Can Public Rankings Improve School Performance?

Evidence from a Nationwide Reform in Tanzania

Jacobus Cilliers
Isaac M. Mbiti
Andrew Zeitlin

ABSTRACT

In 2013, Tanzania introduced "Big Results Now in Education" (BRN), a low-stakes accountability program that published both nationwide and within-district school rankings. Using data from the universe of school performance for 2011–2016, we identify the impacts of the reform using a difference-in-differences estimator that exploits the differential pressure exerted on schools at the bottom of their respective district rankings. We find that BRN improved learning outcomes for schools in the bottom two deciles of their districts. However, the program also led schools to strategically exclude students from the terminal year of primary school.

Jacobus Cilliers is at Georgetown University (ejc93@georgetown.edu). Isaac M. Mbiti is at University of Virginia, J-PAL, BREAD, NBER, and IZA (imbiti@virginia.edu). Andrew Zeitlin is at Georgetown University and Center for Global Development (az332@georgetown.edu). *The authors are grateful to Shardul Oza, Wale Wane, Joseph Mmbando, Youdi Schipper, and Twaweza for their support and assistance in helping assemble the data sets used in the study. For helpful comments and suggestions the authors thank Nora Gordon, James Habyarimana, Billy Jack, Kitila Mkumbo, Lant Pritchett, Justin Sandefur, Richard Shukia, Abhijeet Singh, Jay Shimshack, Miguel Urqiola, and seminar participants at the RISE conference, Twaweza conference in Dar es Salaam, DC Policy Day, Georgetown University, and University of Virginia. Fozia Aman, Ben Dandi, Austin Dempewolff, and Anna Konstantinova provided exceptional research support. This research was funded by the Research on Improving Systems of Education (RISE) initiative of the UK DFID and Australian Aid and administered by Oxford Policy Management. The authors do not have any conflicts of interest to disclose. Data and programs for replication are available through Dataverse at https://doi.org/10.7910/DVN/ABVMCL. The replications files do not include the World Bank SDI data, which are available at https://www.sdindicators.org/. The restricted access version of the SDI data can be obtained by contacting the SDI team at sdi@worldbank.org with a statement of research objectives and a justification.*
[Submitted January 2019; accepted August 2019]; doi:10.3368/jhr.56.3.0119-9969R1
JEL Classification: I21, I25, I28, and O15
ISSN 0022-166X E-ISSN 1548-8004

I. Introduction

School performance rankings based on standardized tests are typically used as the foundation of accountability systems. Such systems are thought to be more effective if school performance is used to sanction or reward schools (Hanushek and Raymond 2005). However, there are concerns that such "high-stakes" systems can encourage unintended behaviors, including gaming, teaching to the test, and neglecting unrewarded tasks or academic subjects (Baker 1992, 2002; Holmstrom and Milgrom 1991). Further, political constraints, such as opposition from teachers, make these systems difficult to implement. As a result, the first accountability systems that are implemented tend to be "low-stakes" systems that focus simply on publicizing information about school performance.[1] Although successful low-stakes accountability reforms have taken place in contexts where parents are willing and able to act on the information provided to them (Andrabi, Das, and Khwaja 2017; Camargo et al. 2018; Koning and Van der Wiel 2012), these systems may also be effective if they create sufficient reputational pressure for higher-level education administrators or school staff (Bruns, Filmer, and Patrinos 2011; Figlio and Loeb 2011). But such pressure could be a double-edged sword, encouraging the same distortions and perverse behaviors that are associated with high-stakes systems.

In this work, we study the intended and unintended consequences of a nationwide accountability system implemented in Tanzania in 2013. In response to growing concerns about school quality, the Government of Tanzania instituted a package of reforms that were branded "Big Results Now in Education" (BRN) (World Bank 2015, 2018b). The BRN education program was a flagship reform that was overseen and coordinated by the Office of the President and implemented by the Ministry of Education. It aimed to improve the quality of education in Tanzania through a series of top-down accountability measures that leveraged the political prominence of the reforms to pressure bureaucrats in the education system.

A key BRN policy was its school ranking initiative, under which the central government disseminated information about individual primary schools' average scores in the Primary School Leaving Exam (PSLE) and their corresponding national and within-district rankings. As this was the most prominent and comprehensively implemented BRN component (Todd and Attfield 2017), we focus our study on examining the impacts of this intervention. We interpret the BRN reforms as a low-stakes accountability reform since there were no financial consequences for low school performance on the PSLE.[2] Prior to the reform, the government only released information about students who passed the PSLE. With the reform, the complete set of PSLE results, including school rankings, was released on a website and shared directly with schools through District Education Officers (DEOs), who supervise public schools within their jurisdictions and have considerable discretion over human and physical resource

1. This pattern is clearly seen in many U.S. states, as well as in the U.K., where the first step towards an accountability system was a school report card initiative (Burgess et al. 2005). Low-stakes accountability reforms built around reputational incentives for schools have also been implemented in Brazil, Chile, Colombia, Mexico, and Pakistan (Figlio and Loeb 2011).

2. We test this assertion in our data and confirm that low-performing schools saw no reductions in school resources. See Section V for more details. This contrasts with studies in other contexts, such as Craig, Imberman, and Perdue (2013, 2015), where school funding was closely tied to school performance.

decisions therein. As there were limited efforts to disseminate this information to parents—and survey data collected under our study confirm that parental awareness of these rankings was minimal—the reforms leveraged bureaucratic incentives of DEOs, head-teachers, and other education officials through top-down pressure.

To study the BRN's impacts, we assemble a novel data set that combines administrative and (matched) sample-based data from several sources to estimate the impact of the BRN reforms on a comprehensive set of school-level outcomes (Cilliers, Mbiti, and Zeitlin 2020). For our main analysis, we construct a panel of administrative data on exam outcomes, covering all Tanzanian primary schools in the period 2011–2016. To shed light on the potential mechanisms through which the reforms affect test scores, we match our administrative data on examinations to data on student enrollments in 2015 and 2016 from the government's Education Management Information System (EMIS), as well as to microdata from the World Bank's Service Delivery Indicators, a nationally representative panel of almost 400 schools in Tanzania, which were collected in 2014 and 2016.

We identify the effects of the BRN's publication of within-district school ranks using a difference-in-differences strategy that exploits the differential pressure faced by schools across the ranking distribution under BRN. In a given year, a school will be ranked using the prior year's PSLE test score. We posit that the publication of school rankings would exert more pressure on schools that fall near the bottom of the distribution in their district, where (relative) failure is more salient, compared to schools in the middle of the distribution. In this respect, our study is similar to prior work that has used the differential pressures generated by accountability reforms to study their consequences for schools and teachers (Chiang 2009; Dee and Wyckoff 2015; Figlio and Rouse 2006; Reback, Rockoff, and Schwartz 2014; Rockoff and Turner 2010; Rouse et al. 2013).[3] We operationalize this hypothesis of differential pressure in a parsimonious manner by comparing how schools respond to being ranked in the bottom two deciles relative to the middle six deciles (our reference category) in the pre-BRN (2012) versus post-BRN (2013–2016) periods.

As the prereform relationship between schools rankings and subsequent school performance may be driven in part by mean reversion (Chay, McEwan, and Urquiola 2005; Kane and Staiger 2002), our difference-in-differences strategy will use pre- and post-reform data to identify the effect of the BRN school ranking program by netting out prereform estimates of mean reversion and other stable processes through which rankings affect school performance.

As better-ranked schools have better test scores, it can be difficult to disentangle the effects of the district rankings from other effects driven by (or associated with) levels of PSLE test scores. To circumvent this potential confounding effect, we exploit between-district heterogeneity, where many schools have the same average test scores, but radically different within-district rankings in both pre- and post-BRN periods.[4] This

3. To the extent that schools elsewhere in the ranking distribution are unaffected by the reform, such a strategy provides not only a test of the existence of a BRN effect, but an estimate of its total impact. This contrasts with settings where "sharp" discontinuities identify local impacts of particular mechanisms within a given policy environment, as in Rockoff and Turner (2010), Mizala and Urquiola (2013), and others.

4. For example, a school with the 2011 national average PSLE mark of 111 could find itself anywhere from the bottom 5 percent to the top 5 percent of its district, depending on the district in which it resided. Reback, Rockoff, and Schwartz (2014) exploit similar heterogeneity in their evaluation of the No Child Left Behind reform.

allows us to compare the response of schools of similar quality (measured by test scores) that are exposed to different within-district rankings. Flexible, year-specific controls for absolute test scores absorb mean reversion or other processes that govern the evolution of absolute test scores in each year through mechanisms other than district ranks. Moreover, the construction of school-level panel data allows fixed-effect estimates to address potential time-invariant confounds.

We find that the BRN school ranking intervention increased average PSLE test scores by approximately 20 percent of a prereform school standard deviation for schools in the bottom decile relative to schools ranked in the middle six deciles. The pass rate for these schools improved by 5.7 percentage points (or 36 percent relative to the pre-BRN pass rate among bottom-decile schools), and on average two additional students from each of these schools passed the PSLE—a 24 percent increase relative to the pre-BRN number of passing students for bottom-decile schools. Our estimated effect sizes for the bottom-decile schools are similar to the experimental estimate of distributing school report cards in Pakistan (Andrabi, Das, and Khwaja 2017). Our estimates are also within the range typically found in evaluations of teacher performance pay programs and accountability schemes (Muralidharan and Sundararaman 2011; Figlio and Loeb 2011).

We explore several potential mechanisms through which the BRN reform led to these test score gains by matching administrative data on test scores with the World Bank's Service Delivery Indicator school-level panel data set. Despite the ability of District Education Officers (and other government officials) to redirect resources to bottom-ranked schools, we do not find any evidence that these schools received additional teachers, textbooks, financial grants, or inspections. In addition, we do not find any evidence of increased effort (measured by school absenteeism or time teaching) among these schools. However, we find that there were two fewer students taking the PSLE in bottom-ranked schools. This is roughly a 4 percent reduction, given prereform average class size of 48 pupils among these schools. Examining administrative enrollment data, we find similar reductions in seventh-grade enrollment in the bottom-ranked schools, suggesting students are induced to leave seventh grade altogether. We do not find statistically significant changes in enrollment in other grades, as would be implied by induced grade repetition. Further, we do not find any evidence that these seventh-grade enrollment reductions reflect students moving from bottom-ranked schools to better-ranked schools. Robustness checks support the interpretation that these school responses—both the positive learning gains and the negative enrollment effects—appear to be driven by the district rankings and not by other, contemporaneous reform components.

