RISE
RESEARCH ON IMPROVING SYSTEMS OF EDUCATION

RISE Working Paper 20/046
September 2020

From Cheating to Learning: An Evaluation of Fraud Prevention on National Exams in Indonesia

Emilie Berkhout, Menno Pradhan, Rahmawati, Daniel Suryadarma, and Arya Swarnata

Abstract

Cheating reduces the signal value of exam data and it might shift the focus of teachers and students away from learning. However, it is difficult to prevent cheating if it is widespread. We evaluate the impact of computer-based testing (CBT) on national exam scores in junior secondary schools in Indonesia, exploiting the phased roll-out of the program from 2015 to 2019. First, we find that test scores decline dramatically after the introduction of CBT with school-level means declining by 0.4 standard deviation. Schools with response patterns that indicate cheating experience an increased drop in their test scores. Second, scores rebound within two years after introducing CBT, suggesting that barriers to cheating provide incentives for learning. Third, we find evidence of spillover effects from CBT within districts. Cheating declines more in schools that have not yet switched to CBT if more schools located in the same districts make the switch, suggesting that CBT not only eliminates cheating but makes it less socially permissible.

JEL Classifications: C23 H52 I21

UKaid from the British people
Australian Aid
BILL & MELINDA GATES foundation
Oxford Policy Management
BLAVATNIK SCHOOL OF GOVERNMENT
UNIVERSITY OF OXFORD

From Cheating to Learning: An Evaluation of Fraud Prevention on National Exams in Indonesia

Emilie Berkhout
University of Amsterdam, Amsterdam Institute for Global Health and Development

Menno Pradhan
University of Amsterdam, Amsterdam Institute for Global Health and Development; Vrije Universiteit Amsterdam, Tinbergen Institute

Rahmawati
Center for Assessment and Learning, Indonesian Ministry of Education and Culture

Daniel Suryadarma
SMERU Research Institute

Arya Swarnata
SMERU Research Institute

Acknowledgements:

This project is funded by the United Kingdom's Foreign, Commonwealth and Development Office (FCDO), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation under the RISE Programme. We thank the Indonesian Ministry of Education and Culture for sharing their data with us. We would like to acknowledge Hessel Oosterbeek, Amanda Beatty, our colleagues at the University of Amsterdam and the Free University Amsterdam, and the audience at the RISE Annual Conference 2019 and the IRSA Conference 2019 for their valuable feedback. Corresponding Author: e.berkhout@aighd.org.

This is one of a series of working papers from "RISE"—the large-scale education systems research programme supported by funding from the United Kingdom's Foreign, Commonwealth and Development Office (FCDO), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation. The Programme is managed and implemented through a partnership between Oxford Policy Management and the Blavatnik School of Government at the University of Oxford.

Please cite this paper as:
Berkhout, E. et al. 2020. From Cheating to Learning: An Evaluation of Fraud Prevention on National Exams in Indonesia. RISE Working Paper Series. 20/046. https://doi.org/10.35489/BSG-RISE-WP_2020/046

1 Introduction

Cheating on high stakes exams is a concern all over the world, from the 'cheating mafia' in India (Safi, 2018) to fraudulent, prestigious high schools in the United States (Anderman, 2015). Cheating is costly to society. Exam results are used by schools, employers and policy makers, yet cheating makes these results misleading as a signal for individual learning achievement. Moreover, when students and teachers know that they can cheat on high stakes exams, they might exert less effort on studying and teaching.

When cheating is widespread it becomes harder to fight. Cheating leads to higher grades, which is what students, teachers and bureaucrats are held accountable for. Therefore, these stakeholders have a reason to allow for cheating and to keep it a secret. Non-cheaters are the ones who lose out from cheating as their results reflect poorly relative to the cheating students and teachers, but when there are few non-cheaters, and cheating cannot be proven, it will be hard for them to stop the cheating practice. A common solution for a central planner, in this case the Ministry of Education, is to send external supervisors to exams (Bertoni et al., 2013), but this does not work if the supervisors are also corrupt (Borcan et al., 2017). This way, widespread cheating can perpetuate for a long time.

Cheating at national exams in Indonesia has been a persistent problem. The practice has been widely reported in the popular press (e.g. Economist, 2011; Sundaryani, 2015; Jong, 2015), but there have been hardly any cases of where it was prosecuted. Reported cheating ranged from students copying each other's answers to teachers and principals providing answer keys to students prior to or on the exam day.

In 2015, the extent of cheating became apparent when the central government started disseminating an "integrity index" at the school level (Rahmawati and Asrijanty, 2016). The integrity index identifies cheating through suspicious answer patterns, a method that has been validated in schools in Chicago (Jacob and Levitt, 2003) and has been used in multiple studies to measure cheating on exams in Italy and Mexico (Battistin et al., 2017; Martinelli et al., 2018). Using this method, classrooms with identical answer strings or counter-intuitive performance

on items of certain difficulty levels, such as scoring high on difficult items while the easier items are incorrect, are given a low integrity index. The index was shared with district governments, who are responsible for ensuring a fair examination in Indonesia's decentralized setting. The results revealed widespread cheating. 33 percent of the schools were flagged by the Ministry of Education as suspicious, compared to 5 percent of the classrooms in Chicago (Jacob and Levitt, 2003) and 5 percent in Italy (Angrist et al., 2017). In Mexico 7 percent of high school exams were flagged as suspicious (Martinelli et al., 2018), and the latter increased to 32 percent after two years of monetary incentives based on test scores for students and teachers.

In this paper, we evaluate the impact of computer-based testing (CBT) for the national examination of junior secondary schools in Indonesia. The Ministry of Education introduced this program in 2015 with the aim to eliminate cheating. With CBT, the test items are drawn directly from a server, so test versions vary across students and across classrooms. This way, it is much harder to cheat as teachers and students do not know the questions beforehand and students have no reason to work together during the exam. In addition, teachers cannot change students' answers, because the computer program grades the exam. Although this type of testing is not new (for instance, see Wang et al. (2008) for CBT use in the United States), implementing CBT at large scale in a developing country is exceptional. The program started with 40 junior secondary schools in 2015. By 2019, 78 percent of Indonesia's junior secondary schools (43,841 schools with 3,554,556 exam takers) implemented the program. Each year, district governments assigned schools to the program depending on willingness to participate and access to required computer facilities.

Our descriptive analysis shows that the program improved the accuracy of exam scores as a measure of learning outcomes. We expect the change in exam scores to be larger for schools where cheating was common practice. We find that the difference in scores between paper-based exams in 2015 and computer-based exams in the first year that CBT was implemented is indeed negatively correlated with the integrity index. In addition, our descriptive analysis shows that cheating was concentrated in certain regions. The correlation between school rankings based on exam scores before and after the introduction of CBT is low. We find that this is mainly

due to a reversal of ranks across districts rather than within districts.

The analysis is based on publicly available data on the average exam score, the variance, the number of students taking the national exam and the integrity index at the school level. By construction, the integrity score is only available for schools using the paper-based test.[1] We apply difference-in-difference methods for each cohort that switched to CBT between 2016 and 2019, for which we compare the change in exam scores between 2015 and the years in which they implement CBT to the change in exam scores of matched control groups drawn from schools that never implemented the program. Matching was done based on pre-2015 test scores and the integrity index. By using these methods rather than an event study model, we avoid concerns which have been raised in recent literature on interpreting the results of event study models (Abraham and Sun, 2018; Chaisemartin and D'Haultfoeuille, 2019; Goodman-Bacon, 2018).

We find that average school level exam scores decreased on average by 5.2 points (0.4 standard deviation) as a result of the introduction of CBT. To confirm that this effect is due to a reduction in cheating rather than a change in the test taking mode, we split the sample by high and low integrity (defining low integrity as being flagged by the Ministry) and ownership of computers in 2015 (used as a proxy for familiarity with working on computers). If this is true, schools with a low integrity index would experience a larger drop in their scores and ownership of computers should not matter much. For schools with low (high) integrity, the average effect is -8.8 (-2.1) points for schools with computers in 2015 while the average effect is -1.6 (-0.6) points larger for schools that did not have computers. The results show that the effect of CBT worked through a reduction in the opportunity to cheat and that familiarity with computers had a relatively small effect. We also find that the variance in test scores at the school level increased with 0.53 standard deviation, suggesting that the CBT method was better able to distinguish between high and low performing students.

In regions where CBT was implemented at a faster pace, the integrity index of control schools in the same district rose faster and their test scores declined. This suggests that the

[1]The algorithm checks whether students copy answers, which is impossible for CBT exams as all students receive different questions.

roll-out of CBT affected local norms with respect to cheating, which caused positive spillovers on non treated schools in the same district. This could arise from the fact that exams are proctored by teachers from other schools in the same district. Teachers from schools that switched to CBT may have become stricter when proctoring schools that conduct paper-based exams to ensure a fair competition. As a robustness check, we correct our impact estimations and allow the trend of the control group to vary with the share of schools that implement CBT in the district and find that it makes little difference.

We believe that the results should be considered a lower bound estimate of the extent of cheating in Indonesian schools in 2015. First, the schools that switched to CBT had on average higher integrity scores than the control schools, so the impact is likely to be larger for those schools that did not convert yet. Second, control schools may have reduced cheating as a result of the dissemination of the integrity scores, which also started in 2015. Third, the roll-out of CBT affected local norms with respect to cheating.

The effect of CBT on test scores becomes smaller as schools implement the program for a longer period. After two years of implementation, the average effect on exam scores becomes insignificant. This suggests that the program may have helped to improve learning. Students and teachers who took the CBT exam in subsequent years had more time to prepare for the exam than those who participated in CBT the first time it was implemented at the school.

This study makes several contributions. First, it adds to a small literature on the effects of programs aimed at reducing cheating in schools. These programs include cameras in classrooms in Romania (Borcan et al., 2017), random assignment of external monitors in Italy (Bertoni et al., 2013), centralization of grading in New York (Dee et al., 2019) and tablet-based testing in India (Singh, 2020). All these studies found that the programs reduced cheating and, in turn, test scores. Our study shows that it is possible for a government of a developing country with widespread cheating to substantially reduce cheating on a high stakes national exam with the utilization of technology. Moreover, the technology decreases the costs of exam administration from about 9.2 million dollars to 2.4 million dollars each year, because printing and distributing the exams on paper is no longer necessary (Siddiq, 2018). In addition,

universities and employers do not have to incur additional costs if they can rely on national exams as an accurate measure of learning achievement. Although the intervention requires an investment in computers, these costs are mostly fixed and the computers can be used for supportive teaching programs outside of exam periods.

Second, we provide the first evidence that CBT supported the transition from a cheating culture to a learning culture. The reduction in the effect of CBT on test scores over time suggests that CBT had an impact on learning. Without additional resources, this must have come from additional effort from students and teachers. This is in line with the finding that the introduction of high stakes testing improves learning outcomes in contexts with little cheating (Jacob, 2005; Angrist and Lavy, 2009). In addition, we find that the introduction of CBT affected local norms with respect to cheating. As more schools were no longer able to cheat, other schools followed suit by also reducing cheating, either through peer pressure or voluntarily. This is in accordance with findings by Bertoni et al. (2013), who found that external monitors in one classroom also reduced cheating in other classrooms in the same school without an external monitor.

