstances?*

(On rules) *Do you obey school rules?*

Regarding self-control, we use the response to the following question:

(On self-control) *Do you study at home by making study plan by yourself?*

Students' responses are ordinal variables: 3 means "yes, absolutely", 2 "yes, tentatively", 1 "no, tentatively", and 0 "no, absolutely". Descriptive statistics of these variables are reported in Table

[35] For recent literature review on the effects of non-cognitive skills on those outcomes, see, for instance, Zhou (2016).

Table 10: Reform and students' non-cognitive skills

	Confidence	**Grit**	**Compassion**	**Rules**	**Control**
	(1)	(2)	(3)	(4)	(5)
Treatment*After	0.001	-0.023	-0.001	0.026	-0.018
	(0.047)	(-0.948)	(-0.056)	(1.018)	(-0.546)
Year fixed effects	Yes	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes	Yes
Student environment	Yes	Yes	Yes	Yes	Yes
School environment	Yes	Yes	Yes	Yes	Yes
City exp & institution	Yes	Yes	Yes	Yes	Yes
Observations	220,224	220,216	220,170	220,209	220,233

Notes: Estimates of β_3 (cf. Eq.1). Dependent variable is a measure of non-cognitive skills. Robust t-statistics in parentheses. Clustered standard errors are used to adjust for correlation of error terms within school. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

16 in Appendix B. We estimate the effects of the reform on these measures, using the model of Eq.1 with all the controls added.

Table 10 presents DD estimates for each of the outcomes. Results show that the reform does not have any significant effect on those measures of non-cognitive traits. To report, we further checked effects on other possible non-cognitive measures available from the questionnaire, such as students' willingness of helping others, practice of keeping promises with friends, and showing interests in events in society, and found that there are no significant effects on them too.[36] Again, it is admittedly difficult to make a strong claim based on this rather speculative exercise. However, the bottom line is that there is no evidence that the information disclosure has an adverse impact on education outcomes beyond test scores.

7 Concluding remarks

This paper studied empirically how publishing school-level results of a national assessment affects student outcomes. Considering the case of Japan, we showed that publishing school-mean scores of these assessments increased test scores. Schools became more accountable to the public, and

[36] Results are available from the authors upon request.

made better use of diagnostic information from previous assessments to improve teaching quality. We further showed that the apparent increase in the school effectiveness was not offset by the worsening of equity in education: students of all ability levels, including ones at the lower end of the distribution, benefited from an improvement in test scores. Besides, there was no indication that non-cognitive skills were adversely affected.

These results potentially have enormous implications. World Bank (2018) highlights the low learning levels of children attending school in many developing countries. In the report, one of the key policy actions recommended to address this learning crisis is to assess learning, using well-designed student assessments as a diagnostic tool. Clearly, one of the key design issues of any large-scale standardized assessment is the choice over the level of aggregation of results disclosed to the public. While more empirical evidence from different institutional/cultural settings is required, we offered one comprehensive evidence to suggest that publication of school-level results, school-mean test results in particular, can have a positive impact on student achievements without unintended adverse effects.

Appendix

A Data sources

Table 11: Data sources

Data (alphabetical order)	Sources
Assessment scores, for all subjects	National Assessment of Academic Ability (NAAA)
Information disclosure system	Our own survey
Number of 6th-grade students, by school[a]	Saitama Prefecture School Handbook (published by Saitama prefectural office)
Number of NAAA attendees, by school[a]	NAAA
Number of primary school students, by municipality[b]	School Basic Survey (published by Saitama prefectural office)
Parental school choice	Our own survey
Population, by municipality and by age[c]	Saitama Population survey (published by Saitama prefectural office)
Primary education expenditures, by municipality[b]	Local Education Expenditure Survey (published by Saitama prefecutral education board)
School autonomy in budget allocations	Survey on Education Boards (conducted by the Ministry of Education)
School autonomy in curriculum development	Survey on Education Boards (conducted by the Ministry of Education)
School behavior[d]	NAAA School Questionnaire
School environment	NAAA School Questionnaire
School inspection by education boards	Survey on Education Boards (conducted by the Ministry of Education)
Students' home environment	NAAA Student Questionnaire
Students' non-cognitive skills	NAAA Student Questionnaire

Notes: (a) To calculate a take-up rate. (b) To obtain municipal education expenditure per primary student. (c) Used for analysis on selective migration. (d) On accountability to the public and the use of NAAA as a diagnostic tool.