Our study contributes to the literature of education system accountability in three distinct ways. First, we show that despite the low-stakes nature of the school ranking initiative and the limited potential parental response to the school rankings, schools did respond to the reform. Given parents' minimal awareness of the rankings and limited scope for school choice, this suggests that top-down pressure and bureaucratic reputational incentives are capable of driving learning improvements. This is a novel finding, given that the existing literature on school ranking interventions has only showed these to be effective when there is sufficient school choice or high-stakes consequences (Andrabi, Das, and Khwaja 2017; Nunes, Reis, and Seabra 2015; Hastings and Weinstein 2008; Koning and Van der Wiel 2013; Figlio and Loeb 2011).

Second, we show that that even low-stakes accountability systems can induce perverse behavioral responses. Previous studies have shown that schools responded to accountability systems by focusing on students near the proficiency margin (Neal and Schanzenbach 2010), excluding lower-performing students from testing by suspending them or categorizing them as disabled (Figlio and Getzler 2002; Figlio 2006) or manipulating exam results outright (Jacob 2005). However, these behaviors are typically associated with high-stakes accountability systems that impose significant consequences on underperforming schools. Despite the limited consequences of their performance, bottom-ranked schools in Tanzania responded by excluding students from the assessment.

Third, we add to the limited evidence base on the effectiveness of nationwide accountability reforms in developing contexts. Our study documents the introduction of such an accountability system and provides evidence of its short-run impacts. The existing evidence on accountability reforms focuses on developed countries, and on the U.S. No Child Left Behind Act of 2002 in particular, where findings generally suggest positive test score impacts (see Figlio and Loeb 2011 for an overview). Evidence of accountability reforms in developing countries is more scarce, and is dominated by evaluations of pilot programs of teacher performance pay systems (Glewwe, Ilias, and Kremer 2010; Muralidharan and Sundararaman 2011) or bottom-up accountability mechanisms (Banerjee et al. 2010; Lieberman, Posner, and Tsai 2014). A potential drawback of these pilot studies is that the nature of their implementation—and, consequently, the incentives they create—may be very different when implemented by government at scale, and, more broadly, the estimates they deliver may fail to capture general equilibrium effects (Bold et al. 2018; Muralidharan and Niehaus 2017). Unfortunately, larger-scale reforms often have not been evaluated due to lack of a suitable control group (Bruns, Filmer, and Patrinos 2011).[5] Our work fills an important gap in this literature by documenting the promise and perils of a low-stakes bureaucratic accountability system, at scale, in a developing country.

II. Context and Reform

Following the introduction of free primary education, net enrollment rates in Tanzania steadily increased from 59 percent in 2000 to 80 percent in 2010 (Valente 2015; Joshi and Gaddis 2015). However, the surging enrollment raised concerns about the quality of education. The basis for this perceived learning crisis is readily seen in results of the Primary School Leaving Exam (PSLE), which is taken at the end of Grade 7 and serves to certify completion of primary education and to determine progression to secondary school.[6] PSLE pass rates declined from about 70

5. An important exception is the case of Chile, where, in the context of a voucher system (Hsieh and Urquiola 2006), Mizala and Urquiola (2013) use a regression discontinuity design to estimate the effect of receiving a good ranking on schools near the cutoff. However, their study focuses on market outcomes, such as enrollment and tuition. They do not examine test scores, and their study focuses on parental choice as a driving mechanism.

6. The PSLE tests students in mathematics, science, English, Kiswahili, and social studies. The maximum score on each subject is 50, and individual subject scores are then totaled for each student, up to a maximum of 250 points. A score of 100 or above constitutes a passing score. For more details see https://www.necta.go.tz/psle (accessed November 25, 2020).

percent in 2006 to an all-time low of 31 percent in 2012 (Todd and Attfield 2017; Joshi and Gaddis 2015). At the same time, independent learning assessments highlighted that only about one in four children in the third grade could read at a second-grade level (Twaweza 2012; Jones et al. 2014).

In response to these challenges facing the education sector, the Government of Tanzania launched a large-scale, multipronged education reform called Big Results Now in Education (BRN) in 2013, which aimed to raise pass rates on the PSLE to 80 percent by the year 2015.[7] The BRN reforms emphasized exam performance and created pressure to demonstrate improvements across the education system. Government ministries were instructed to align their budgets to the reforms, which were coordinated through the Presidential Delivery Bureau, a specialized division of the Office of the President. In a public ceremony, the Education Minister and the Minister for Local Government both pledged that their ministries would endeavor to meet the BRN targets. In addition, national, regional, and district officials signed performance contracts to signal their commitment to the reforms (Todd and Attfield 2017). District officials who oversaw implementation were required to submit regular progress reports to higher level officials, as well as the Delivery Bureau. The Bureau would then provide feedback and recommendations on the reports. Overall, this structure ensured that there was sufficient political pressure on government bureaucrats to implement the reforms.

The BRN education reforms comprise nine separate initiatives that ranged in approach from school management–focused interventions to initiatives aimed at improving the accountability environment, such as the school ranking program. In spite of their systemic national ambitions, as we document in Table B.1 in the Online Appendix, most of these initiatives were either implemented on a limited scale (as with the remedial education program), rolled out slowly (as with the capitation grant reform), or designed in such a way that they did not create effective incentives for schools (as with the incentives for year-on-year test score gains).[8] Other initiatives focused on early grade learning and were thus not relevant for our study. We therefore focus on the component that was not only implemented universally and immediately, but also central to BRN's emphasis on learning outcomes—the school ranking initiative.

BRN's school ranking initiative disseminated information about each school's average score on the PSLE relative to all other schools in the country (the national ranking), as well as rankings relative to schools in the district (the district ranking). Prior to BRN, District Education officers (DEOs) and schools were only provided with the list of students who passed the PSLE. For the first time, the reform distributed the complete set of results for all students and included school rankings and average scores. This information was posted on the internet, published in print media, and distributed to DEOs. Nearly all DEOs disseminated exam results and district rankings to the 90–110 primary schools within their district (Todd and Attfield 2017). Surveys we conducted with DEOs in 38 districts show that in 2016 more than 90 percent of DEOs who had received the ranking information by the time of our survey held meetings with head teachers to inform them of their rankings and to discuss ways to improve their performance.

7. The BRN program was introduced as part of larger set of reforms modeled on Malaysia's Big Fast Results program.

8. Many of the implementation problems were due to a lack of funding (Todd and Attfield 2017).

Phone surveys conducted with a nationally representative sample of 435 head teachers and 290 school management committee chairs in 45 districts in 2016 reveal that head teachers were better informed about their school's within-district rank compared to their national rank.[9] Almost 75 percent of surveyed head teachers were able to provide an estimate of their school's district rank, but fewer than 20 percent of head teachers could provide an estimate of their national rank. For this reason, and based on evidence we present in Section VI, we focus on estimation of the impacts of BRN's district rankings in particular.

Unlike head teachers and DEOs, parents were not well informed about the school's rankings. Only 2 percent of surveyed school management committee chairs—who are plausibly the most informed parents in the school—were able to provide an estimate about their school's national rank; 22 percent of committee chairs could provide an estimate of their school's district rank. A recent qualitative study reached a similar conclusion, finding that parental understanding of the BRN school ranking initiative was much lower compared to head teachers, DEOs, and teachers (Integrity Research 2016). This is not surprising, since there was no national dissemination strategy to parents, and the district-level dissemination was targeted at head teachers, not parents (Integrity Research 2016). As a result, any estimated effects of the ranking program were arguably driven by school and district responses, rather than parents.

In Tanzania's partially decentralized education system, there are many possible channels through which public sharing of district ranks can improve school performance, even when parental awareness is limited. DEOs faced substantial pressure to demonstrate improvement in their districts.[10] DEOs receive funds from the central government to finance services and projects, and they are also responsible for monitoring, staff allocations, and the transfer of capitation grants to schools.[11] In principle, they could use the district ranks to reallocate resources towards underperforming schools, provide additional support, or place pressure on schools to improve. Our interviews with DEOs revealed that some DEOs would gather all head teachers in a room and publicly berate bottom-ranked schools, and a recent report provides anecdotal evidence that some DEOs organized additional training on how to conduct remedial and exam preparation classes (Integrity Research 2016). For their part, head teachers—driven by career concerns, professional norms, or competition with other district schools—could have taken a variety of steps to improve exam performance in order to demonstrate their competence and work ethic to their peers or superiors. Though our

9. During the phone surveys we asked DEOs and head teachers to provide us with their school's most recent district and national rank. Many respondents could not even provide an estimate of their ranks. For the respondents who did provide an estimate, we then compared their reports to the actual data to assess the accuracy of their reports. Respondents' accuracy was higher for district rankings compared to national rankings. More information about the data collection can be found in a report by the RISE Tanzania Country Research Team (2017).

10. A DFID report concluded that "[t]here was a clear sense that…district officials would be held accountable for these pass rates, with some officials stating that BRN should stand for 'Better Resign Now' because they did not believe it would be possible to achieve these ambitious targets with their current level of resourcing" (Todd and Attfield 2017, p. 22).

11. Prior to 2016, the central government sent capitation grants to districts, which then transferred these funds to schools. According to government policy, each school was supposed to receive 10,000 TSh, per student per year, roughly US$4.5, but in practice there was wide variation in the amount disbursed and uncertainty over the disbursement schedule (Twaweza 2013, 2010).

data provide limited opportunities to test among alternative—and potentially complementary—channels, we will analyze the impacts of school-level responses on learning, as well as several margins through which this might be achieved, in Section V.

III. Data and Descriptive Statistics

We compiled and matched multiple sources of administrative and survey data to estimate the impact of the BRN school ranking initiative on primary school outcomes. To begin, we scraped the National Examinations Council of Tanzania (NECTA) website for school-level data on performance in the Primary School Leaving Examination (PSLE). The data include average school test scores and national and district rankings for all primary schools for 2011–2016, linked over time. We do not have access to student-level or subject-specific scores, or any data prior to 2011.[12] In addition to test scores, the NECTA data set also contains information on the number of test-takers at each school, the number of students who passed the exam, and the pass rate for the school—although these measures are only available for the years 2012–2016.