The rest of the paper proceeds as follows. In the next section we provide background information on the Indonesian national examination. Section 3 describes the data and section 4 explains our empirical strategy. We report on the impact of CBT on exam scores in section 5 and we discuss the results in the final section.

2 The Indonesian National Examination

Indonesia has a high stakes national examination at the end of junior and senior secondary school (grade 9 and 12, respectively). Students take exams in Indonesian, English, mathematics and science. In 2010, they had to score higher than 55 out of 100 on average across four subjects to graduate. Between 2011 and 2014, the schools gained more autonomy in the graduation of their students when a composite score of the national exam and school exams determined graduation. Graduation has been independent from the national exam since 2015. However,

performing well in the national exam is still important because the national exam score is used to determine admission into higher education levels. This is especially true for the grade 9 exam, which we focus on in this paper. Admission in senior secondary schools depends on the grade 9 exam, while universities have their own entrance exams.

High exam scores are not only important for the students, but also for the schools and district governments. The score contributes substantially to school and local government achievement indicators (Economist, 2011). Although there is no legislation for holding schools accountable on their exam scores, local governments consider performance on the national exam as a matter of prestige. They put pressure on school principals and teachers to achieve high grades.

As argued by Neal (2013), using one assessment system to measure student achievement and school quality creates incentives to cheat for both the students and the educators. Anecdotal evidence indicates that cheating on national exams is indeed widespread (Economist, 2011; Jong, 2015). Students copy each other's answers or use answer sheets, which they illegally buy or receive from the teacher, and the teachers allow for these cheating practices to take place. In some cases, teachers provide the answers. The exam answer sheets are collected and scanned at the provincial level and graded centrally by the Ministry of Education and Culture (MoEC), but the teacher could still interfere with the answer sheets beforehand.

Before 2015, the Government of Indonesia (GoI) tried to prevent cheating on the national exam by increasing the number of unique booklets in an exam room from two to five in 2011, and from five to 20 in 2013. However, students and teachers still managed to cheat. Therefore, since 2015, the GoI took additional measures with the aim to identify and reduce cheating.

First, the Center for Assessment and Learning (Pusmenjar or *Pusat Asesmen dan Pembelajaran*) of MoEC developed an algorithm that generates a score to identify cheating at the school level, called the integrity index. The index detects suspicious response patterns across students within the same schools and districts (Rahmawati and Asrijanty, 2016).[2] Response patterns are seen as an indication of answer copying and therefore bring about suspicion of

[2]The exam only contains multiple choice questions.

cheating. The index is constructed using a combination of previously developed methods to identify answer copying (Hanson et al., 1987; Widiatmo, 2006; Van Der Linden and Sotaridona, 2006). To confirm that the answer pattern is unexpected, the algorithm checks if the school's results across items are in line with the items' difficulty parameter. The additional check is performed because identical wrong answers could also result from that the teacher incorrectly taught the concept that the item tests, which is not an indication of cheating.

Pusmenjar also checks if the school's exam results do not deviate too much from the school's exam results in previous years, school accreditation reports and school quality reports of the provincial government. It would be considered as evidence of cheating when a school with a track record of poor performance achieved uniformly correct answers. In addition, the school-level integrity index is validated against a qualitative measure of school quality determined by respective provincial governments.

The index measures cheating on a continuous scale between 0 and 100, where a lower score means that there is more evidence for cheating. Pusmenjar considers an integrity index below 70 as low integrity, between 70 and 80 as fair integrity and above 80 as high integrity.[3] The integrity index is robust to type 1 errors, but it is prone to type 2 errors. This means that when the score is low, there is compelling evidence for cheating. At the same time, exam scores of schools with high integrity could still include cheating (Rahmawati and Asrijanty, 2016).

The GoI shares the results of the integrity index with district governments to signal that they do not only care about high grades on the national exam, but also about how the exam scores are achieved. However, the GoI does not implement sanctions based on the integrity index.

Second, the most rigorous policy change to prevent cheating is the implementation of Computer-Based Testing (CBT). Students receive the exam items directly from a server with an item bank containing 30,000 items per subject each year. The system draws items from this

[3]These threshold values are based on the correlation between a change in the integrity index and a change in exam scores between 2015 and 2016. Pusmenjar found for schools with an integrity index above 80 in both years that exam scores do not vary much over time, while the exam scores of schools that started with an integrity index below 70 in 2015 and had an integrity index above 80 in 2016 declined substantially (Rahmawati and Asrijanty, 2016).

bank that are given to the students in random order. Randomization happens both horizontally (i.e. different items across forms) and vertically (i.e. different order of items), such that each student in the exam room has a unique test version.

CBT prevents cheating in a number of ways. The test versions vary across students, classrooms and schools. This makes copying answers ineffective for students. In addition, neither teachers, school principals nor students have access to the test beforehand and answer sheets of the paper-based exams are useless. Finally, grading is done automatically as soon as a student completes an exam and encrypted student responses are sent directly to the central server of the GoI, so modification of the student responses by other parties is impossible.

Some parts of the test procedure remained the same as with paper-based testing (PBT). Both paper-based and computer-based exams are monitored by teachers from other schools in the district, who are randomly assigned by the district government. The teacher is not allowed to be in the classroom with his or her own students during the exam. The paper-based and computer-based exams test the same competencies and are the same across Indonesia. The items for each of the 20 paper-based test versions are taken from the same item bank as the computer-based test versions.[4]

CBT is rolled out in phases, starting from 2015. In that year 40 junior secondary schools switched from PBT to CBT. Implementation then ramped up. A total of 43,841 junior secondary schools (78%) implemented CBT in 2019 (see table 1). Schools apply for CBT, after which the district governments determine if these schools can do the exam on computers.[5] Schools are also allowed to use computers in neighboring schools. Once a school switches to CBT, their integrity index is not calculated anymore.[6]

In this paper, we estimate the impact of CBT on school average exam scores. We expect the school mean exam scores to decline with CBT implementation and we expect that

[4]Five test versions are shuffled in four different ways to create 20 test versions per classroom.

[5]The district governments check if the schools have a sufficient number of computers and stable electricity supply. Schools with computers, but without a stable internet connection can download the exams and conduct the exams offline. The questions are only revealed once the students commence the exam.

[6]It is impossible to calculate the integrity index in a comparable way as it is partly based on how often students copy each other's answers. With every student receiving different questions, this method is not possible anymore.

the spread of student exam scores within schools to be larger when they cannot cheat. Low-performing students are likely to benefit more from cheating; the high performing students would have achieved good scores without cheating as well. However, if schools that allow cheating have a homogeneous group of low-performing students the effect on the spread of the exam scores would be minimal.

3 Data and Descriptive Statistics

We use publicly available administrative data from Pusmenjar. The data is called Pamer (*Pengoperasian Aplikasi Laporan Pemanfaatan Hasil Ujian Nasional*) and it reports the national examination results. The data set contains exam score means, the number of students taking the exam, the standard deviations and the integrity index at the school level . We have access to the exam score mean data from 2010 to 2019, standard deviations from 2010 to 2018, and the integrity index from 2015 to 2018.[7] In addition, we know which schools switched to CBT between 2015 and 2018. The exam scores are between zero and 100, and the final exam score is the average score in mathematics, Indonesian, English and science. We complement this data with information on school resources in 2015 from datasets called *Dapodik* and *Sekolah Kita*.

The exam information is available for all junior secondary schools in Indonesia. These include public and private schools. The private school students also take the national exam because it determines continuation to senior secondary school. Our sample consists of 56,500 schools; for our analysis we focus on 50,124 schools that participated in the national exam each year between 2015 and 2019.[8] The data on school resources is only available for 34,412 junior secondary schools that fall under MoEC. We do not have school resource information for private schools under the Ministry of Religious Affairs.

Schools selected into the treatment. Table 1 shows the average exam score and integrity

[7] A CD with the national examination data, including the integrity index, can be requested from the Ministry of Education and Culture. Exam score data between 2015 and 2019, but not the integrity index, can also be accessed at https://hasilun.Pusmenjar.kemdikbud.go.id/.

[8] There are 188 panel schools that switched to CBT but switched back to PBT before 2019. We leave these schools out of the analysis.

index and access to electricity, internet and computers in 2015 for all junior secondary schools, grouped by the year in which the schools switched to CBT. CBT implementation started in 2015. In the first two years, only a small percentage of schools took exams on computers. From 2017 onward, large groups of schools switched. In 2019, there were 10,750 schools that still took exams on paper. We use these schools as our control group (applying the appropriate weights). Schools without access to computers, electricity and internet in 2015 tend to switch to CBT later, as expected. Schools are allowed to use the computers of neighboring schools. Of the junior secondary schools that implemented CBT in 2019, 29 percent did not own the computers themselves (Kemdikbud, 2019). Schools that switched later also had lower average exam scores and integrity in 2015.

Table 1: Staggered Adoption of CBT

CBT Cohort (Year)	(1) Schools (#)	(2) Cumulative (%)	(3) Exam Score (Mean 2015)	(4) Integrity (Mean 2015)	(5) Electricity[1] (% 2015)	(6) Internet[1] (% 2015)	(7) Computers[1] (% 2015)
2015	40	0.1	78.4	-	100	89.7	92.3
			(8.0)		(0.0)	(30.7)	(27.0)
2016	856	1.8	69.6	77.5	100	96.6	82.9
			(12.5)	(10.6)	(0.0)	(18.1)	(37.7)
2017	9,377	20.5	62.1	75.2	99.9	91.0	68.4
			(12.3)	(11.7)	(3.5)	(28.6)	(46.5)
2018	16,183	52.8	62.3	67.4	98.6	82.4	47.2
			(13.5)	(16.9)	(11.6)	(38.1)	(49.9)
2019	12,963	78.6	59.6	68.2	96.8	75.3	34.7
			(12.5)	(16.0)	(17.7)	(43.2)	(47.6)
No CBT	10,705	21.4	58.7	65.6	86.4	62.6	18.1
			(12.4)	(17.9)	(34.3)	(48.4)	(38.5)

Note: The table includes 50,124 panel schools. Standard deviations are provided between parentheses.
[1] We only have this information for schools that fall under the Ministry of Education. These are 33,331 schools in total, or from the first to the last row: 39, 766, 7,335, 8,004, 7,296 and 9,891 schools.