B Supplementary results

Table 12 shows all the coefficients of control variables in Column 4 of Table 4 (though coefficients on year and school fixed effects and constant are not shown for brevity). We here comment briefly on coefficients on those additional controls. First, coefficients on student's home environment variables are all positive and highly significant regardless of the subjects: scores of students who eat breakfast, wake up and sleep at a regular time, and use a tutoring school are higher. Regarding coefficients on school environment variables, when a proportion of students who receive financial support is higher, test scores of students are lower. While there is some indication that a reduction in a class size is related to higher scores, the proportion of teachers with limited experience is not associated with scores. Turning to city-level time-varying factors, there is a positive and significant association between city-level personnel expenditure per student and scores particularly in Japanese (Basics) and Mathematics (Applications). Note that taking account of the possibility that schools given autonomy in allocation of education-related budget may spend it more efficiently, the remaining current and capital expenditures are interacted with autonomy in budgetary allocation (personnel spending is normally not authorized at the school level, thus not interacted). And it is interesting to see that the interaction between capital expenditure and school autonomy in budget is often significantly positive. The frequency of school visit by education board members does not seem to be related to test scores. However, there is some indication that schools having autonomy in curriculum are associated with higher scores. Last, having tested if city-level controls (including both expenditure and institutional variables) are jointly significant, we confirm that they generally are, with the p-values being less than 0.1 except for Math (Applications), for which they are only marginally insignificant with the p-value of 0.14.

Table 12: Detailed results of Column 4 of Table 4

	Japanese		**Mathematics**	
	Basics	Applications	Basics	Applications
Treat*After	1.122***	1.165***	0.813***	0.886***
	(3.441)	(3.853)	(2.823)	(3.211)
Treat	-2.503***	-1.118***	-1.975***	0.359
	(-9.255)	(-4.410)	(-6.664)	(1.475)
Student's home environment				
Eat breakfast	2.733***	2.472***	2.870***	2.432***
	(51.306)	(51.201)	(50.352)	(50.254)
Wake up at a regular time	0.865***	0.897***	0.901***	0.833***
	(21.282)	(23.257)	(21.889)	(22.742)
Sleep at a regular time	0.959***	0.940***	0.989***	0.913***
	(27.962)	(28.073)	(28.953)	(28.727)
Use a tutoring school	0.453***	0.078	0.850***	0.220***
	(8.057)	(1.502)	(13.372)	(3.622)
School environment				
Prop of students with fin support	-0.021***	-0.020***	-0.025***	-0.022***
	(-3.614)	(-3.753)	(-4.042)	(-4.083)
Class size in the previous year	-0.016*	-0.018**	-0.013	-0.014*
	(-1.849)	(-2.089)	(-1.385)	(-1.839)
Prop of teachers exp under 5 yrs	0.001	0.000	-0.003	0.002
	(0.493)	(0.086)	(-1.146)	(0.894)
City expenditure and institutions				
Personnel exp	0.012**	0.006	0.010	0.014**
	(2.151)	(0.953)	(1.404)	(2.520)
Other current exp	0.005	0.002	0.001	-0.001
	(1.356)	(0.588)	(0.289)	(-0.395)
Capital exp	0.000	0.000	-0.001	-0.000
	(0.049)	(0.112)	(-0.880)	(-0.650)
Other current exp*Auto in budget	0.009	0.004	-0.006	0.007
	(1.231)	(0.563)	(-0.934)	(0.980)
Capital exp*Auto in budget	0.000	0.004**	0.005**	0.001
	(0.080)	(1.998)	(2.197)	(0.696)
School visit	0.005	-0.012	0.006	-0.017
	(0.206)	(-0.534)	(0.207)	(-0.691)
Autonomy in budget	-0.476	-0.423	0.678	-0.505
	(-0.746)	(-0.677)	(1.116)	(-0.834)
Autonomy in curriculum	0.173*	0.153**	0.011	0.067
	(1.746)	(2.010)	(0.110)	(0.826)
Joint significance, Exp & Inst (p-value)	0.0137	0.0126	0.0784	0.1422
Year fixed effects	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes
Adjusted R2	0.0663	0.0597	0.0639	0.0582
Observations		220,345		