We augment the data on PSLE test scores with administrative data from the Education Management Information System (EMIS), which contains data on student enrollment and the number of teachers for all schools in 2015 and 2016.[13] Since EMIS uses a different unique identifier, we manually matched schools from EMIS to NECTA data, using information on school location and school name. We were able to match 98.9 and 99.7 percent of schools from the NECTA data in 2015 and 2016, respectively.

In addition, we use micro-data from the World Bank Service Delivery Indicators (SDI) survey in Tanzania to supplement our analysis. The SDI is a nationally representative panel survey of 400 schools in Tanzania collected in 2014 and 2016 (World Bank 2016a,b,c).[14] The survey measures teaching practices, teacher absence, and school resources. In addition, these data include the results of a learning assessment administered to a representative sample of fourth-graders in SDI schools.

Table 1 shows basic descriptive statistics of our different data sources: NECTA, EMIS, and SDI. The NECTA data in Panel A show that average test scores and national pass rates, both computed at the school level, dropped dramatically from 2011 to 2012, but steadily increased from 2013 to 2016. Despite population growth and an expansion in the number of primary schools from roughly 15,000 in 2011 to more than 16,000 in 2016, our data show the number of test-takers monotonically decreased from 2011 to 2015, with only a small increase in 2016 relative to 2015.[15] Over our study period, the

12. The government issued schools with a unique ID for examination purposes. To validate that this ID consistently relates to the same school over time, we also performed a fuzzy merge by school location and school name, across all the years for which we have data. We identified a small number of schools where the unique IDs were "recycled" from schools that had closed down, or where the unique ID did not consistently relate to the same school. We constructed our own unique IDs for these cases.

13. The ministry responsible for housing the EMIS data changed in 2015, and consequently, data prior to that year are not available.

14. There is some very minor attrition in the data as we are able to match 99 percent (395/400) of schools in the SDI data to the NECTA data. Only 388 and 392 of these schools were in operation in 2012 and 2013, respectively.

15. The closing down of schools is rare: 99.2 percent of schools that existed in 2011 remain in our sample in 2016.

Table 1
Descriptive Statistics

	2011	2012	2013	2014	2015	2016
Panel A: NECTA						
Average marks	111.31	85.96	102.76	108.65	120.41	119.17
	(26.76)	(22.28)	(24.51)	(26.14)	(30.17)	(27.60)
Average number of test-takers		56.34	53.97	49.92	47.44	48.29
		(37.82)	(37.92)	(36.44)	(36.08)	(37.49)
Average number of passed candidates		17.31	27.31	28.45	32.18	33.97
		(22.89)	(28.95)	(28.67)	(30.48)	(31.42)
Average pass rate		0.29	0.49	0.56	0.67	0.69
		(0.25)	(0.28)	(0.28)	(0.27)	(0.25)
Total number of test-takers	1,010,084	865,534	844,921	792,118	763,602	789,236
National pass rate	0.58	0.31	0.51	0.57	0.68	0.70
Number of schools	14,939	15,362	15,656	15,867	16,096	16,344
Number of districts	136	136	151	166	166	184
Panel B: EMIS Data						
Private school					0.05	0.05
					(0.22)	(0.22)
Grade 4 enrollment					67.32	63.46
					(50.76)	(49.45)
Grade 5 enrollment					63.33	61.83
					(47.68)	(46.71)

(*continued*)

Table 1 (*continued*)

	2011	2012	2013	2014	2015	2016
Grade 6 enrollment					61.68	61.58
					(47.23)	(45.93)
Grade 7 enrollment					49.36	48.82
					(38.99)	(38.11)
Total number of schools					16,178	16,351
Panel C: Service Delivery Indicators						
Average classroom presence				0.52		0.59
				(0.23)		(0.22)
Average number of teachers				17.47		17.48
				(1.73)		(2.39)
Capitation grants received (TSh/student)			5,134		5,966	
			(3,891)		(17,796)	
Textbooks received per student			0.12		0.63	
			(0.45)		(0.84)	
Number of inspections			1.56		1.45	
			(2.65)		(1.74)	
Number of schools				395		395

Notes: Panel A shows school-average examinations performance, collected by the National Examination Council of Tanzania (NECTA); Panel B reports enrollment data, according to the Education Management Information System (EMIS). School-level EMIS data are only available for 2015 and 2016. Panel C reports the summary data for key variables used from the Service Delivery Indicators (SDI) data set. The SDI data were collected in 2014 and 2016, but some variables were collected using the previous year (2013 or 2015) as the reference period. Means and standard deviations reported, unless otherwise noted. Capitation (or per-pupil) grants are reported in nominal Tanzanian shillings (TSh).

government increased the number of districts from 136 in 2011 to 184 in 2016.[16] Due to this redistricting, we construct district ranks for each school using the district that the school belonged to corresponding to the PSLE exam year.[17] Our results are qualitatively similar if we construct the district rank based on the district the school belongs to in the following year.

EMIS data in Panel B show that average Grade 7 enrollment is similar to the average number of test-takers the same year reported in Panel A.[18] Almost all students enrolled in Grade 7 therefore end up taking the exam. The data also show a large drop in enrollment between Grades 6 and 7, implying a dropout rate of approximately 20 percent between 2015 and 2016.

Panel C reports summary statistics from the SDI data. Although the SDI was collected in 2014 and 2016, the reference period for certain data, such as inspections and school resources, was the previous calendar year (2013 and 2015, respectively). Other data, such as enrollment figures, were collected on a contemporaneous basis. The data show teachers were often not in their classrooms. During unannounced visits to schools in 2014, only 52 percent of teachers were in their classrooms; this improved slightly to 59 percent in 2016. Capitation grant receipts by schools rose from just over 5,000 Tanzanian shillings (TSh) per student in 2013 to almost TSh 6,000 per student in 2015. This is substantially lower than the stated government policy of providing schools with grants of TSh 10,000 per student per year. In addition, there is a high degree of variation in the amount of grant funding received: 8 percent of schools reported that they had received the full amount, and 4 percent reported that they had received nothing. Finally, the data show that schools were inspected on average 1.56 and 1.45 times in 2013 and 2015, respectively, and roughly 70 percent of schools received at least one inspection.

IV. Empirical Strategy

We exploit the differential pressure exerted by BRN on schools at the low end of their district ranking to test for and estimate the impacts of BRN's district ranking initiative on school outcomes. This strategy addresses a fundamental challenge to evaluating the school ranking program, as all schools in the country were simultaneously exposed to this policy. In this respect, our approach is similar in spirit to recent studies that adopt such a strategy to evaluate the effects of accountability reforms in the U.S. context (Chiang 2009; Dee and Wyckoff 2015; Figlio and Rouse 2006; Reback, Rockoff, and Schwartz 2014; Rockoff and Turner 2010; Rouse et al. 2013). Using a panel data set comprising the universe of primary schools and spanning multiple years both before and after the reform, we estimate the impact of BRN school rankings with

16. We account for the changes in district boundaries by including year-specific district fixed effects in our analysis. We take this approach since it better conforms to the key mechanisms such as District Education Officer actions, which would not be well captured if we aggregated our district fixed effects to the original larger districts.

17. The motivation for our chosen definition is that schools face pressure to improve in the subsequent year, on the basis of their performance in a given year. We opted for this definition because district ranks reported in a phone survey of head teachers more closely corresponded to this construction.

18. On average, schools had 2.6 and 0.7 more Grade 7 students enrolled than exam-sitters in 2015 and 2016, respectively. The 2016 data are more reliable.

a difference-in-differences model. This allows us to test and account for possible consequences of district ranks in the pre-BRN period.

Our empirical strategy exploits the remarkable across-district heterogeneity in test scores. Two schools (in different districts) with the same average test scores could have very different district rankings due to this heterogeneity. This allows us to identify the effects of a school's within-district ranking separately from its absolute performance. Naturally, absolute performance levels (or average test scores) are a potential confound for the effects of within-district rank. Not only do absolute performance levels contain information about persistent aspects of school quality, but it may also be the case that schools that perform (say) poorly in absolute terms in a given year may rebound in the subsequent year due to mean reversion. As illustrated in Online Appendix Figure A.1, the heterogeneity across Tanzania's districts creates almost complete overlap in the support of district rankings at several values of absolute test scores in each of the years 2011–2016. In 2011, for example, a school with the national average mean PSLE score of 111 could have ranged from being in the bottom 5 percent of schools in its district to being in the top 5 percent of schools, depending on the district in which it was located. This allows us to condition flexibly on absolute test scores while estimating the effects of district rankings. This set of flexible controls allows school test scores to follow different paths from year to year, as the relationship between test scores in year $t-1$ and t is allowed to vary with t.

Given this variation, we use a difference-in-differences model to estimate the impact of being in the extreme deciles of the district ranking in the post-BRN period on schools' exam performance in the subsequent year. This specification is provided in Equation 1:

$$(1)\quad y_{sdt} = f_t(y_{s,t-1}) + \sum_{q\in\{1,2,9,10\}} \alpha_q I_q(r_{s,t-1}) + \sum_{q\in\{1,2,9,10\}} \beta_q Post_t \cdot I_q(r_{s,t-1}) + \gamma_{dt} + e_{sdt}.$$

Here, y_{sdt} is the mean exam performance (test score) of school s in district d and year t. The function $f_t(y_{s,t-1})$ represents a set of year-specific flexible controls for the lagged test score (up to a fourth-order polynomial) of school s. We allow this relationship between past and current performance to vary year by year, to account for possible differences in the scaling and predictive content of each year's exams. We denote by $r_{s,t-1}$ the within-district rank of school s in the prior year. I_q is an indicator function that is equal to one if a school is ranked in decile q in its district, and $Post_t$ indicates whether the outcome is in the postreform period.[19] We include district-by-year fixed effects, γ_{dt}, and $e_{s,d,t}$ denotes an idiosyncratic error term. In this and subsequent specifications, we cluster the standard errors at the district level to allow arbitrary patterns of correlation across schools within a district or within schools over time.

The difference-in-differences model in Equation 1 compares the relationship between within-district school rankings and subsequent school outcomes in both post-BRN periods with the same relationship in the pre-BRN period. In both the pre- and post-BRN period, we compare performance in the bottom two deciles of district ranks with that of schools falling in the middle 60 percent of schools. Although our empirical specification

19. An uninteracted $Post_t$ indicator would be absorbed by the combination of our controls. Therefore, we do not include it in our specification.

includes indicators for both bottom- and top-ranked schools, we focus our discussion on the results pertaining to the bottom-ranked schools for conciseness.[20] We attribute the impact of the BRN reform only to the *difference* in the estimated relationship between ranks and subsequent performance for the pre- versus post-BRN periods. This is estimated by the set of parameters β_q.