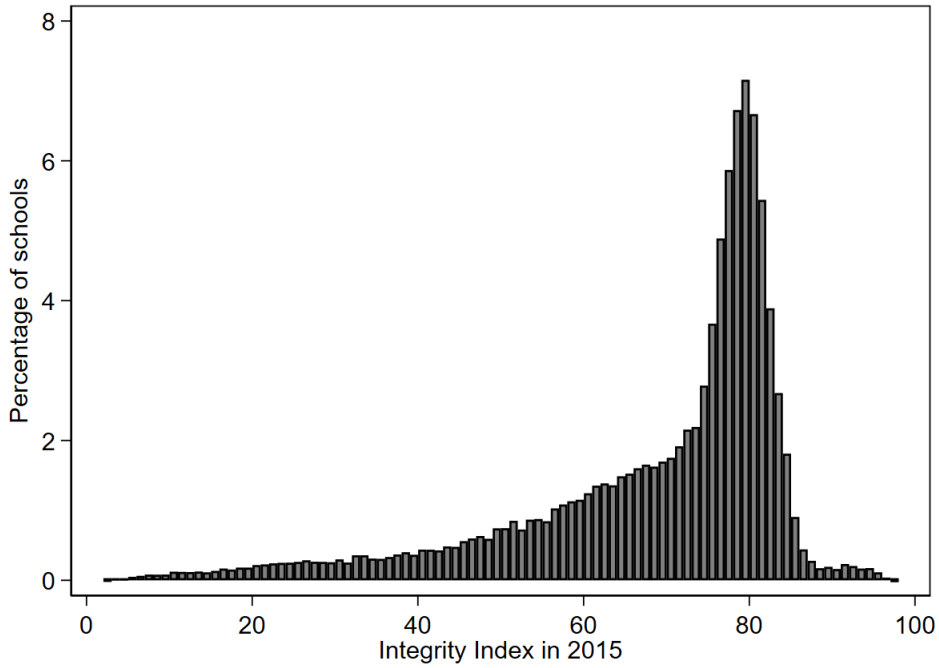
Figure 1 presents the distribution of the integrity index in 2015 and box plots of the 2015 exam scores of schools grouped by their integrity. The integrity index ranges between 0 and 100. Only 24 percent of schools achieved high integrity above 80 in 2015. Moreover, a third of the schools scored below 70, which Pusmenjar uses as a threshold for sufficient evidence for cheating. This is more than in Italy and Chicago, where a similar algorithm flagged the exams of about 5 percent of classrooms as compromised (Angrist et al., 2017; Battistin et al., 2017),

but less than in Andhra Pradesh, India, where a similar algorithm flagged 38 to 43 percent of classrooms (Singh, 2020). The results confirm that cheating was widespread when CBT was introduced.

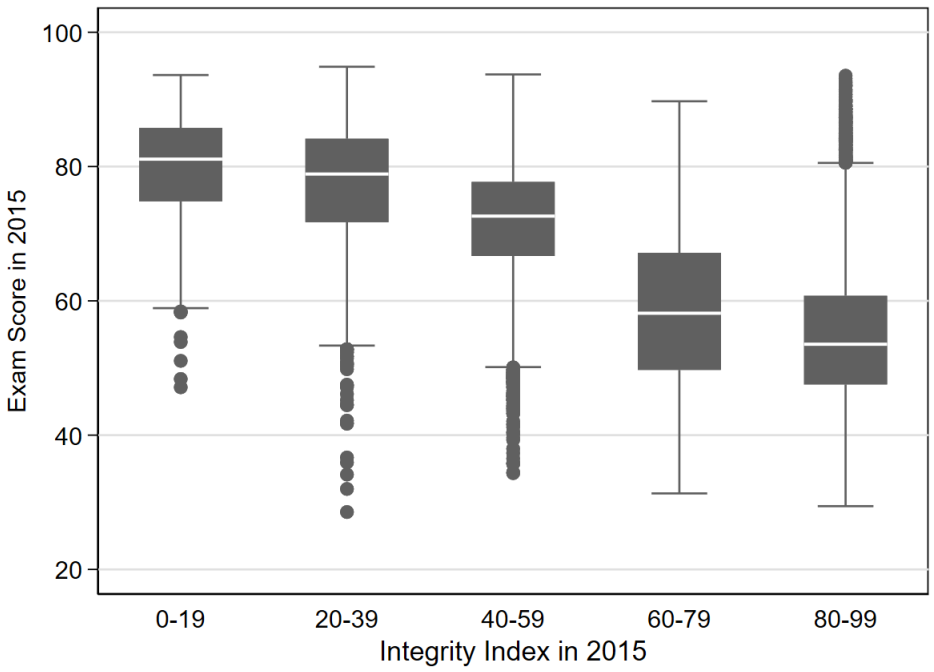
The box plots of the exam scores in figure 1b show that the lower the integrity index, the higher the exam scores.[9] In addition, it shows a high variance in the exam scores of schools with high integrity, meaning that there were both schools with high and low exam scores that did not cheat. The graph also shows that a high school average exam score does not automatically translate into a high integrity index. As shown by the box plots, the integrity index can distinguish between high scoring schools that do and do not cheat.

[9]The pairwise correlation coefficient of the integrity index and exam scores in 2015 is -0.6 and is statistically significant at the 1 percent level.

Figure 1: Integrity Index Distribution and Correlation with Exam Scores



(a) Distribution of Integrity Index

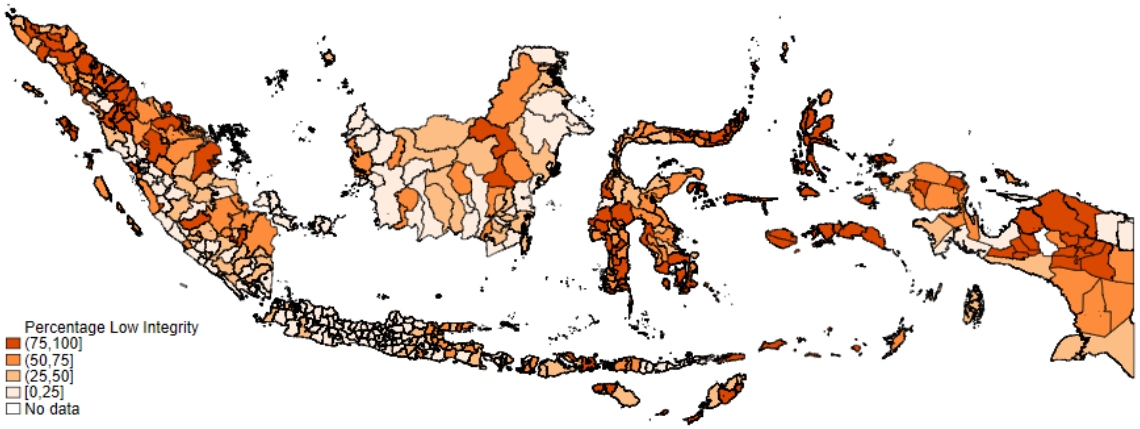


(b) Exam Score Distribution by Integrity

Note: Figures include 44,186 schools for which the 2015 integrity index is non-missing. Panel (a) has a band width of 1. Panel (b) shows the median, the 25th and the 75th percentile, the upper and lower adjacent values and outliers.

There is a strong regional dimension to cheating in Indonesia. Figure 2 shows the percentage of schools that had an integrity index below 70 in 2015 by district. Districts with many low integrity schools were often located next to each other. The regional concentration of cheating was also apparent in Italy, where most cheating took place in the southern provinces (Angrist et al., 2017).

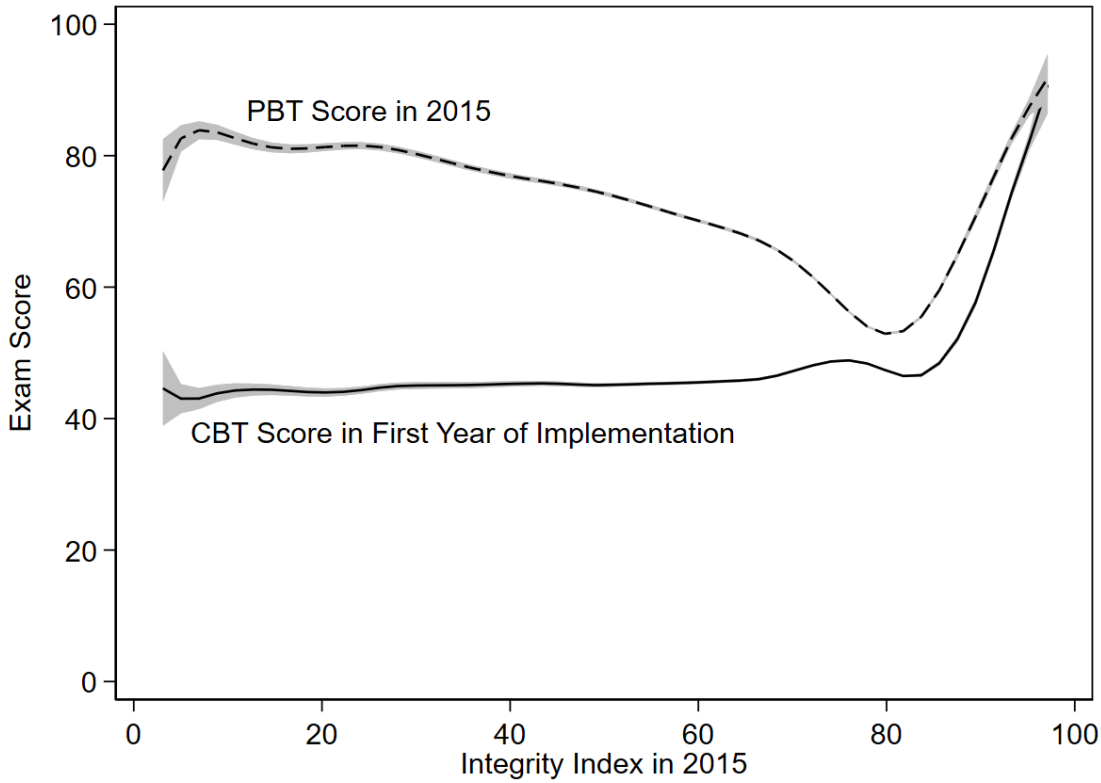
Figure 2: Regional Variation in Integrity in 2015



Note: Figure shows the percentage of schools with an integrity index below 70 (called low integrity) in each district.

To get an idea of how CBT affected the exam scores we plot the 2015 exam score and the exam score in the first year the school implemented CBT as function of the integrity score in 2015 (see figure 3). The figure indicates that in 2015, high scores could be either obtained through cheating or in an honest way. After switching to CBT, however, the schools that did so by cheating saw their exam score drop substantially. For schools with an integrity score below 70, the exam score dropped by 27 points on average. For honest schools we also observe a drop, but it is much more modest. There are various explanations for the latter phenomena. The exam taking method, changes in the difficulty of the exam over time or the fact that integrity score is prone to type 2 errors could all have contributed to the drop. In the impact analysis in section 5 we correct for some of these effects.

Figure 3: Difference between Paper-Based and Computer-Based Exam Scores of Treatment Schools by Integrity



Note: The lines represent smoothed results of a local polynomial regression. The figure includes 34,783 out of 39,379 treatment schools for which the 2015 integrity index is non-missing. The CBT score polynomial regression result combines the exam scores of all treatment schools in the first year of CBT implementation, which is between 2016 and 2019. 95% confidence interval in grey.