Notes: Estimates of Eq.1. Constant and coefficients on year and school fixed effects are not shown for brevity. Robust t-statistics in parentheses. Clustered standard errors are used to adjust for correlation of error terms within school. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 13: Descriptive statistics of school behavior

Variable	Mean	Std. Dev.	Min.	Max.	N
Accountability	1.13	0.62	0	2	1871
Diagnostic tool	1.4	0.55	0	2	2171
Take-up rate	0.98	0.03	0.81	1	1805

Notes: For accountability and diagnostic tool, 2 means "very well", 1 "reasonably well", 0 "hardly", to the aforementioned questions from the NAAA's school questionnaire. Take up rate is the share of the number of assessment takers to the total number of students in the school.

Table 14: Using Cities A-D as treatment cities

	Treatment group: consists of				
	All 4 cities	**City A only**	**City B only**	**City C only**	**City D only**
Japanese (Basics)					
Treat*After	0.573***	1.700***	0.596	0.469	0.157
	(3.111)	(3.780)	(1.514)	(1.608)	(0.553)
Japanese (Applications)					
Treat*After	0.454**	1.396***	1.010**	0.127	0.094
	(2.416)	(3.208)	(2.545)	(0.370)	(0.368)
Math (Basics)					
Treat*After	0.143	0.935*	0.713***	-0.038	-0.312
	(0.610)	(1.843)	(2.806)	(-0.082)	(-0.883)
Math (Applications)					
Treat*After	0.188	0.563	1.202***	-0.260	0.000
	(1.010)	(1.252)	(4.884)	(-0.794)	(0.001)
Year fixed effects	Yes	Yes	Yes	Yes	Yes
School fixed effects	Yes	Yes	Yes	Yes	Yes
Student environment	Yes	Yes	Yes	Yes	Yes
School environment	Yes	Yes	Yes	Yes	Yes
City exp & institutions	Yes	Yes	Yes	Yes	Yes
Observations	244,110	214,417	214,905	221,319	220,400

Notes: DD estimates of β_3 (cf. Eq.1). Dependent variable is standardized test score. Robust t-statistics in parentheses. Clustered standard errors are used to adjust for correlation of error terms within school. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 15: Analysis on selective migration

	Dependent variable: city-population of		
	11 years old	**7-11 years old**	**0-11 years old**
Treat*After	-20.209	-155.469	-369.447
	(-0.822)	(-1.657)	(-1.639)
Year fixed effects	Yes	Yes	Yes
City fixed effects	Yes	Yes	Yes
Observations	276	276	276

Notes: DD estimates. Robust t-statistics in parentheses. Clustered standard errors are used to adjust for correlation of error terms within city. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 16: Descriptive statistics of non-cognitive traits

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Self-confidence	2.04	0.90	0	3	220,224
Grit (perseverance)	2.53	0.86	0	3	220,216
Compassion	2.76	0.55	0	3	220,170
Respect for rules	2.35	0.65	0	3	220,209
Self-Control	1.80	0.96	0	3	220,233

Notes: 3 means "yes, absolutely", 2 "yes, tentatively", 1 "no, tentatively", and 0 "no, absolutely", to the aforementioned questions from the NAAA's student questionnaire.

References

ANDRABI, T., J. DAS, AND A. I. KHWAJA (2017): "Report cards: The impact of providing school and child test scores on educational markets," *American Economic Review*, 107, 1535–1563.

BERGBAUER, A. B., E. A. HANUSHEK, AND L. WOESSMANN (2018): "Testing," NBER Working Paper 24836 (revised November 2019).