A correlation between district rank and subsequent-year PSLE performance in the prereform period might arise for several reasons. DEOs might have exerted pressure on bottom-ranked schools even before BRN was introduced, and schools may have also been intrinsically motivated to improve their performance. In the specification of Equation 1, these relationships are captured in the estimated prereform ranking effects, α_q. Although this model (due to the inclusion of lags) can only be estimated for one prereform year, it does allow a useful test. Specifically, if the prereform rank effects, α_q, are jointly insignificant, then this would suggest that within-district ranks had little empirical consequence prior to the reform.

For a set of secondary, exam-related outcomes—the pass rate, number passed, and number of exam-sitters—we have data for one only one prereform period (2012), so we cannot control for the lagged dependent variable when estimating a difference-in-difference model. In these cases, we continue to control flexibly for lagged average exam scores, $f_t(y_{s,t-1})$, as in Model 1, to identify the effect of district rank separately from absolute academic performance. To address any remaining potential confounds from persistent school-specific characteristics, such as the size of the school's catchment population, we augment this model to allow for school fixed effects, to estimate

$$x_{sdt} = f_t(y_{s,t-1}) + \sum_{q\in\{1,2,9,10\}} \alpha_q I_q(r_{s,t-1}) + \sum_{q\in\{1,2,9,10\}} \beta_q Post_t \cdot I_q(r_{s,t-1}) + \gamma_{dt} + \mu_s + e_{sdt}. \quad (2)$$

In this specification, x_{sdt} are these secondary outcomes, and μ_s refers to school-level fixed effects. These school fixed effects absorb any time-invariant sources of correlations between schools' district ranks and the outcome of interest that might not be fully captured by the lagged test score. Standard errors for this specification are also clustered at the district level.

Finally, we use SDI and EMIS data on a range of outcomes, such as enrollment, school resources, and teacher absences, to test for mechanisms. All of these data sets contain multiple years of data, but they do not contain data from the prereform period. In these cases, we estimate a fixed-effects model of the form in Equation 2. While we are unable to difference out the prereform ranking effects, q, we continue to control flexibly for lagged test scores in order to isolate a plausible causal effect of district rankings, taking advantage of the fact that schools with the same test score can have different rankings in alternative districts.[21] Conservatively, the estimates for these outcomes can

20. Our interest here is primarily in impacts of BRN-related exposure on low-ranked schools, but by including indicators for the top two deciles, we avoid relying on the assumption that these can be pooled with schools in the middle of their district distribution. To demonstrate that point estimates are not driven by this choice of category, we also present results graphically where only the fifth and sixth decile serve as the omitted category in Figure 1. We use the parsimonious specification to increase the statistical power of our analysis, but this choice does not substantively effect point estimates.

21. Reback, Rockoff, and Schwartz (2014) use a similar empirical strategy to evaluate the accountability pressure generated by the No Child Left Behind reforms in the United States.

be interpreted as the sum of prereform ranking effects and the postreform change (that is, $\alpha_q + \beta_q$ from Equation 1). If the prereform coefficients, α_q, are truly zero for these outcomes, then these estimates can be interpreted as BRN effects. Our confidence in this assumption will be strengthened to the extent that prereform ranking effects, α_q, are jointly insignificant for test score outcomes. We can also be more confident in this assumption if we find similar results in closely related but longer time series. For example, we can compare and confirm that the results using the number of exam-takers as an outcome are comparable to those using enrollment as an outcome (see Table 3). Moreover, in Section VI, we will present tests that split the sample by school size in both double-difference and single-difference models, to test for mean reversion. We find no evidence that effects are stronger for smaller schools, where mean reversion is expected to be stronger (Kane and Staiger 2002).

V. Results

Below we present results for impacts of the reform. After providing evidence that BRN improved exam results for schools in the bottom of their district ranks, we explore several mechanisms that may underlie this result. We test for impacts on the number of test-takers and on enrollment, finding evidence consistent with BRN-induced drop-out. We find no evidence of impact on a range of educational inputs or on learning in lower grades.

A. Pressure on Low-Ranked Schools Improves Low-Ranked Schools' Average Exam Performance and Pass Rates

Figure 1 illustrates the basis for the difference-in-difference estimates, focusing on the estimation of Equation 1 for two main outcomes of interest: the average PSLE score, which is the basis for the ranking itself, and pass rate, which motivated the adoption of the reform. In Figure 1, Panels A and C, we show estimates and 90 percent confidence intervals for the impacts of being in decile q of the district rank, *relative* to the schools between the 40th and 60th percentiles in their district, on a school's average test scores and pass rates, respectively, in the subsequent year. We estimate coefficients separately for the prereform period (the light lines and squares, corresponding to parameters α_q in Equation 1) and postreform periods (the dark lines and circles, corresponding to the sum of parameters $\alpha_q + \beta_q$). Estimates control for a fourth-order polynomial in the absolute lagged test score, as well as district–year indicators. Our estimates of the consequences of the reform for the relative performance of bottom- and top-performing schools within their districts is given by the *difference* between these pre- and postreform ranking coefficients. These differences, and associated confidence intervals, are shown directly in Figure 1, Panels B and D.

There is no statistically significant relationship between within-district ranks and subsequent performance in the pre-BRN period. As Panels A and C in Figure 1 illustrate, point estimates are closer to zero for the pre-period. Regression estimates of Equation 1 confirm that these pre-BRN ranking effects are both individually and jointly insignificant. An F-test for the joint significance of the coefficients α_q fails to reject

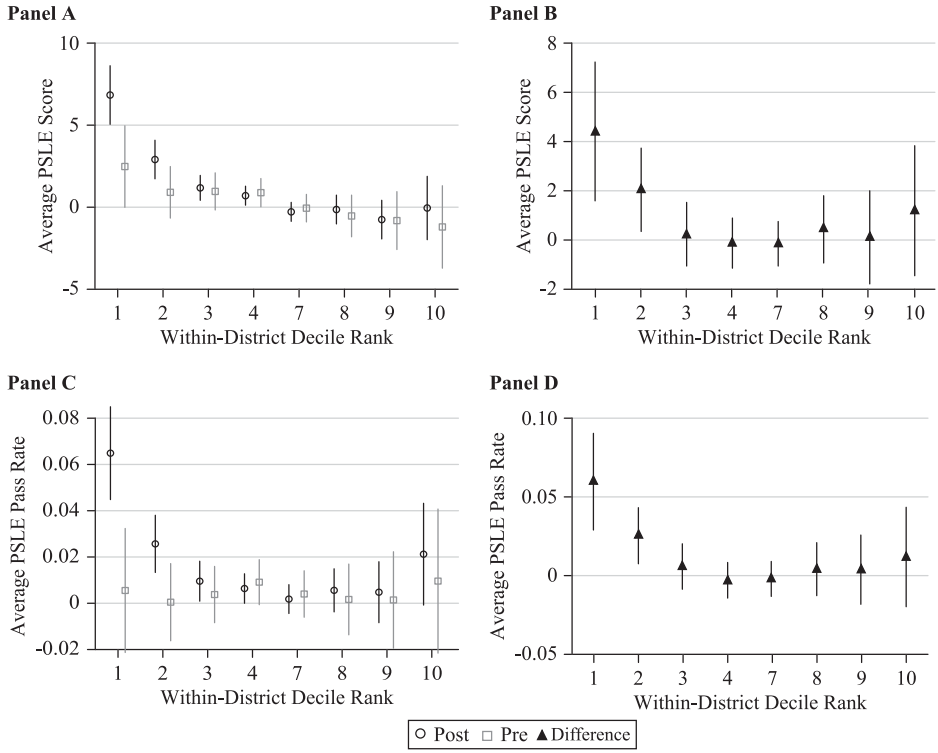

Figure 1

Exam Performance by Within-District Decile Rank—Pre- vs. Postreform

Notes: All panels show regression coefficients and 90 percent confidence intervals, estimated using Equation 1. In Panels A and C the light lines refer to the prereform period ($\hat{\alpha}_q$), and the dark lines refer to the postreform period ($\hat{\alpha}_q + \hat{\beta}_q$). Panels B and D shows results for the *difference* in these ranking decile effects between pre- and post periods (β_q). In both the pre- and postreform periods, schools are compared to schools in the middle two deciles of the district rank. In Panels A and B the outcome is a school average exam performance, scaled from zero to 250; in Panels C and D the outcome is pass rate.

the null that they are equal to zero, with a p-value of 0.28 and 0.78 when the outcome variable is the average score and the pass rate, respectively. We do not need to assume these prereform ranking effects are zero to identify test score impacts, since the difference-in-difference specification of Equation 1 would also difference out any prereform relationship.[22] But given that our analysis of potential mechanisms will rely on data sets that do not include the prereform period, this absence of a prereform relationship strengthens our confidence in a causal interpretation of those estimates as well.

22. If this difference-in-differences approach, combined with the flexible control for lagged test scores, failed to address mean reversion in the postreform period, then one would expect to see stronger effects in smaller schools, where the variance of average test scores is greater (Kane and Staiger 2002). As we will show in Table 5, if anything, larger schools exhibit stronger effects.

Table 2
Impacts of the Reform on School Exam Performance

	Marks		Pass Rate		Number Passed	
	(1)	(2)	(3)	(4)	(5)	(6)
0–10th percentile in previous year	4.406*** (1.004)	6.147*** (0.849)	0.058*** (0.012)	0.079*** (0.011)	1.828* (1.026)	2.180*** (0.754)
10–20th percentile in previous year	2.049*** (0.563)	2.382*** (0.578)	0.024*** (0.006)	0.030*** (0.007)	0.930 (0.845)	0.852** (0.379)
Diff-diff	Yes	Yes	Yes	Yes	Yes	Yes
School fixed effects	No	Yes	No	Yes	No	Yes
Control lagged exam score	Yes	Yes	Yes	Yes	Yes	Yes
Control mean, post BRN	109.46	109.46	0.58	0.58	30.82	30.82
Observations	77,731	77,431	77,731	77,431	77,731	77,431
R^2	0.655	0.801	0.607	0.763	0.425	0.912

Notes: Each column represents a separate regression. All specifications include district-by-year fixed effects, flexible controls for lagged test scores, and indicators for prereform associations between district-rank deciles and subsequent outcomes. Reported coefficients correspond to the differential effect of being ranked in the associated decile of within-district performance in the post- (vs. pre-) reform period, compared to the middle six deciles. In even columns, the specification is augmented with school fixed effects. In Columns 1 and 2 the outcome is the average PSLE score (ranging from 0–250), in Columns 3 and 4 it is the pass rate, and in Columns 5 and 6 it is the number of pupils who passed. 300 singleton schools are dropped when results are estimated using school fixed effects. Standard errors are clustered at the district level.