The implementation of CBT was accompanied by a stark reversal of the rankings across schools and districts. Table 2 presents the rank correlations of the average exam score of 2015, 4 years before and 4 years after, for 226 out of 514 districts in which all schools implemented CBT by 2019.[10] The first three columns look at the rank correlation of the national ranking. In years before the implementation of CBT, the rank correlation varied between 0.45 and 0.61 with higher rank correlations closer to the base year. In the years after CBT, the same pattern is observed, but the rank correlation dropped to 0.18 in 4 years. Column 2 presents the average rank correlations of schools within districts. Interestingly, the opposite pattern arises. Average

[10]We performed the same exercise on the full sample and found similar but less distinct patterns, see table A1 in the Appendix.

rank correlations after CBT are higher than they were before CBT started. As shown in column 3 and 4, this pattern is driven by districts with high average integrity. On the other hand, the rank correlations (based on the average score of the districts) across districts dropped sharply after the implementation of CBT. While the rank correlation was in the range of 0.53 to 0.63 before the start of CBT, it turned even negative in years thereafter. The evidence shows that the loss in rank correlation is mostly resulting from rank reversals across districts, and less so from rank reversal of schools within districts.

Table 2: Rank Correlation over Time for Districts with Full CBT Implementation by 2019

	(1)	(2)	(3)	(4)	(5)
	School Percentile	School Decile Within District			District Rank
		All	District Integrity < 70	District Integrity >= 70	
2011	0.45	0.33	0.24	0.38	0.53
2012	0.50	0.36	0.26	0.42	0.59
2013	0.59	0.49	0.34	0.59	0.58
2014	0.61	0.56	0.42	0.65	0.65
2015	1	1	1	1	1
2016	0.67	0.64	0.50	0.73	0.67
2017	0.48	0.52	0.34	0.65	0.51
2018	0.23	0.43	0.17	0.61	0.06
2019	0.18	0.44	0.16	0.62	-0.1
Observations	24,028	24,028	9,536	14,492	226

Note: Table shows the Pearson pairwise correlation coefficient of the rank in each year with the rank in 2015. It includes 24,028 panel schools from 226 (out of 514) districts in which all schools implemented CBT by 2019. None of these schools implemented CBT in 2015, 3 percent in 2016, 35 percent in 2017 and 80 percent in 2018. There are between 11 and 952 schools in a district (228 on average). The integrity categories in column 3 and 4 are determined based on average district integrity in 2015.

4 Empirical Strategy

To measure the impact of CBT on test scores, we conduct a series of difference-in-difference (DiD) estimations. We do this separately for each group of schools that switched to CBT in

a different year (we call these cohorts). We use the schools that still took exams on paper in 2019 as the control group to determine the general trend in exam scores. We compare the exam scores of the treatment schools (that switch to CBT in any year between 2016 and 2019) and the control schools (that did not switch until 2019) between 2015 and the years in which they implemented CBT. We leave the first 40 schools that switched to CBT in 2015 out of our analysis.

Our analysis depends on two important assumptions. First, we assume that the average student ability within schools is stable over time. Each year a different group of students took the exam. We are only able to attribute a difference in exam scores over time at the school level to CBT if the underlying ability of the students remained the same. The assumption would be violated if students changed schools because of CBT. We argue that this is unlikely. Students enrolled three years before they took the exam, so they could not anticipate whether their school would opt into CBT. Moreover, it is difficult for students to change junior secondary schools once they are enrolled due to the centralized school admission system for public junior secondary schools. To provide evidence for a stable student composition over time, we estimate the impact of CBT on the number of exam participants of each school and find no effect (see table A3 in the Appendix). We discuss these results further in section 5.1.

The second assumption is the common trend assumption. We assume that the trend in exam scores of the treatment and control group would have been the same if the exams would have remained on paper. If the trend in exam scores before 2015 differs between the treatment and control schools, it would be implausible to assume that the exam score trend would have been the same after 2015. We found that the exam score trend between 2010 and 2015 was somewhat different between the control group and each of the treatment cohorts (figure 4a). Moreover, the level of exam scores of the control schools was lower than that of the treatment cohorts.

To increase the probability that the common trend assumption holds, we match each treatment cohort with the control group based on their average exam score between 2010 and 2015 and their integrity index in 2015. Studies have shown that matching the control group to

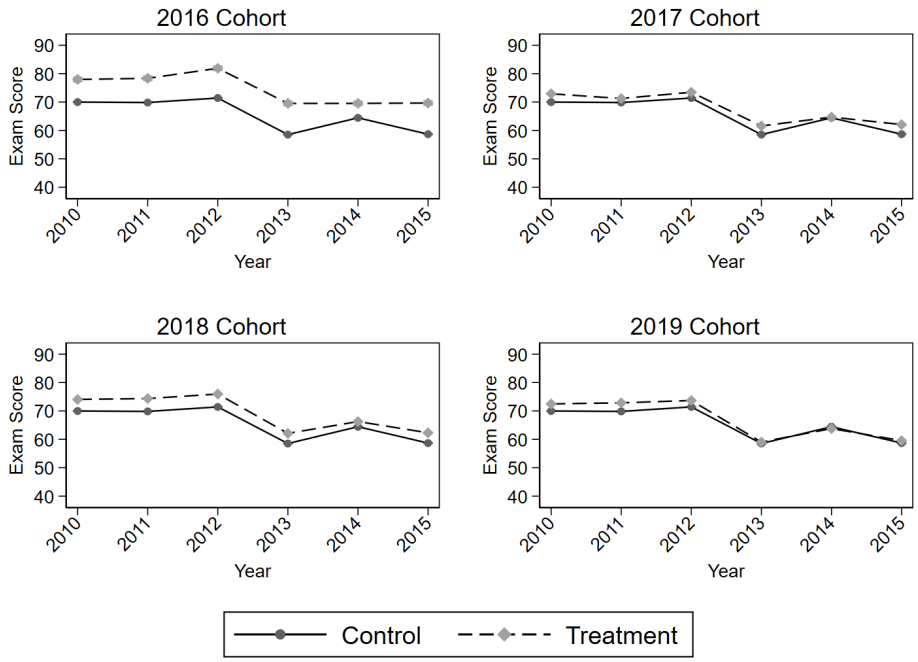
the treatment group based on pre-intervention characteristics improves the performance of the DiD estimator (Abadie et al., 2010; Linden and Adams, 2011; Ryan et al., 2019). We follow the propensity score-based weighting model by Linden and Adams (2011).[11] We estimate the propensity to be treated using a logit model of CBT implementation on the exam scores between 2010 and 2015 and the integrity score in 2015, including dummy variables for missing values. We weight the control group exam scores with the inverse of this propensity score.

Comparing the distributions of the predicted propensity score between the control group and each of the treatment cohorts, we find that there is substantial overlap for schools that switched to CBT in 2017 and later (see figure A1 in the Appendix), but there is a lack of common support for the 2016 cohort. The schools in the 2016 cohort had higher average integrity and higher average exam scores than the control group (see table 1). Because of the lack of common support, we do not report impact estimates for the 2016 cohort (2 percent of treated schools).

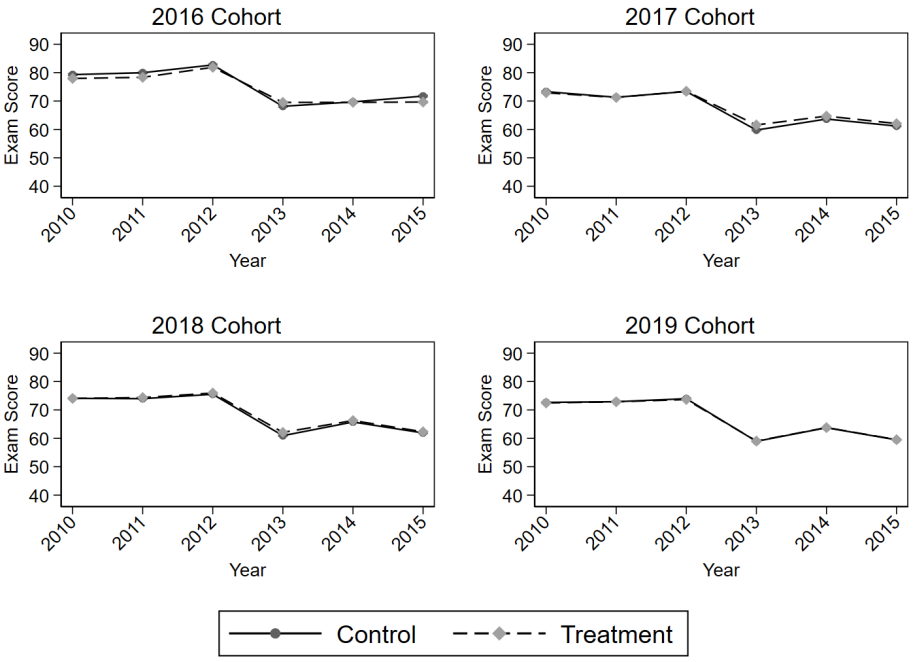
Figure 4b plots the weighted mean exam scores for each cohort. The matching worked well, since the weighted average exam scores of the control group are virtually the same as those of each of the treatment cohorts between 2010 and 2015. A detailed balance table is provided in table A2 in the Appendix.

[11]Note that this method is different from the synthetic control approach as in Abadie et al. (2010) The synthetic control approach is developed for an analysis with one treated unit and it is complicated to expand this method to multiple treated units, as in our analysis. Linden and Adams (2011) show that their propensity score-based weighting model generates similar results as the synthetic control approach by Abadie et al. (2010) with multiple treated units.

Figure 4: Exam Score Trend Comparison for Each Cohort



(a) Raw Exam Score Trends of Treatment and Control Schools



(b) Matched Exam Score Trends of Treatment and Control Schools

Note: Mean exam scores by year and cohort for panel junior secondary schools before they implemented CBT. The control group in the figures with raw mean exam scores is the same across the cohorts in panel (a). In panel (b), control school means are weighted with the inverse of the propensity score for each cohort separately.

We are interested in the effect of CBT on school mean exam scores and the spread of the exam scores within schools.[12] We estimate the following model for each of the treatment cohorts separately using data between 2015 and 2019,

$$Y_{sdt} = \sum_{y=2016}^{2019} \beta_y \cdot \mathbf{1}\{t = y\} + \sum_{y=2016}^{2019} \gamma_y \cdot \mathbf{1}\{t = y\} \times T_{sd} + u_s + \epsilon_{sdt} \tag{1}$$

where Y is the school average exam score or the standard deviation of student exam scores within a school for school s in district d in time t. $\mathbf{1}\{t = y\}$ is an indicator for time t being any of the years in y (using 2015 as the baseline year), T indicates the treatment status of schools and u are school fixed effects. The treatment effect is captured by coefficient γ. We expect γ to be zero in years that the treatment schools still took the exam on paper. In the years that the treatment schools implement CBT, we expect γ to be negative for the school mean exam scores and positive for the within school standard deviation of the exam scores.