BLAZAR, D. AND M. A. KRAFT (2017): "Teacher and Teaching Effects on students' Attitudes and Behaviors," *Educational Evaluation and Policy Analysis*, 39, 146–170.

BURGESS, S., D. WILSON, AND J. WORTH (2013): "A natural experiment in school accountability: The impact of school performance information on pupil progress," *Journal of Public Economics*, 106, 57–67.

CAMARGO, B., R. CAMELO, S. FIRPO, AND V. PONCZEK (2018): "Information, market incentives, and student performance: Evidence from a regression discontinuity design in Brazil," *Journal of Human Resources*, 53, 414–444.

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

CULLEN, J. B. AND R. REBACK (2006): "Tinkering Toward Accolades: School Gaming under a Performance Accountability System," in *Improving School Accountability Check-Ups or Choice*, ed. by T. Gronberg and D. Jansen, Advances in Applied Microeconomics 14, 1–34.

FIGLIO, D. AND L. GETZLER (2006): "Accountability, Ability and Disability: Gaming the System?" in *Improving School Accountability Check-Ups or Choice*, ed. by T. Gronberg and D. Jansen, Advances in Applied Microeconomics 14, 35–49.

FIGLIO, D. AND S. LOEB (2011): "School Accountability," in *Handbook of the Economics of Education*, ed. by E. A. Hanushek, S. Machin, and L. Woessmann, Amsterdam: Elsevier, vol. 3, 383–421.

HANUSHEK, E. A. AND M. E. RAYMOND (2005): "Does school accountability lead to improved student performance?" *Journal of policy analysis and management*, 24, 297–327.

HASTINGS, J. S. AND J. M. WEINSTEIN (2008): "Information, school choice, and academic achievement: Evidence from two experiments," *Quarterly Journal of Economics*, 123, 1373–1414.

KONING, P. AND K. VAN DER WIEL (2012): "School responsiveness to quality rankings: An empirical analysis of secondary education in the Netherlands," *De Economist*, 160, 339–355.

——— (2013): "Ranking the schools: How school-quality information affects school choice in the Netherlands," *Journal of the European Economic Association*, 11, 466–493.

MIZALA, A. AND M. URQUIOLA (2013): "School markets: The impact of information approximating schools' effectiveness," *Journal of Development Economics*, 103, 313–335.

NEAL, D. AND D. W. SCHANZENBACH (2010): "Left behind by design: Proficiency counts and test-based accountability," *Review of Economics and Statistics*, 92, 263–283.

OECD (2015): "Education at a Glance 2015: OECD Indicators," OECD Publishing, Paris, http://dx.doi.org/10.1787/eag-2015-en.

REBACK, R. (2008): "Teaching to the rating: School accountability and the distribution of student achievement," *Journal of Public Economics*, 92, 1394–1415.

REBACK, R., J. ROCKOFF, AND H. L. SCHWARTZ (2014): "Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind," *American Economic Journal: Economic Policy*, 6, 207–241.

ROCKOFF, J. AND L. J. TURNER (2010): "Short-run impacts of accountability on school quality," *American Economic Journal: Economic Policy*, 2, 119–147.

ROSENKVIST, M. A. (2010): "Using student test results for accountability and improvement: A literature review," OECD Education Working Papers, No. 54.

ROUSE, C. E., J. HANNAWAY, D. GOLDHABER, AND F. DAVID (2013): "Feeling the Florida Heat? How low-performing schools respond to voucher and accountability pressure," *American Economic Journal: Economic Policy*, 5, 251–281.

WOESSMANN, L. (2005): "The effect heterogeneity of central examinations: evidence from TIMSS, TIMSS-Repeat and PISA," *Education Economics*, 13, 143–169.

WORLD BANK (2018): *World Development Report 2018: Leaning to realize education's promise*, Washington, DC: World Bank.

ZHOU, K. (2016): "Non-cognitive skills: Definitions, measurement and malleability," Paper commissioned for the Global Education Monitoring Report 2016, Education for people and planet: Creating sustainable futures for all, UNESCO.