In Column 1 of Table 2, we present regression estimates of our primary difference-in-differences model from Equation 1, for the impacts of post-BRN district rankings on subsequent average exam scores. The first two rows present our main coefficients estimates, β_q, which compare differential post-BRN outcomes for schools in the bottom two deciles to schools in the middle six deciles of their district. These results indicate that being in the bottom decile in the post-BRN period is associated with a rise of more than four points on the average PSLE mark in the subsequent year, relative to schools in the middle six deciles, *over and above* any relationship that existed in the prereform period. This corresponds to an impact of just over 0.25 standard deviations in the distribution of school means for bottom-decile schools—a substantial effect size for what is essentially an informational intervention.[23,24] There is a smaller, but still statistically significant impact for schools in the second decile of school performance.[25]

23. Pre-BRN summary statistics for each performance decile are available in Online Appendix Table B.2.
24. See, for example, McEwan (2013) for a review. All families of interventions studied there have average effect sizes smaller than 0.2 standard deviations on student-level test scores.
25. Figure 1 shows that there are no detectable differences among schools that fall in the middle six deciles, in a specification that allows different effects for each decile. This affirms our decision to compare the extreme deciles with schools in the middle 60 percent, rather than only the middle 20 percent, for reasons of statistical power.

Columns 3–6 of Table 2 expand these results to two additional metrics of exam performance: the pass rate and the number passed. We first discuss the results in the odd-numbered columns, which are estimated using our main specification, Equation 1. The impacts on pass rates, shown in Column 3, are substantial. The estimated impact of 5.77 percentage points for schools in the bottom decile implies a 38 percent increase relative to the pre-BRN pass rate of 15 percent among bottom-decile schools.[26] In Column 5, we find that the reform increased the number of students that passed the PSLE by 1.8 students—a 21 percent increase relative to the pre-BRN level. In Online Appendix Table B.3, we show that these BRN-induced increases in exam scores also translate into reductions in the probability that schools in the bottom decile or quintile remain in that group in the subsequent year.

As a robustness check, the even-numbered columns show results on each measure of exam performance, estimated using the school fixed-effects model of Equation 2. It is encouraging that each of these results remain qualitatively unchanged. The coefficients for all three outcomes are, in fact, slightly larger, and the results on number passed are also now more precisely estimated.[27] This provides us more confidence to use this specification in subsequent analyses where data unavailability prevents us from constructing the lags of dependent variables.

B. Pressure on Low-Performing Schools Decreases Test-Takers and Enrollment

Although our estimates show that school-ranking initiative lead to increases in learning for (some) students in bottom-ranked schools, schools could have artificially boosted their average exam scores by strategically reducing the number of students who sit the exam. Here, we test for and estimate this mechanism directly.

In Column 1 of Table 3, we estimate BRN ranking impacts on the number of exam-sitters, using Equation 2. Schools that fell in the bottom decile in their district have more than two fewer test-takers the following year, compared to schools that fell in the middle six deciles. This equates to roughly a 4 percent reduction in the number of test-takers, since the average number of test-takers in the pre-BRN period in bottom quintile schools is 47.[28] Likewise, schools in the second-to-bottom decile also appear to reduce the number of test-takers by nearly two students.[29]

26. The pre-BRN summary statistics are available in Online Appendix Table B.2.

27. This is because lagged test scores are not as highly correlated with the number passed, compared to average score. Note that the *R*-squared in Column 5 is lower than in Columns 1 and 3. The introduction of fixed effects allows us to mop up additional variation.

28. The pre-BRN summary statistics are available in Online Appendix Table B.2. Online Appendix Table B.7 shows that these impacts are borne equally by male and female students.

29. Not shown in the table, we also find marginally significant impacts on schools at the top of the distribution, though in the opposite direction—for schools in the top decile of their district distribution, there is a marginally statistically significant increase of more than 1.6 students in the following year. There is no statistically significant change in test-takers in the second highest decile. Given that this is not mirrored in enrollment figures, and that it is the only statistically significant finding on enrollment outcomes for the top two deciles, out of ten coefficients estimated, we do not place much emphasis on interpretation of this finding. As will be shown in our analysis of enrollment sorting within schooling markets below, we also see no evidence of outflows from poorly performing schools to better nearby schools.

Table 3
Number of Test-Takers and Enrollment

	PSLE Data—Exam-Sitters		EMIS Data—Enrollment			
	All years (1)	2015 and 2016 (2)	Grades 4–6 (3)	Grade 6 (4)	Grade 7 (5)	Grade 7/ Grade 6 (6)
0–10th percentile in previous year	−2.039***	−1.676***	−0.773	−0.329	−1.646**	−0.028**
	(0.763)	(0.589)	(1.787)	(0.882)	(0.737)	(0.013)
10–20th percentile in previous year	−1.846***	−1.068**	1.929	0.865	−0.479	−0.008
	(0.635)	(0.423)	(1.334)	(0.659)	(0.480)	(0.010)
Diff-diff	Yes	No	No	No	No	No
Control lagged exam score	Yes	Yes	Yes	Yes	Yes	Yes
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Control mean	52.01	50.07	197.14	64.10	51.55	0.84
Observations	77,431	31,520	31,150	31,150	31,150	15,828
R^2	0.910	0.953	0.968	0.930	0.931	0.127

Notes: Each column represents a separate regression. All specifications include flexible controls for lagged test scores and school and district-by-year fixed effects. Column 1 is estimated on outcomes from 2012–2016, including indicators for prereform associations between district-rank deciles and subsequent outcomes. Reported coefficients in that column correspond to the differential effect of being ranked in the associated decile of within-district performance in the post- (vs. pre-) reform period, compared to the middle six deciles. In Columns 2–6, data are restricted to postreform years 2015–2016 in which EMIS data are available; this does not allow for a difference-in-difference specification. In Columns 1 and 2 the outcome variable is the number students sitting the exam, using PSLE data. In Columns 3–6 outcomes are different constructions of enrollment, based on EMIS data. The outcome in Columns 3–5 is the number of students enrolled in the corresponding grade(s). In Column 6 the outcome indicator is Grade 7 enrollment in 2016 divided by Grade 6 enrollment in 2015. Standard errors are clustered at the district level.

We next explore the fate of the students who were excluded from testing due to the BRN school rankings initiative. The answer is important to the welfare implications of the BRN ranking. We consider three possibilities: students could have repeated Grade 6 in the same school, switched to a different (potentially better) school, or dropped out of school altogether. The welfare loss of fewer exam-sitters will be greater if the reduction is driven primarily by students dropping out, rather than switching schools or repeating Grade 6.

To test whether BRN-induced pressure led to greater repetition rates, we apply Equation 2 with grade-specific enrollment numbers from EMIS data as the outcome (see Columns 3–6 in Table 3). Because EMIS data are only available for the years 2015 and 2016, we are not able to estimate this relationship in the pre-BRN period. However, we continue to take advantage of the fact that schools with the same test score can have different rankings in other districts, in order to identify a plausible causal effect of district rankings (see Reback, Rockoff, and Schwartz 2014 for more details on this

identification strategy). In this specification, we impose that the relationship between rankings and subsequent outcomes was flat in prior to BRN, implying coefficients $\alpha_q = 0$, for all quantiles q. To shed light on whether this restriction is likely to be consequential, we first reestimate impacts on the number of test-takers, restricting attention to 2015–2016 and therefore only using post-BRN data. As shown in Column 2 of Table 3, the estimated impacts on test-takers remains statistically significant, although slightly smaller in magnitude.[30] This adds to our confidence that the assumption that prereform ranking effects on EMIS outcomes are zero, or that violations of this assumption appear likely to attenuate observed results for enrollment-related outcomes.

In Column 5 of Table 3, we show that estimated impacts on Grade 7 enrollment are similar to those observed for PSLE exam-taking. In particular, the estimated loss of more than 1.6 students in Grade 7 mirrors almost exactly the effect size estimated on exam-takers for data from the same period.[31] However, Column 4 shows that there is no concurrent increase in Grade 6 enrollment in these same schools. We therefore find no evidence that these students are merely repeating Grade 6—they are either dropping out or switching schools.

Next, two tests allow us to rule out the interpretation that students are switching to other, potentially better-performing schools. First, we estimate the impact on enrollment in earlier grades. If students switched out of low-ranking and into high-ranking schools in their district because they updated their beliefs about school quality, then this effect need not be limited to Grade 7. Students in earlier grades could also be induced to switch as well. However, we have already seen in Column 4 of Table 3 that there is no evidence of an enrollment impact in Grade 6. To provide an additional—and potentially more powerful—test for such a phenomenon, we estimate impacts of within-district deciles on the pooled numbers of students in Grades 4–6. As shown in Column 3 of Table 3, there is no evidence that the enrollment effects observed in Grade 7 are also found in lower grades.[32]

Second, we test if school enrollment is also a function of neighboring schools' performance. If Grade 7 students from schools in the bottom decile of their district are switching to other schools, then there should be a *positive* effect on Grade 7 enrollment when a school is surrounded by other schools that perform poorly. To operationalize this, we use a school's ward designation—a geopolitical unit just below the district that contains on average four schools—to define a local school "market." Wards are good approximations for school markets because they are typically used to define catchment areas for secondary schools. We then augment the specification used in Table 3 by including ward means of the within-district decile indicators, to estimate an equation of the form:

30. Note that it is entirely possible that the *true* treatment effect of BRN is smaller in these later years, as the reform was losing momentum.

31. These results suggest that the test-taking results are not simply due to students being absent from school during the testing period. Rather, the students are not enrolling in Grade 7.