We perform a separate DiD estimation for each CBT cohort, because recent studies found that estimating the weighted sum of the average treatment effects in each group and period is difficult when there are heterogeneous treatment effects. The weights of some groups could be negative in that case, such that the result cannot be causally interpreted (Abraham and Sun, 2018; Goodman-Bacon, 2018; Chaisemartin and D'Haultfoeuille, 2019). Because of the integrity differences between the CBT cohorts, we expect heterogeneous treatment effects and estimate the average treatment effect for each cohort separately. For ease of discussion, we also present the weighted average of the DiD results. As weights we use the share in the number of schools of each cohort, following suggestions by Goodman-Bacon (2018) and Chaisemartin and D'Haultfoeuille (2019). These results can be interpreted as a sample-weighted average treatment effect on the treated schools for the immediate effect. For the lagged effects, it is important to realize that different combinations of cohorts contributed to the estimates.

We present heterogeneous treatment effects by the school's integrity level in 2015, and

[12]Since the exam has not been a graduation requirement since 2015, we cannot analyze passing rates.

by whether the school had access to computers in 2015.[13] Both are included to test whether the effects we observe from CBT are indeed resulting from a reduction in cheating, and not from other factors associated with the method of exam taking. If it is due to a reduction in cheating, the effect should be greater for schools with a low integrity score. On the other hand, if it is due to the switch from paper-based to computer-based testing, the effect should be smaller for schools that already had computers in 2015 and where students were more familiar with working on computers.

Spillovers could arise if the roll-out of CBT in a district results in a norm change with respect to the acceptability of cheating. For example, exam supervision is organized by district governments, which allocate teachers from different schools as proctors to supervise exams. If these proctors came from schools that switched to CBT, they may have been stricter than usual because their school had no option to cheat anymore. Allowing the other school to cheat would lead to an unfair competition between schools. In addition, the distribution of answer sheets among students and teachers may have been disrupted because the answer sheets are of no use to the ones that took the exam on computers, such that the probability that a student that took the exam on paper acquired an answer sheet becomes smaller as more schools switch to CBT. To investigate whether this hypothesis is true, we estimate equation 2 using data from the control group schools only

$$Y_{sdt} = \sum_{y=2016}^{2019} \beta_y \cdot \mathbf{1}\{t = y\} + \eta_1 \cdot \overline{T}_{dt} + u_s + \epsilon_{sdt} \tag{2}$$

where Y is the mean exam score or the integrity score in the years 2015 to 2019 of the schools that had not implemented CBT yet by 2019. $\overline{T}_{dt}$ is the fraction of schools that implemented CBT in district d in year t. We expect that a higher fraction of schools that implemented CBT in the district is correlated with lower exam scores in the control group, and a higher integrity score.

As a robustness check, we also correct the main estimation results for potential spillover

[13]This analysis includes 30,198 out of 50,124 schools for which both the integrity score and school resource information are available.

effects by allowing the test scores of the control schools to vary with the share of schools that switched to CBT in equation 1. We estimate the following model

$$Y_{sdt} = \sum_{y=2016}^{2019} \beta_y \cdot \mathbf{1}\{t = y\} + \sum_{y=2016}^{2019} \gamma_y \cdot \mathbf{1}\{t = y\} \times T_{sd} + \eta_1 \cdot (1 - T_{sd}) \times \overline{T}_{dt} + u_s + \epsilon_{sdt} \quad (3)$$

which is the same as equation 1, but with the addition of $\overline{T}_{dt}$ interacted with a dummy variable that indicates control schools. Conditioning on the fraction of treated schools in the district, we expect the negative treatment effect to be larger because we expect a more downward trend in exam scores of control schools in districts with a higher fraction of treated schools. Because the speed of the roll-out of CBT in a district may be endogenous, we present the spillover analysis as a robustness check rather than as our preferred estimate.

5 Results

5.1 School Average Exam Scores

CBT resulted in a drop of about 5 points in the first year of implementation on average. In figure 5 we plot the estimated treatment effect in each year for each cohort (The detailed regression results can be found in table A4 in the Appendix). The null estimates in years prior to opting into CBT can be seen as placebo estimations and confirm that the effect arises in the year of opting in.

The results also indicate that the CBT effect reduced over time, suggesting a positive effect on learning. The effect is clearly visible for the 2017 cohort but small for the 2018 cohort. It is unlikely that this is due to new cheating methods. While cheating practices on paper-based exams were discussed at length in newspaper articles, there have been few reported cases of

cheating on the computer-based exams (Biantoro and Arfianti, 2019).[14] Without the possibility to cheat, the only way to achieve high grades in through studying. Students participating in later years had more time to prepare for the exam under the new rules. Another explanation is that schools needed time to become familiar with computerized test administration. We do not believe that this is a major factor as the heterogeneous treatment results, reported later, show that the impact of CBT varied little between schools with and without computers.

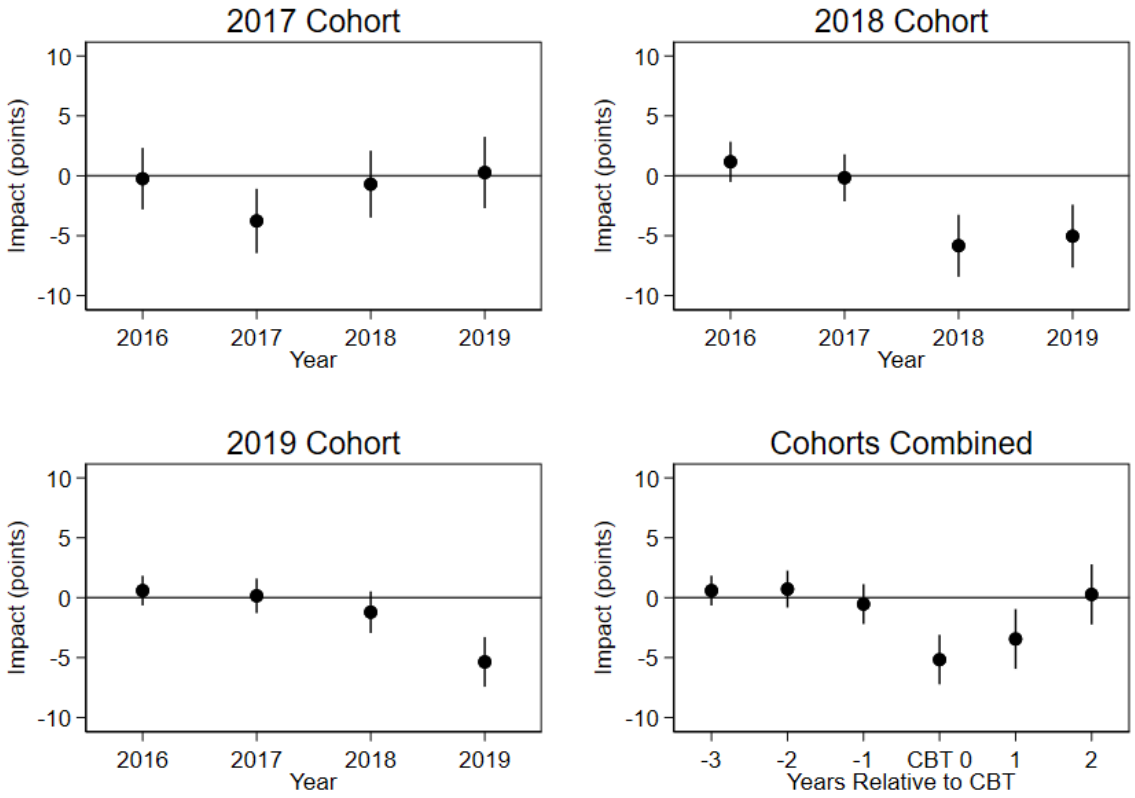
To provide evidence that these results were not due to students moving from treatment to control schools, we also estimate the impact on the number of exam participants in each school. One could be concerned that the improvement in exam scores over time comes from students that were likely to score low on the computer-based exam moving to schools that still took the exam on paper. In that case, we would expect to see a negative impact of CBT on the number of exam participants. The results in table A3 in the Appendix confirm that this is not the case.

Table A5 presents the impact estimation results on exam scores in terms of standard deviations. We used the within-school standard deviation of the test scores, the school level mean exam scores and the number of students that took the exam to calculate the student level mean and standard deviation of the control group exam scores in each year and used these to standardize the exam scores. [15] Average school level exam scores drop with 0.4 standard deviation in the first year of CBT implementation (see table A5 in the Appendix).

[14]There was one teacher that managed to connect his computer with those of the students such that he could control their computers from a distance (Abdi, 2019) and there were some students who took photos of the computer screen during the exam to share questions with others (Alfons, 2019). These cheating methods cannot explain the rise in exam scores, because each student gets a unique test version and because hacking requires advanced knowledge of computer systems that most Indonesian teachers do not have.

[15]We do not have access to the within school standard deviation of exam scores in 2019, so for that year we assume that the ratio between the sum of squares across groups and the sum of squares within groups is the same as in 2018.

Figure 5: Impact Estimation Result on School Exam Scores

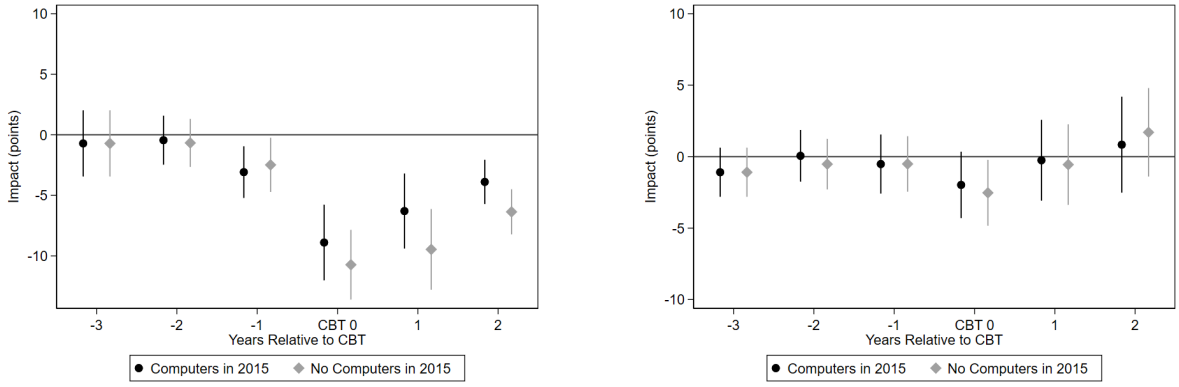


Note: Plot of point estimates of γ_y in equation 1 with 95% confidence interval, estimated separately for each cohort and weighted by the inverse of the propensity score. Standard errors are corrected for clustering at the district level. The 'cohorts combined' figure shows the sample-weighted average effect across cohorts. Detailed results are available in table A4.

Schools with low integrity and those without computers in 2015 were more affected by the switch to CBT. The effect of integrity is much larger than the effect of having computers, indicating that the effect of the CBT mainly operated through a reduction in cheating rather than through the change in test taking mode (from paper to computers). Figure 6 plots the estimates separately for schools with an integrity index below 70 and above 70, and with and without computers in 2015. We focus on the combined estimates reported in table A6 in the Appendix. For low integrity schools, CBT resulted in a 9-point drop in exam scores while for high integrity schools the drop was only 2 points. Not having computers resulted in a 1.6-point drop for low integrity schools but had no significant effect for high integrity schools

indicating that for the latter group, familiarity with computers did not drive the small drop in exam scores.[16] This result is in line with evidence from the US (Wang et al., 2008) and India (Singh, 2020), that show that computer-based testing yields similar results as compared to paper-based testing when there is no scope for cheating in either.

Figure 6: Impact Estimation Result on School Exam Scores by Baseline Integrity and Computer Ownership



(a) Schools with 2015 Integrity < 70

(b) Schools with 2015 Integrity >= 70

Note: Plot of point estimates of γ_y in equation 1 with 95% confidence interval, interacted with a dummy variable for ownership of computers in 2015, and estimated separately for each cohort and integrity category and weighted by the inverse of the propensity score. The figure includes 30,198 schools for which the integrity index and computer information is available in 2015. Figure shows sample-weighted average effect across cohorts that switched to CBT in 2017, 2018 or 2019. The integrity categories are based on the integrity index in 2015. Standard errors are corrected for clustering at the district level. Detailed results are available in table A6.

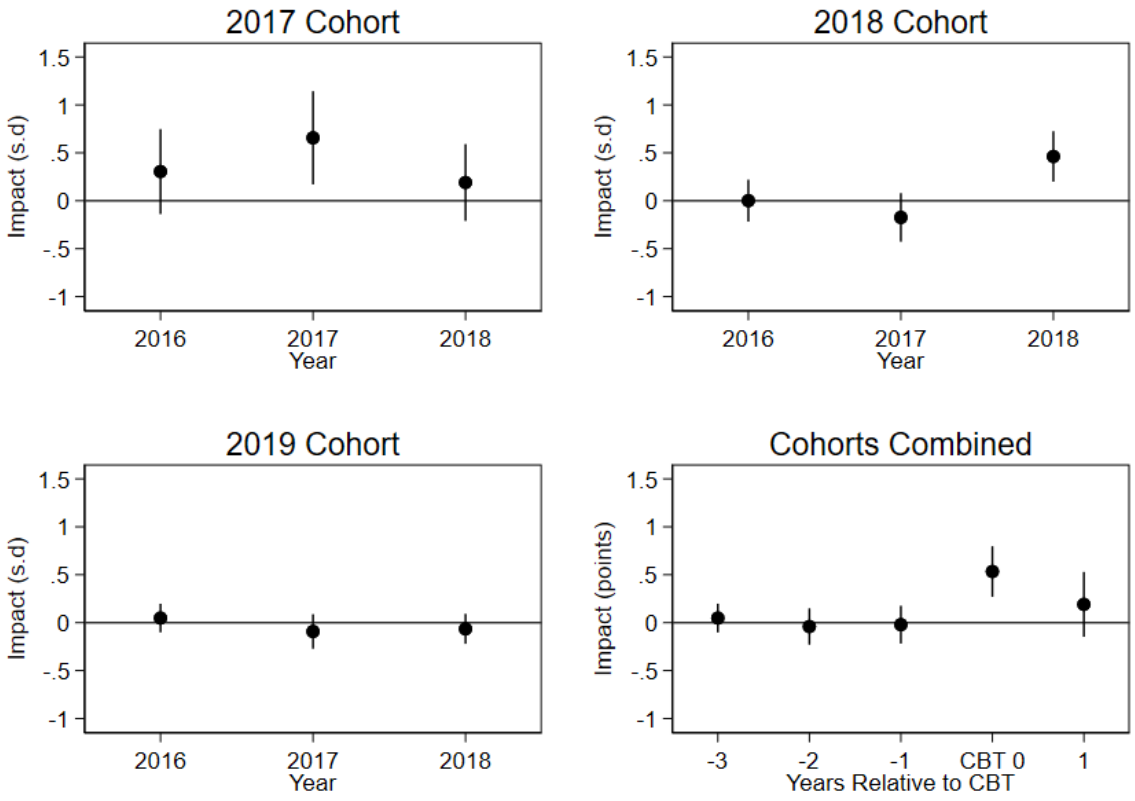
5.2 Variance of the Exam Scores Within Schools

As expected, the standard deviation of exam scores within schools increased with 0.7 and 0.5 for the 2017 and 2018 cohorts, respectively, when these schools switched to CBT (figure 7). This is equal to 12 and 8 percent of the control group mean standard deviation of exam scores within schools (see table A7 in the Appendix). The standard deviation within schools did not significantly change before 2019 for the schools that switched to CBT in 2019.

[16]The drop in exam scores of high integrity schools could be due to the integrity index being conservative. The creators of the index are confident that the exams include cheating when the integrity index is lower than 70, but they are not certain if the exams of schools with an integrity index above 70 do not include cheating (Rahmawati and Asrijanty, 2016). Our results suggest that there were some schools with integrity above 70 that cheated but were not detected by the algorithm.

As described before, it is likely that lower performing students benefited more from cheating before CBT. The disappearance of the treatment effect on the standard deviation in the second year of CBT implementation suggests that lower performing students improved their test scores more than higher performing students.

Figure 7: Impact Estimation Result on Standard Deviation of Exam Scores Within Schools



Note: Plot of point estimates of γ_y in equation 1 with 95% confidence interval, estimated separately for each cohort and weighted by the inverse of the propensity score. Standard errors are corrected for clustering at the district level. The 'cohorts combined' figure shows the sample-weighted average effect across cohorts. Detailed results are available in table A7.

5.3 Correction for Spillover Effects as Robustness Check

To investigate the spillovers, we first look at the correlation between a change in the fraction of schools in the district that implemented CBT and a change in exam scores and the integrity index of the control schools (table 3). We look at the district level, because education policy is

determined at that level and the proctors are assigned within the district. Recall that it was not a whole district that switched, but schools within districts opted in. The average district has 97 junior secondary schools.

The more schools in a district switched to CBT, the lower the exam scores of the control schools and the higher their integrity. Only the exam scores of schools with integrity below 70 significantly decreased as more schools in the district switched to CBT, suggesting that the exam score difference was due to a reduction in cheating practices. Note that the speed in which a district converts to CBT may be endogenous. One should thus be careful to interpret these estimates as causal.

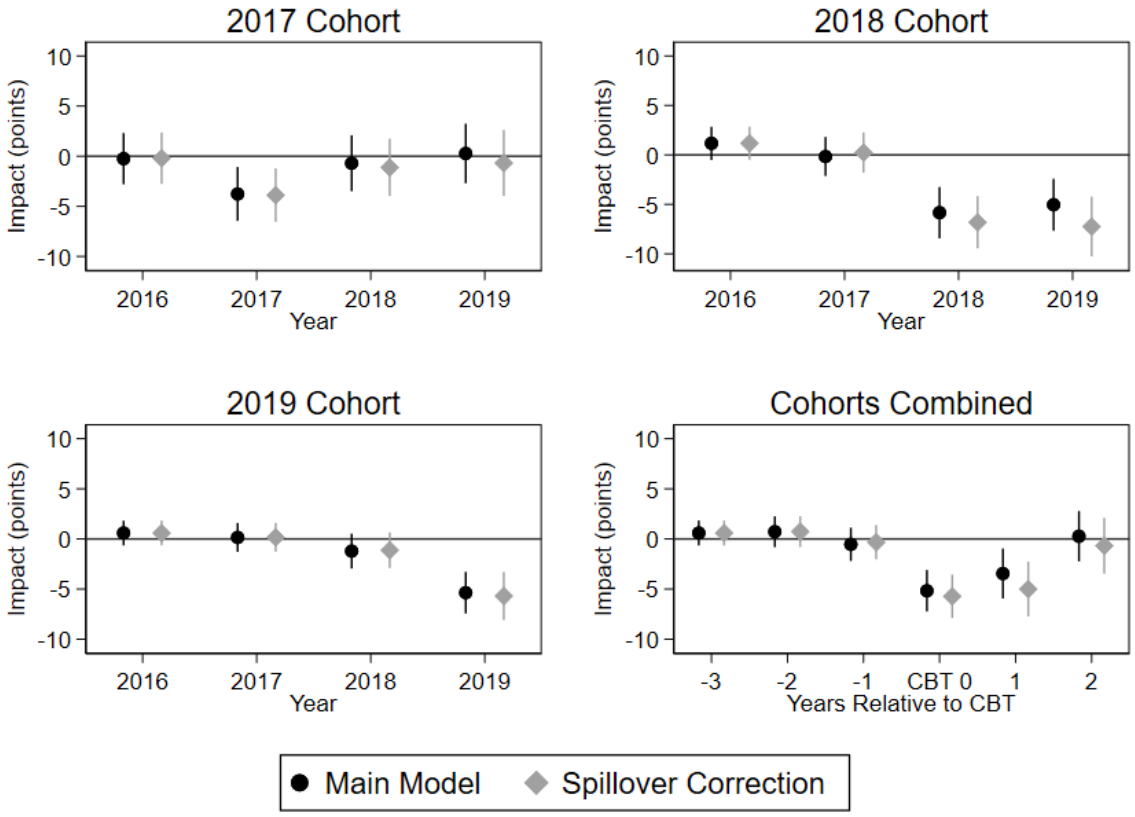
Table 3: Correlation between a Change in the Fraction CBT in District and a Change in the Exam Scores and Integrity Index of Control Schools

	(1)	(2)	(3)	(4)
		Integrity < 70	Integrity >= 70	
	Exam Score	Exam Score	Exam Score	Integrity Index
Fraction CBT in the district	-1.65	-9.78	0.83	10.34
(excluding the observed school)	(1.77)	(2.77)**	(1.51)	(2.87)***
School Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	53,525	21,360	25,655	40,352
Number of Schools	10,705	4272	5131	10,671
R^2	0.17	0.30	0.09	0.07

Note: Plot of point estimates of η_1 in equation 2 with 95% confidence interval, estimated on the control schools. Standard errors between parentheses and corrected for clustering that the district level. Each regression includes year and school fixed effects. Column 2, 3 and 4 have less observations than column 1 due to missing values of the integrity index. * p<0.10 ** p<0.05 *** p<0.01

In figure 8 we present the main results with and without the correction for spillover effects as discussed in equation 3. The effect of the spillover correction on the estimated effect of CBT, as reported in table A4 in the Appendix, is small. Although the coefficients of the treatment effect in 2018 and 2019 for the 2018 cohort are significantly different between the two models when performing a Wald test (results not reported), the difference does not alter our conclusion. Hence, our impact estimates are robust against controlling for the spillover effects.

Figure 8: Impact Estimation Result With Correction for Spillover Effects



Note: Plot of point estimates of γ_y in equation 1 (black) and 2 (grey) with 95% confidence interval, estimated separately for each cohort and weighted by the inverse of the propensity score. Standard errors are corrected for clustering at the district level. The 'cohorts combined' figure shows the sample-weighted average effect across cohorts. Detailed results are available in table A4.

6 Discussion and Conclusion

Cheating on Indonesian national exams was a widespread problem, so much that the Ministry of Education and Culture began to measure the magnitude of the problem by developing an "integrity index" that used answer patterns to detect cheating. In 2015, 33 percent of the junior secondary schools had an integrity index below 70, a threshold that indicates strong evidence of cheating.