32. Given the salience of the PSLE, it is likely much harder to switch schools at Grade 7 (relative to earlier grades) as schools would be reluctant to admit students who may perform poorly on the PSLE. Thus, we would expect greater switching to occur in earlier grades. The fact that we do not find any evidence of school switching suggests that either there is insufficient school choice or that parents were not sufficiently informed.

$$(3)\quad x_{swdt} = f_t(y_{s,t-1}) + \sum_{q\in\{1,2,9,10\}} \alpha_q I_q(r_{sw,t-1}) + \sum_{q\in\{1,2,9,10\}} \beta_q I_q(r_{sw,t-1}) \cdot Post_t + \sum_{q\in\{1,2,9,10\}} \phi_q \bar{I}_{qw,t-1} + \sum_{q\in\{1,2,9,10\}} \pi_q \bar{I}_{qw,t-1} \cdot Post_t + \gamma_{dt} + \mu_s + e_{dst}$$

As before, x_{swdt} denotes enrollment in school s, which is in ward w and district d, in year t. The key difference in relation to Equation 2 is that we include ward-level means of the district rank indicators (or the share of schools in the ward within each district rank), defined as $\bar{I}_{qw,t-1} = \frac{1}{|w|} \sum_{s\in w} I_q(r_{sw,t-1})$, where $|w|$ is the number of schools in ward w. We include these directly and interacted with a post-BRN indicator. Coefficients π_q capture BRN-induced sorting in enrollment—that is, induced transfers to other schools within the ward.

If all BRN-induced exits from schools are in fact transfers to other schools in the same ward, then we would expect the coefficients on ward averages of the district rank indicators to be equal in magnitude and opposite in sign to the corresponding school-level effects, so that $\beta_q + \pi_q = 0$ for each district ranking decile, q.[33] For example, recall that in Table 3 we estimate that a school in the bottom decile of its district will be induced to shed two test-takers in the following year. If these were purely transfers, then we would expect the sorting coefficient (π_q) for the bottom decile to equal two. On the other hand, if there were no BRN-induced transfers, then the coefficients on ward means would be equal to zero.

Estimates of Equation 3, presented in Online Appendix Table B.4, suggest that enrollment sorting is not important here. We are able to reject the implication of the pure sorting model, that $\beta_q + \pi_q = 0$ for all q, with a p-value of 0.054. Point estimates for the share of ward means in the bottom decile—where we see induced dropouts—are *negative*, rather than positive, as would be predicted by an empirical model of pure sorting. Moreover, the sorting coefficients, π_q, are always statistically insignificant.

Taking this direct test of sorting together with the absence of enrollment effects on lower grades, we conclude that the estimated impact of receiving a low district ranking on the number of test-takers (and enrollment) is unlikely to be driven by strategic grade repetition or by students switching schools away from low-ranked schools. Instead, it appears that the pressure of receiving a low ranking in the district leads schools to respond by inducing some students who would otherwise enroll in this exam-taking year to drop out altogether.

As the number of excluded students is relatively small, we examine the extent to which this strategy could have driven the observed increase in schools' average test scores. To do so, we reestimate the analysis of the program's effect on test scores (Table 2, Column 1), bounding the consequences of the exclusion effect. Specifically, we can

33. Intuition for the coefficients on these ward-level means, $\bar{I}_{qw,t-1}$, can be understood by comparison with estimation of the analog of Equation 2 entirely at the ward level, as in Hsieh and Urquiola (2006). In such a ward-level regression, we would expect the coefficients on ward averages of the district rank indicators to be equal to zero if the school-level impacts of district rank on the number of takers had been purely driven by students switching schools. But for each decile q of district ranking, the coefficients on the corresponding ward-level averages, $\bar{I}_{qw,t-1}$, are equivalent to the sum of coefficients on school indicators, β_q, and their corresponding ward averages, π_q, in Equation 3. Consequently, a case in which BRN induces pure student switching across schools would be a case in which $\beta_q + \pi_q = 0$.

compute the adjusted school's average test score by adding back the excluded students (using the coefficients in Table 3) and making an assumption about what these students would have scored on the PSLE. The positive and significant test score effects hold unless the excluded students had PSLE average test scores below 27 (out of 250). This level of performance is approximately equivalent to the average performance of the worst school in the data.[34] Thus, our analysis shows that exclusion of just two students can meaningfully inflate a school's average test score—giving school administrators an incentive to pursue this strategy—especially if schools can correctly identify students who are likely to fail.

This strategy of exclusion only affects the school's average test score and pass rate. It will not directly affect the absolute number of students who pass the PSLE.[35] Thus, our results on the absolute number of students passing the PSLE are not an artifact of such gaming and are instead a reflection of real improvements in learning. Taken together, our results on the number passed and the number of test-takers suggest that low-ranked schools responded to the BRN district ranking initiative by both excluding students and also exerting effort to raise the performance of the remaining students.

C. No Evidence of Impacts on Monitoring, Resources, Teacher Effort, or Learning in Other Grades

We next turn to examining potential mechanisms underlying the estimated improvements in exam performance. Schools at the bottom deciles could have improved performance if they received more resources from government or the community, used existing resources more efficiently, increased overall levels of teaching effort, or reallocated existing resources and efforts towards preparing Grade 7 students for the exams. In this section, we use the World Bank Service Delivery Indicators (SDI) panel data set to test for some of these mechanisms. Specifically, these data allow us to test for impacts on numbers of teachers, stocks of textbooks, school finances, district inspections, and teacher presence. The SDI data from a sample of Grade 4 pupils further allow us to test for learning impacts on earlier grades.

As a starting point for this analysis, we show that the main results hold with the reduced sample of schools and years in which the SDI data collection took place. Online Appendix Table B.5 replicates the main results for the reduced sample and the years corresponding to the SDI outcomes: 2013–2016.[36] The estimated impacts for this sample are in fact much higher and remain statistically significant.

Columns 1–3 in Table 4 show that bottom-decile schools did *not* receive any more resources from government.[37] There is no statistical significant difference in the number of teachers, the number of textbooks (per student), or the per-student value of capitation

34. In this school all students failed and received the worst possible letter grade ("E") on the PSLE.

35. There could be indirect effects of strategic exclusion on the number passed. Potential pathways include reductions in class size and reductions in within-classroom heterogeneity.

36. Data collection took place in 2014 and 2016. However, some questions—such as receipt of resources and school inspections—were asked using the previous year as the reference period.

37. Recall that this model is estimated with school fixed effects, and so is identified off switches in ranking status across years.

Table 4
District Ranking Impacts on Monitoring, Teacher Effort, Resource Spending, and Allocation

	Teachers (1)	Textbooks (2)	Capitation Grants (3)	Inspections (4)	Teacher Presence (5)
0–10th percentile in previous year	0.921 (0.665)	0.110 (0.174)	0.202 (0.976)	–0.298 (0.544)	–0.036 (0.082)
10–20th percentile in previous year	–0.202 (0.500)	–0.139 (0.084)	–1.052 (–0.669)	0.782 (0.552)	–0.015 (0.057)
Post-BRN mean: 20–80th percentile	17.88	0.38	3857.17	1.49	0.54
Observations	760	754	756	758	760

Notes: Each column represents a separate regression, estimated using SDI data, with flexible controls for lagged test scores and district-by-year and school fixed effects. Coefficients correspond to the effect of being ranked in the associated decile of within-district performance in the postreform period, compared to the middle six deciles. The SDI data were collected in 2014 and 2016 (the postreform period), but some variables were collected using the previous year (2013 or 2015) as the reference period. For each column, only two years of data are available: Columns 1, 4, and 5 use outcomes for the years of 2014 and 2016, and Columns 2 and 3 use outcomes for the years 2013 and 2015. The dependent variables in Columns 2 and 3 are inverse hyperbolic sine transformations (an approximation for the natural logarithm) and calculated at a per-student level, using enrollment data from 2014. Data from Column 3 are reported in Tanzanian shillings. The mean values reported in the penultimate row is of the untransformed outcome. We adjust for outliers in the following way: (i) we adjust downwards the per-student capitation grant to the maximum that a school can receive, 10,000 Tanzanian shillings; (ii) we set as missing one school that reported receiving 600 textbooks per student. Since the specifications include school fixed effects, schools with only one observation are dropped. Standard errors are clustered at the district level.

grants received over the year.[38] Thus, we do not find evidence that schools faced punitive consequences for poor performance.

Column 4 shows that schools were no more likely to receive a school inspection if they were in the bottom decile in the preceding year. We therefore have no evidence that government provided additional supervisory or support visits to schools. This does not conclusively rule out top-down pressure from the district officials, though, since inspections are performed by the Quality Assurance Department of the Ministry of Education, and this variable does not capture visits by the District or Ward Education Officers. In addition, these data do not capture other potential methods education officials could have used pressure schools, such as the stakeholder meetings documented in the qualitative reports (Integrity Research 2016).

Column 5 shows that there is no impact on teacher presence. Thus, it is unlikely that increases in school-presence among teachers could have driven the observed learning gains. However, this leaves open the question about the means used by schools to

38. We take an inverse hyperbolic sine transformation of the textbooks and grants.

improve learning outcomes. It is possible that schools responded to the pressure by offering remedial courses or spending more time preparing Grade 7 students for the PSLE. Unfortunately, such forms of effort were not measured in the SDI data.[39] It is also possible that the strategic removal of some students could actually *cause* learning outcomes to improve for those who remain, especially if those removed from class are particularly disruptive.[40]

The SDI data also allow us to look at the impacts of the reform on student learning in other (nonincentivized) grades. The impact of the reform on learning in other grades could go in two directions. On the one hand, it is possible that increased effort levels could lead to positive spillovers on learning in earlier grades. On the other hand, learning in earlier grades could have suffered if existing school resources were redirected to students in Grade 7.

The SDI data collection included an assessment of a random sample of 20 Grade 4 students in three different subject areas: mathematics, English, and Kiswahili. All the measures are standardized to have a mean of zero and standard deviation of one. Online Appendix Table B.6 reports results on these outcomes. Although results are somewhat imprecise, there is no detectable positive or negative impact on learning. The gains in learning are therefore likely restricted to Grade 7 students.