To prevent this problem, the Ministry of Education and Culture decided in 2015 to use computer-based testing (CBT) for the national exam (taken in grade 9 and 12). CBT

implementation was staggered for exams in grade 9, so we were able to compare schools that implemented it years before other schools. We find that CBT caused a decline in scores. By comparing the treatment effect of high and low integrity schools and schools with and without computers in 2015, we confirm that the decline in exam scores was mainly driven by a reduction in cheating. Exam scores improved again over time, such that the effect of CBT becomes statistically insignificant after two years of implementation. In addition, we find that in regions where CBT was introduced faster, cheating in schools which conducted paper-based exams reduced more. Our finding that cheating was regionally concentrated points to the existence of a "cheating culture" in certain regions, which is in accordance with findings of Bertoni et al. (2013) and Martinelli et al. (2018) that indicate the existence of a cheating culture in certain schools where this behavior seemed to be the norm.

The Indonesian experience suggests that technology can contribute to the transition from a cheating culture to a learning culture. A cheating culture is a low equilibrium. It is perpetuated because all parties, students, teachers and bureaucrats alike, have an incentive to cheat if everyone else does so as well. But if one can easily obtain high marks by cheating there is no incentive to put in effort for the final exam. In this way, the cheating culture may have contributed to the lack of progress on learning outcomes observed in Indonesia (Beatty et al., 2018).

While CBT makes it very difficult to cheat, no technology is perfect. Without a change in culture, the parties will likely find new ways to cheat. It is therefore encouraging that we found that the cheating norms are local and adjust with the implementation of CBT. High equilibria, with low cheating and a focus on learning, also existed in 2015. The fact that we found that the introduction of CBT also reduced cheating among schools that were still administering paper-based exams gives us hope that CBT helped to move some districts from the low to the high equilibrium where technology will no longer be needed to fight cheating.

References

Abadie, A., Diamond, A., Hainmueller, and Jens (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abdi, A. P. (2019). Kemendikbud temukan kasus kecurangan terstruktur saat unbk smp. *Tirto.id*.

Abraham, S. and Sun, L. (2018). Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects. *SSRN Electronic Journal*.

Alfons, M. (2019). 126 siswa curang saat unbk 2019, kemendikbud: Otomatis nilai nol. *detikNews*.

Anderman, E. M. (2015). India's 'cheating mafia' gets to work as school exam season hits. *The Conversation*.

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review*, 99(4):1384–1414.

Angrist, J. D., Battistin, E., and Vuri, D. (2017). In a small moment: Class size and moral hazard in the Italian Mezzogiorno. *American Economic Journal: Applied Economics*, 9(4):216–249.

Battistin, E., De Nadai, M., and Vuri, D. (2017). Counting rotten apples: Student achievement and score manipulation in Italian elementary Schools. *Journal of Econometrics*, 200(2):344–362.

Beatty, A., Berkhout, E., Bima, L., Coen, T., Pradhan, M., and Suryadarma, D. (2018). Indonesia Got Schooled: 15 Years of Rising Enrolment and Flat Learning Profiles. *RISE Working Paper 18/026*.

Bertoni, M., Brunello, G., and Rocco, L. (2013). When the cat is near, the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics*, 104:65–77.

Biantoro, B. and Arfianti, A. (2019). Issues in the Implementation of Computer-based National Exam (CBNE) in Indonesian Secondary Schools. *Advances in Social Science, Education and Humanities Research*, 353:399–403.

Borcan, O., Lindahl, M., and Mitrut, A. (2017). Fighting corruption in education: What works and who benefits? *American Economic Journal: Economic Policy*, 9(1):180–209.

Chaisemartin, C. d. and D'Haultfoeuille, X. (2019). Two-way fixed effects estimators with heterogeneous treatment effects. *SSRN Electronic Journal*.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. (2019). The causes and consequences of test score manipulation: Evidence from the New York regents examinations. *American Economic Journal: Applied Economics*, 11(3):382–423.

Economist (2011). More cheating, or else! scandals in the classroom. *The Economist*.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. *NBER Working Paper No. 25018*.

Hanson, B. A., Harris, D. J., and Brennan, R. L. (1987). A comparison of several statistical methods for examining allegations of copying. *ACT Research Report Series 87-15*.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5-6):761–796.

Jacob, B. A. and Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics*, 118(3):843–877.

Jong, H. N. (2015). New exam system fails to prevent cheating, leaks. *The Jakarta Post*.

Kemdikbud (2019). Ujian nasional berbaris komputer statistics. `https://unbk.kemdikbud.go.id/#statistik`. Accessed: 2020-05-30.

Linden, A. and Adams, J. L. (2011). Applying a propensity score-based weighting model to interrupted time series data: Improving causal inference in programme evaluation. *Journal of Evaluation in Clinical Practice*, 17(6):1231–1238.

Martinelli, C., Parker, S. W., Pérez-Gea, A. C., and Rodrigo, R. (2018). Cheating and incentives: Learning from a policy experiment. *American Economic Journal: Economic Policy*, 10(1):298–325.

Neal, D. (2013). The consequences of using one assessment system to pursue two objectives. *Journal of Economic Education*, 44(4):339–352.

Rahmawati and Asrijanty (2016). Integrity Index of National Exam: An effort to gain precise information on achievement of curriculum standards. In *Conference Proceedings of the 42nd Conference of the International Association for Educational Assessment*.

Ryan, A. M., Kontopantelis, E., Linden, A., and Burgess, J. F. (2019). Now trending: Coping with non-parallel trends in difference-in-differences analysis. *Statistical Methods in Medical Research*, 28(12):3697–3711.

Safi, M. (2018). Why students at prestigious high schools still cheat on exams. *The Guardian*.

Siddiq, T. (2018). Kemendikbud: Unbk tekan anggaran ujian nasional hingga 70 persen. *Tempo*.

Singh, A. (2020). Myths of Official Measurement: Auditing and Improving Administrative Data in Developing Countries. *RISE Working Paper 20/042*.

Sundaryani, F. S. (2015). Students get high scores by cheating. *The Jakarta Post*.

Van Der Linden, W. J. and Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31(3):283–304.

Wang, S., Jiao, H., Young, M. J., Brooks, T., and Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, 68(1):5–24.

Widiatmo, H. (2006). Metode untuk mendeteksi penyontekan jawaban pada tes pilihan ganda: studi kasus smp di kabupaten garut. *Pusat Penelitian Pendidikan, Balitbang Diknas*, pages 219–226.

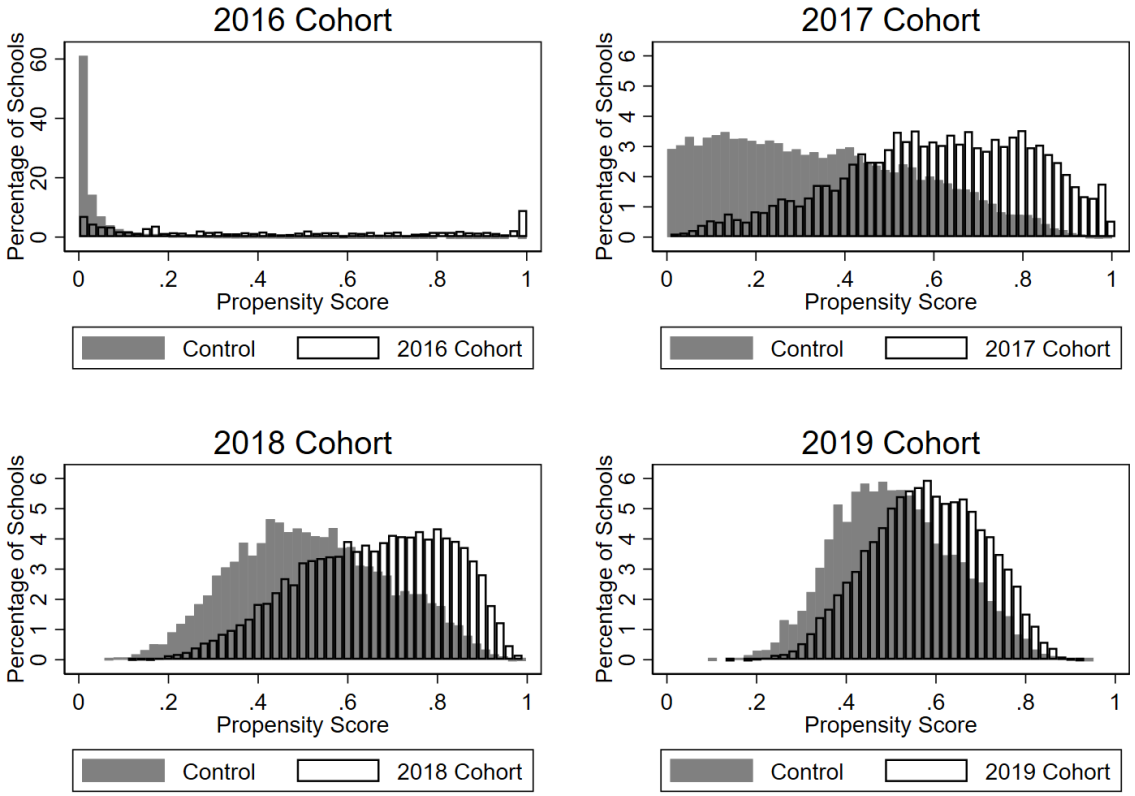
A Appendix

Table A1: Rank Correlation over Time for All Districts

	(1)	(2)	(3)	(4)	(5)
	School Percentile	School Decile Within District			District Rank
		All	District Integrity < 70	District Integrity $>= 70$	
2011	0.37	0.27	0.23	0.30	0.42
2012	0.43	0.29	0.25	0.32	0.50
2013	0.54	0.43	0.33	0.50	0.60
2014	0.62	0.52	0.43	0.59	0.69
2015	1	1	1	1	1
2016	0.65	0.60	0.50	0.68	0.71
2017	0.50	0.50	0.36	0.60	0.61
2018	0.31	0.40	0.23	0.54	0.32
2019	0.24	0.37	0.18	0.51	0.21
Observations	50,084	50,084	21.636	28,448	514

Note: Table shows the Pearson pairwise correlation coefficient of the rank in each year with the rank in 2015. It includes 50,084 panel schools from 514 districts, only excluding 40 schools that switched to CBT in 2015. None of the schools in the table implemented CBT in 2015, 2 percent in 2016, 20 percent in 2017, 53 percent in 2018 and 79 percent in 2019. There are between 6 and 952 schools in a district (107 on average). The integrity categories are determined based on average district integrity in 2015.

Figure A1: Common Support Between Control Schools and Each Treatment Cohort



Note: Propensity score is estimated using the school average exam score in each year from 2010 to 2015 and the integrity index in 2015. Size of bins is 0.02. Y-axis scale of the first histogram that compares the control schools to the schools that switched to CBT in 2016 deviates from the scale of the other figures.