VI. Robustness

The publication of within-district school rankings was only one part of a suite of reforms undertaken under the heading of BRN (see Table B.1 in the Online Appendix for more details). Many of these reforms were unlikely to impact Grade 7 outcomes during the study period (for example, the early grade curriculum reforms), and the implementation of most of BRN components were delayed due to the lack of funding. For example, the capitation grant reform was only launched in 2016—the last period of our study. The school ranking program was the first component launched and one of the few that was consistently implemented throughout our study period. Other initiatives, such as the Student Teacher Enrichment Programme (STEP), were implemented starting in 2014 and may drive our results. To assuage these concerns, we conduct several robustness checks.

We first examine whether our results are driven by pressure generated from national rankings or from district rankings. As Online Appendix Table B.1 shows, the BRN reforms included a national ranking system, and these could have been more salient than the district rankings. However, schools' national rankings are a one-to-one, monotonic increasing function of the school-level average test scores, which are already flexibly included in our specifications. Thus, our empirical specifications arguably already

39. In general, it is very difficult to capture teacher effort accurately. However, a number of experimental studies on teacher incentives found increases in learning outcomes without any corresponding increase in teacher presence (for example, see Muralidharan and Sundararaman 2011; Mbiti et al. 2019). This suggests that teachers can increase effort in ways that are difficult to capture using our conventional methods.

40. The reductions in class size are arguably too small to reduce within-class heterogeneity significantly. Thus, given the small reduction in the number of students, disruption effects are a plausible mechanism.

control for a sufficient statistic of the national ranking. However, to further assuage concerns about the potential role played by the national component of the school ranking, we conduct additional empirical checks in Table 5. First, we split our sample by the average district performance in Columns 1 and 2 of Table 5. Bottom-ranked schools in below-average districts would be the worst schools nationally, while bottom-ranked schools in above-average districts would not necessarily fall in the bottom of the national rank distribution.[41] Conversely, top schools in the above-average districts would be among the best schools nationally, while top schools in the below-average districts would not necessarily be among the best schools. To the extent that national rankings play a role in driving district-ranking results, then bottom-ranked schools in the better districts face less pressure than bottom-ranked schools in the below-average districts. Thus, we would not find any effects in Column 2. However, our results show that bottom-ranked schools in both types of districts saw subsequent increases in performance, suggesting that the district rankings were the primary driver.

In the original project design, the best-performing schools would receive nonmonetary rewards, such as certificates, public ceremonies, and media coverage. For schools that saw the greatest improvement, the government planned to grant three to five million Tanzanian shillings ($\backsim$US$1,800–3,000) to the 300 most improved primary schools and one to two million Tanzanian shillings ($\backsim$US$600–1200) to 2,700 other primary schools that were most improved (Government of Tanzania 2014). However, despite these pledges, the government significantly scaled back the program such that by October 2016 only 120 primary schools had ever received incentive grants (World Bank 2018a). Given that less than 1 percent of schools actually won such prizes, it is unlikely that our results are driven by the school incentive grants. Moreover, qualitative reports suggest that the incentive program was not well understood (Integrity Research 2016).

Since the best-performing schools nationally tend to be in above-average districts, the school incentive grant would likely induce top schools from the better districts to respond. However, we did not find any statistically significant effects among the top schools in the best districts (coefficients not shown). In addition, schools that experienced large negative shocks in the previous year would be better placed to earn a reward, as they could leverage mean reversion to boost their test score improvement metric. As discussed in more detail below, we show that our results are robust if we exclude schools that experienced large declines in test scores, suggesting that the results are not driven by the school incentives program.

The BRN program also included the Student-Teacher Enrichment Program (STEP), which was a remedial education program that was aimed at improving test scores in the PSLE. The program trained teachers across districts to identity and support low-performing students who were preparing for the PSLE exam. The STEP program also trained teachers to conduct diagnostic tests to identity students who were at risk of failing. These students would receive extra remedial instruction and exam coaching sessions to prepare them for the exams (Government of Tanzania 2014). Given the limited capacity of the government to roll out this program across the entire country, the implementation was targeted toward districts that had large numbers of failing students

41. The average performance of schools in the bottom decile within below-average districts is roughly equivalent to the 25th percentile of all schools, while the average school in the bottom decile within above-average districts is roughly at the median of all schools.

Table 5
Robustness Checks

	District Performance		STEP Program		Negative Shock		Small School	
	Below (1)	Above (2)	STEP (3)	Other (4)	Shock (5)	None (6)	Smallest (7)	Large (8)
0–10th percentile in previous year	4.155***	5.572***	5.871***	4.313***	−18.558	2.940***	3.817*	4.329***
	(1.207)	(1.636)	(1.509)	(1.305)	(13.720)	(0.482)	(2.263)	(0.953)
10–20th percentile in previous year	1.833***	2.770***	2.764***	1.942**	−3.379	0.917***	2.068	1.722***
	(0.679)	(0.856)	(0.702)	(0.763)	(33.090)	(0.344)	(1.453)	(0.551)
Diff-diff	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Fixed effects	No	No	No	No	Yes	Yes	No	No
Observations	41,096	36,635	25,395	52,336	932	56,463	15,108	62,623
R^2	0.580	0.642	0.669	0.650	0.864	0.815	0.656	0.664

Notes: Each coefficient refers to the decile of within-district performance rank, compared to the middle six deciles. The outcome variable is average school performance, which can take values of 0–250). In Column 1 the sample is restricted to the bottom half of districts, in terms of a district's average school performance on the previous year exam; in Column 2 the sample is restricted to the top half. In Column 3 the sample is restricted to districts where the STEP remedial education training took place; in Column 4 it is restricted to districts where it did not take place. In Column 5, the sample is restricted to schools that dropped 30 percentiles in its national rank between year t and $t-1$. Schools in Column 5 did not experience such test score declines. In Columns 5 and 6 we do not difference out the baseline relationship between rank and performance, since we do not have data for performance in 2010 so do not know which schools in 2011 experienced a large drop. In Column 7, the sample is restricted to smallest quintile of schools—measured in the number of test-takers in 2012. Column 8 is the complement of Column 7. Standard errors are clustered at the district level.

in previous years, as well as districts that had experienced a large drop in performance in the pre-BRN period (Government of Tanzania 2014, 2015). Using the implementation plan outlined in official reports (see, for example, Government of Tanzania 2014), we identify the districts that were targeted for the STEP program, and compare the results for the STEP districts to the non-STEP districts in Columns 3 and 4. Overall, we find similar results in both sets of districts, suggesting that the STEP program is not biasing our results.

An additional concern is that reversion to the mean may drive our results despite our ability to include flexible controls for past performance and district fixed effects. To address such concerns, we split our sample into different categories based on their potential to experience mean reversion. We first compare schools that experienced a large reduction in exam performance the previous year to those that did not. Specifically, we compare schools that saw a reduction of at least 30 percentile points nationally to those that did not.[42] Since schools that experienced a negative shock are more likely to "bounce back," this comparison provides an additional test of the potential for mean reversion to bias our findings. We find no evidence that the treatment impacts are concentrated in schools that experienced a shock in Column 6 of Table 5. The treatment impact remains large after excluding these schools, and there is in fact no detectable impact of the reform on the subset of schools who experienced such a shock, although the sample size in Column 5 is small. We repeat this exercise using different percentile decline thresholds and find similar patterns. For instance, among the group of almost 10,000 schools that experienced a ten percentile drop, we did not find any statistically significant effects of being in the bottom rank on subsequent performance (see Online Appendix Table B.8).

To the extent that smaller schools both are more likely to experience mean reversion and face stronger incentives on a "gains" metric of school performance (Kane and Staiger 2002), we also split the sample by school size in Column 7 and 8 of Table 5. We compare the smallest fifth of schools (Column 7), as measured by the number of test-takers in the prereform period (2012) to their larger counterparts (Column 8). We find similar results in both sets of schools, suggesting that mean reversion is not driving our results. In Online Appendix Table B.9, we repeat this exercise for test score outcomes using only postreform data—that is, in a model that does not difference out prereform decile effects. To the extent that the ability to difference out prereform decile effects is important to addressing mean reversion, this might drive estimates using EMIS and SDI data from the postreform period. However, we instead find that larger schools exhibit, if anything, stronger responses, even in this model that uses only postreform data. These findings strengthen our confidence that mean reversion does not drive our results.

As smaller schools and schools that experienced negative shocks are also more likely to see larger performance gains in subsequent exam performance, and face stronger incentives to bring these gains about, these robustness checks also serve as additional checks about the potential for our results to be driven by the school incentive program. Since we generally find that our results are robust when we exclude small schools and

42. We find qualitatively similar patterns using a variety of thresholds including a ten percentile point reduction.

schools that experienced shocks, we can be more confident that our results are primarily reflect the district ranking component of BRN.

VII. Discussion

Tanzania's Big Results Now in Education program has been touted as a "promising society-wide collaborative [approach] to systematically improving learning" (World Bank 2018b). With the full backing of the office of the president, the program was highly visible both nationally and internationally and attracted US$257 million in donor funds.

This study presents evidence that such low-stakes accountability programs, which do not provide any direct financial incentives can lead to improvements in performance, even in the absence of a parental response. In Tanzania, there was an overall improvement in the exam performance for schools in the bottom deciles of their district, who faced additional pressure to improve. There was also a net increase in the total number passed. It is unlikely that parental responses to information provided these incentives, since parents were on the whole unaware of their school's district rank. The mechanism is most likely a combination of pressure exerted by District Education Officers, who themselves had incentives to demonstrate in their district, and a mix of professional norms and competitive desires among head teachers, seeking to avoid poor performance in an environment in which results had become more salient.

However, this study also tells a cautionary tale of the negative unintended consequences of policies. Our results show that the BRN reform had mixed effects on student outcomes. On one hand, the reform improved learning among students in bottom-ranked schools, resulting in almost two additional students passing the PSLE. On the other hand, the reform pushed out roughly two students from Grade 7 in bottom-decile schools, and evidence suggests that these students dropped out rather than repeating or switching schools. Thus, the overall welfare effect of the program is unclear and will depend on the structure of policymaker preferences.

Arguably, the value of educational gains experienced by the 1.8 students per school who were induced to pass their PSLE exam are substantial, as nearly all students who pass the PSLE progress to secondary school.[43] Moreover, on average, students who enter secondary school (Form 1) have a 73 percent chance of completing lower secondary school and receive an additional 3.8 years of schooling.[44] Of course, marginal PSLE passers are likely among the least prepared students for secondary schools, so these typical attainment levels likely represent an upper bound on those achieved by students induced to pass the PSLE by the BRN ranking.

These benefits have to be compared to the *reduction* in acquired human capital for the roughly two students per school who were excluded from the PSLE as a result of the reform. Since our analysis suggests that they dropped out of school altogether rather

43. The number of students enrolled in the first year of secondary school typically *exceeds* the number of students passing the PSLE in the prior year.

44. The transition rates between the years 2016 and 2017 for Grade 9 (Form 2) to Grade 13 (Form 6) were 97 percent, 83 percent, 91 percent, 22 percent, and 95 percent, respectively.

than repeating a grade or transferring to another school, we conclude that these students lose the acquired human capital of Grade 7 altogether.[45]

Weighing these positive and negative effects, we conclude that it is likely that BRN's public ranking of schools resulted in a net increase in total grade attainment in schools ranking in the bottom of their districts, but that this came at the expense of losses in the human capital of low-performing students. An inequality-averse policymaker may reject this tradeoff in spite of the positive effect on average years of schooling.

Nonetheless, this reform highlights that even reputational incentives—if they are sufficiently powerful to induce a behavioral response—can induce strategic responses that are inconsistent with policymakers' intent. School stakeholders respond to incentives on the margins they judge most effective. The consequences of those behavioral responses can be a double-edged sword.

References

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2017. "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets." *American Economic Review* 107(6):1535–63.

Baker, George. 1992. "Incentive Contracts and Performance Management." *Journal of Political Economy* 100(3):596–614.

Baker, George. 2002. "Distortion and Risk in Optimal Incentive Contracts." *Journal of Human Resources* 37(4):727–51.

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2010. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India." *American Economic Journal: Economic Policy* 2(1):1–30.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2018. "Experimental Evidence on Scaling up Education Reforms in Kenya." *Journal of Public Economics* 168(December):1–20.

Bruns, Barbara, Deon Filmer, and Harry Patrinos. 2011. *Making Schools Work: New Evidence on Accountability Reforms*. Washington, DC: World Bank.

Burgess, Simon, Carol Propper, Helen Slater, and Deborah Wilson. 2005. "Who Wins and Who Loses from School Accountability? The Distribution of Educational Gain in English Secondary Schools." CEPR Discussion Paper 5248. Washington, DC: CEPR.

Camargo, Braz, Rafael Camelo, Sergio Firpo, and Vladimir Ponczek. 2018. "Information, Market Incentives, and Student Performance: Evidence from a Regression Discontinuity Design in Brazil." *Journal of Human Resources* 53(2):414–44.

Chay, Kenneth, Patrick McEwan, and Miguel Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." *American Economic Review* 95(4):1237–58.

Chiang, Hanley. 2009. "How Accountability Pressure on Failing Schools Affects Student Achievement." *Journal of Public Economics* 93(9–10):1045–57.

Cilliers, Jacobus, Isaac Mbiti, and Andrew Zeitlin. 2020. "Replication Data for: 'Can Public School Rankings Improve School Performance? Evidence from a Nation-Wide Reform in Tanzania' Journal of Human Resources (2020)." Harvard Dataverse, V1, UNF:6:0ZSD7w 3CCeTcLRuBzvKPnQ==[fileUNF]. https://doi.org/10.7910/DVN/ABVMCL

45. Because estimated impacts on the absolute number of PSLE passers are net of these dropouts, any continuation value of subsequent years of schooling by those induced to drop out is already accounted for above.

Craig, Steven, Scott Imberman, and Adam Perdue. 2013. "Does It Pay to Get an A? School Resource Allocations in Response to Accountability Ratings." *Journal of Urban Economics* 73(1):30–42.
Craig, Steven G., Scott Imberman, and Adam Perdue. 2015. "Do Administrators Respond to Their Accountability Ratings? The Response of School Budgets to Accountability Grades." *Economics of Education Review* 49(December):55–68. https://doi.org/10.1016/j.econedurev.2015.07.005
Dee, Thomas, and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy* 34(2):267–97.
Figlio, David. 2006. "Testing, Crime, and Punishment." *Journal of Public Economics* 90(4–5): 837–51.
Figlio, David, and Lawrence S. Getzler. 2002. "Accountability, Ability, and Disability: Gaming the System." NBER Working Paper 9307. Cambridge, MA: NBER.
Figlio, David, and Susanna Loeb. 2011. "School Accountability." In *Handbook of the Economics of Education*, ed. Eric Hanushek, Stephen Machin, and Ludger Woessmann, 383–421. Amsterdam: North Holland.
Figlio, David, and Cecilia E. Rouse. 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics* 90(1–2):239–55.
Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2(3):205–27.
Government of Tanzania. 2014. "Big Results Now! Education Lab Storyline." Presidential Delivery Bureau, United Republic of Tanzania.
———. 2015. "Big Results Now! Annual Report 2013/14." Presidential Delivery Bureau, United Republic of Tanzania.
Hanushek, Eric, and Margaret Raymond. 2005. "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management: Journal of the Association for Public Policy Analysis and Management* 24(2):297–327.
Hastings, Justine, and Jeffrey Weinstein. 2008. "Information, School Choice, and Academic Achievement: Evidence from Two Experiments." *Quarterly Journal of Economics* 123(4): 1373–414.
Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7:24–52.
Hsieh, Chang-Tai, and Miguel Urquiola. 2006. "The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program." *Journal of Public Economics* 90:1477–503.
Integrity Research. 2016. "P4R Analysis of the Tanzania School Ranking Initiative." London: Integrity.
Jacob, Brian. 2005. "Accountability, Incentives, and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89(5–6):761–96.
Jones, Sam, Youdi Schipper, Sara Ruto, and Rakesh Rajani. 2014. "Can Your Child Read and Count? Measuring Learning Outcomes in East Africa." *Journal of African Economies* 23(5): 643–72.
Joshi, Arun, and Isis Gaddis. 2015. Preparing the Next Generation in Tanzania: Challenges and Opportunities in Education. Washington, DC: World Bank.
Kane, Thomas, and Douglas Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives* 16(4):91–114.
Koning, Pierre, and Karen Van der Wiel. 2012. "School Responsiveness to Quality Rankings: An Empirical Analysis of Secondary Education in the Netherlands." *De Economist* 160(4): 339–55. https://doi.org/10.1007/s10645-012-9194-9

Koning, Pierre, and Karen Van der Wiel. 2013. "Ranking the Schools: How School-Quality Information Affects School Choice in the Netherlands." *Journal of the European Economic Association* 11(2):466–93.
Lieberman, Evan, Daniel Posner, and Lily Tsai. 2014. "Does Information Lead to More Active Citizenship? Evidence from an Education Intervention in Rural Kenya." *World Development* 60(Supplement C):69–83. https://doi.org/10.1016/j.worlddev.2014.03.014
Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *Quarterly Journal of Economics* 134(3):1627–73.
McEwan, Patrick. 2013. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." Working paper. Wellesley, MA: Wellesley College.
Mizala, Alejandra, and Miguel Urquiola. 2013. "School Markets: The Impact of Information Approximating Schools' Effectiveness." *Journal of Development Economics* 103:313–35.
Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31(4):103–24.
Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119(1):39–77.
Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92(2):63–83.
Nunes, Luis, Ana Balcão Reis, and Carmo Seabra. 2015. "The Publication of School Rankings: A Step toward Increased Accountability?" *Economics of Education Review* 49(December): 15–23.
Reback, Randall, Jonah Rockoff, and Heather L Schwartz. 2014. "Under Pressure: Job Security, Resource Allocation, and Productivity in School Under No Child Left Behind." *American Economic Journal: Economic Policy* 6(3):207–41.
RISE Tanzania Country Research Team. 2017. "Monitoring the Big Results Now in Education Program." Initiative on Innovation, Development, and Evaluation, Technical Report. Washington, DC: Georgetown University. https://tinyurl.com/y58am75x (accessed November 25, 2020).
Rockoff, Jonah, and Lesley Turner. 2010. "Short-Run Impacts of Accountability on School Quality." *American Economic Journal: Economic Policy* 2(4):119–47.
Rouse, Cecilia, Jane Hannaway, Dan Goldhaber, and David Figlio. 2013. "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." *American Economic Journal: Economic Policy* 5(2):251–81.
Todd, Robin, and Ian Attfield. 2017. "Big Results Now! In Tanzanian Education: Has the Delivery Approach Delivered?" Unpublished.
Twaweza. 2010. "Capitation Grant for Education: When Will It Make a Difference?" Uwazi Policy Brief 8. Dar es Salaam, Tanzania: Twaweza.
———. 2012. "Are Our Children Learning? Literacy and Numeracy in Tanzania." Uwezo National Report. Dar Es Salaam, Tanzania: Uwezo.
———. 2013. "Capitation Grants in Primary Education: A Decade since Their Launch, Does Money Reach Schools?" Sauti za Wananchi brief 3. Dar es Salaam, Tanzania: Twaweza.
Valente, Christine. 2015. "Primary Education Expansion and Quality of Schooling: Evidence from Tanzania." IZA Discussion Paper. Bonn, Germany: IZA.
World Bank. 2015. "TZ Big Results Now in Education Program (P147486)." Washington, DC: World Bank. http://documents1.worldbank.org/curated/en/225831468131397200/pdf/P4R-ISR-Disclose-P147486-04-04-2015-1428166774259.pdf (accessed November 19, 2020).
———. 2016a. "Education Service Delivery in Tanzania." Education Technical Report, Tanzania 2014 Service Delivery Indicators, International Bank for Reconstruction and Development. Washington, DC: World Bank.

———. 2016b. "Tanzania Education Program for Results." Implementation Status and Results Report P147486 5. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/267111481912693598/pdf/1481912691830-0000A8056-ISR-Disclosable-P147486-12-16-2016-1481912679228.pdf (accessed November 25, 2020).
———. 2016c. "Tanzania Education Program for Results." Implementation Status and Results Report P147486 4. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/324721468116335677/pdf/ISR-Disclosable-P147486-04-26-2016-1461726439658.pdf (accessed November 25, 2020).
———. 2018a. "Tanzania Education Program for Results." Implementation Status and Results Report P147486 7. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/752351517941488259/pdf/ISR-Disclosable-P147486-02-06-2018-1517941475608.pdf (accessed November 25, 2020).
———. 2018b. *World Development Report 2018: Learning to Realize Education's Promise*. Washington, DC: International Bank for Reconstruction and Development, World Bank.