Table A2: Balance Table With and Without Weights

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		2016 Cohort				2017 Cohort			
	Control	Treatment	Diff 1-2	Weighted Control	Diff 2-4	Treatment	Diff 1-6	Weighted Control	Diff 6-8
Exam 2010	69.98	77.95	7.97***	79.31	-1.36***	72.92	2.93***	73.31	-0.40***
	(6.23)	(6.81)	[0.00]	(7.94)	[0.00]	(6.88)	[0.00]	(6.79)	[0.00]
Exam 2011	69.83	78.36	8.53***	79.96	-1.61***	71.25	1.42***	71.32	-0.07
	(9.36)	(7.96)	[0.00]	(8.77)	[0.00]	(10.50)	[0.00]	(9.59)	[0.64]
Exam 2012	71.41	81.87	10.46***	82.72	-0.85***	73.44	2.03***	73.36	0.08
	(10.04)	(8.15)	[0.00]	(7.11)	[0.00]	(10.95)	[0.00]	(9.97)	[0.61]
Exam 2013	58.52	69.52	11.00***	68.14	1.38***	61.58	3.06***	59.78	1.80***
	(10.81)	(11.10)	[0.00]	(9.77)	[0.00]	(11.51)	[0.00]	(10.75)	[0.00]
Exam 2014	64.44	69.54	5.10***	69.68	-0.15	64.71	0.269	63.65	1.05***
	(12.20)	(11.82)	[0.00]	(14.10)	[0.57]	(11.28)	[0.11]	(12.40)	[0.00]
Exam 2015	58.68	69.65	10.97***	71.75	-2.10***	62.06	3.38***	61.19	0.87***
	(12.35)	(12.47)	[0.00]	(13.84)	[0.00]	(12.31)	[0.00]	(12.24)	[0.00]
Integrity 2015	65.62	77.49	11.87***	80.39	-2.91***	75.19	9.57***	74.81	0.38**
	(17.91)	(10.59)	[0.00]	(11.99)	[0.00]	(11.70)	[0.00]	(10.41)	[0.02]
Observations	10,705	856	11,561	10,705	11,561	9,377	20,082	10,705	20,082

	(1)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
		2018 Cohort				2019 Cohort			
	Control	Treatment	Diff 1-10	Weighted Control	Diff 10-12	Treatment	Diff 1-14	Weighted Control	Diff 14-16
Exam 2010	69.98	74.06	4.07***	74.08	-0.02	72.46	2.48***	72.64	-0.18*
	(6.23)	(6.40)	[0.00]	(6.52)	[0.81]	(6.27)	[0.00]	(6.42)	[0.07]
Exam 2011	69.83	74.36	4.53***	73.98	0.39***	72.86	3.03***	72.90	-0.04
	(9.36)	(9.08)	[0.00]	(8.34)	[0.00]	(8.88)	[0.00]	(8.54)	[0.78]
Exam 2012	71.41	75.95	4.54***	75.52	0.44***	73.67	2.26***	73.97	-0.30**
	(10.04)	(10.23)	[0.00]	(8.89)	[0.00]	(10.27)	[0.00]	(9.40)	[0.03]
Exam 2013	58.52	62.10	3.58***	60.94	1.16***	59.05	0.53***	58.95	0.11
	(10.81)	(12.09)	[0.00]	(10.91)	[0.00]	(10.48)	[0.00]	(10.91)	[0.48]
Exam 2014	64.44	66.26	1.82***	65.75	0.51***	63.79	-0.65***	63.74	0.06
	(12.20)	(12.77)	[0.00]	(12.43)	[0.00]	(12.36)	[0.00]	(12.52)	[0.74]
Exam 2015	58.68	62.31	3.63***	61.90	0.40**	59.55	0.87***	59.47	0.09
	(12.35)	(13.50)	[0.00]	(12.62)	[0.01]	(12.49)	[0.00]	(12.49)	[0.60]
Integrity 2015	65.62	67.37	1.75***	68.81	-1.45***	68.23	2.61***	68.90	-0.68***
	(17.91)	(16.88)	[0.00]	(15.27)	[0.00]	(15.98)	[0.00]	(15.43)	[0.00]
Observations	10,705	16,183	26,888	10,705	26,888	12,963	23,668	10,705	23,668

Note: Standard deviations between parentheses and p-values between brackets. * p<0.10 ** p<0.05 *** p<0.01

Table A3: Impact Estimation Result for Exam Participants

Dependent Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exam Participants	2017 Cohort		2018 Cohort		2019 Cohort		Combined	
CBT -3					0.278	0.277	0.278	0.277
					(0.542)	(0.541)	(0.542)	(0.541)
CBT -2			-0.078	-0.072	0.578	0.698	0.214	0.271
			(0.605)	(0.604)	(0.714)	(0.698)	(0.550)	(0.542)
CBT -1	0.429	0.574	0.095	0.306	-0.432	0.157	-0.001	-0.321
	(0.980)	(0.953)	(0.933)	(0.916)	(0.969)	(0.874)	(0.655)	(0.611)
CBT 0	1.973	1.576	-1.103	-1.617	-0.314	-2.418	-0.089	-1.109
	(2.057)	(1.987)	(0.898)	(0.973)*	(0.906)	(1.194)**	(0.811)	(0.929)
CBT 1	-0.038	-1.415	-1.163	-2.324			-0.750	-1.990
	(1.152)	(1.509)	(0.981)	(1.211)*			(0.968)	(1.221)
CBT 2	-0.454	-3.614					-0.454	-3.614
	(1.379)	(2.181)*					(1.379)	(2.181)*
Fraction CBT in District		-6.353		-2.405		-4.429		-4.047
× (1-T)		(2.364)***		(1.201)**		(1.227)***		(1.266)***
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	100,410	100,410	134,440	134,440	118,340	118,340	353,190	353,190
Number of Schools	20,082	20,082	26,888	26,888	23,668	23,668	70,638	70,638
R^2	0.005	0.006	0.003	0.003	0.003	0.004		
Control Mean		81.3		78.5		71.1		76.6
CBT 0 (weighted)		(89.9)		(90.6)		(80.2)		(87.1)
Control Mean		85.5		76.6				79.8
CBT 1 (weighted)		(96.8)		(90.6)				(91.8)
Control Mean		83.7						83.7
CBT 2 (weighted)		(95.8)						(95.8)

Note: Standard errors between parentheses and corrected for clustering that the district level. The 'combined' columns show the sample-weighted average effect across cohorts. * p<0.10 ** p<0.05 *** p<0.01

Table A4: Impact Estimation Result for Raw Exam Scores

Dependent Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exam Score	2017 Cohort		2018 Cohort		2019 Cohort		Combined	
CBT -3					0.592	0.592	0.592	0.592
					(0.629)	(0.629)	(0.629)	(0.629)
CBT -2			1.169	1.178	0.154	0.172	0.718	0.732
			(0.856)	(0.855)	(0.734)	(0.731)	(0.785)	(0.784)
CBT -1	-0.244	-0.200	-0.164	0.237	-1.211	-1.119	-0.536	-0.3261
	(1.303)	(1.305)	(0.849)	(1.034)	(0.885)	(0.899)	(0.845)	(0.872)
CBT 0	-3.766	-3.885	-5.836	-6.813	-5.360	-5.688	-5.172	-5.722
	(1.368)***	(1.351)***	(1.320)***	(1.342)***	(1.058)***	(1.220)***	(1.054)***	(1.102)***
CBT 1	-0.700	-1.114	-5.041	-7.246			-3.448	-4.997
	(1.420)	(1.453)	(1.338)***	(1.537)***			(1.272)***	(1.394)***
CBT 2	0.272	-0.679					0.272	-0.679
	(1.512)	(1.670)					(1.512)	(1.670)
Fraction CBT in District		-1.912		-4.568		-0.691		-2.617
× (1-T)		(2.317)		(1.960)**		(1.588)		(1.804)
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	100,410	100,410	134,440	134,440	118,340	118,340	353,190	353,190
Number of Schools	20,082	20,082	26,888	26,888	23,668	23,668	70,638	70,638
R^2	0.297	0.297	0.345	0.347	0.269	0.269		
Control Mean	53.7		50.9		51.0		51.6	
CBT 0 (weighted)	(11.2)		(11.9)		(11.2)		(11.6)	
Control Mean	51.0		51.6				51.4	
CBT 1 (weighted)	(11.7)		(11.4)				(11.5)	
Control Mean	51.7						51.7	
CBT 2 (weighted)	(11.1)						(11.1)	

Note: Standard errors between parentheses and corrected for clustering that the district level. The 'combined' columns show the sample-weighted average effect across cohorts. * p<0.10 ** p<0.05 *** p<0.01

Table A5: Impact Estimation Result for Standardized Exam Scores

Dependent Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standardized Exam Score	2017 Cohort		2018 Cohort		2019 Cohort		Combined	
CBT -3					0.042	0.042	0.042	0.042
					(0.045)	(0.045)	(0.045)	(0.045)
CBT -2			0.081	0.082	0.013	0.015	0.051	0.053
			(0.059)	(0.059)	(0.055)	(0.055)	(0.056)	(0.056)
CBT -1	-0.015	-0.011	-0.009	0.024	-0.092	-0.081	-0.038	-0.020
	(0.090)	(0.090)	(0.072)	(0.074)	(0.064)	(0.065)	(0.060)	(0.062)
CBT 0	-0.286	-0.297	-0.435	-0.516	-0.427	-0.467	-0.396	-0.446
	(0.098)***	(0.097)***	(0.093)***	(0.095)***	(0.079)***	(0.092)***	(0.076)***	(0.080)***
CBT 1	-0.047	-0.084	-0.394	-0.578			-0.267	-0.397
	(0.100)	(0.103)	(0.096)***	(0.111)***			(0.091)***	(0.101)***
CBT 2	0.031	-0.053					-0.031	-0.053
	(0.110)	(0.121)					(0.110)	(0.121)
Fraction CBT in District		-0.170		-0.382		-0.085		-0.230
× (1-T)		(0.175)		(0.145)***		(0.121)		(0.135)*
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	100,410	100,410	134,440	134,440	118,340	118,340	353,190	353,190
Number of Schools	20,082	20,082	26,888	26,888	23,668	23,668	70,638	70,638
R^2	0.037	0.038	0.069	0.072	0.050	0.050		
Control Mean	-0.14		-0.06		-0.03		-0.07	
CBT 0 (weighted)	(0.88)		(0.89)		(0.89)		(0.89)	
Control Mean	-0.14		-0.03				-0.07	
CBT 1 (weighted)	(0.86)		(0.90)				(0.89)	
Control Mean	-0.11						-0.11	
CBT 2 (weighted)	(0.88)						(0.88)	

Note: Standard errors between parentheses and corrected for clustering that the district level. Outcome is standardized using the student-level control group mean and standard deviation in each year. The 'combined' columns show the sample-weighted average effect across cohorts. * p<0.10 ** p<0.05 *** p<0.01

Table A6: Heterogeneous Impact Estimation Result for Raw Exam Scores

Dependent Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exam Score	2017 Cohort		2018 Cohort		2019 Cohort		Combined	
	Integrity < 70	Integrity >= 70	Integrity < 70	Integrity >= 70	Integrity < 70	Integrity >= 70	Integrity < 70	Integrity >= 70
CBT -3					-0.711	-1.091	-0.711	-1.091
					(1.260)	(0.901)	(1.260)	(0.901)
CBT -2			1.140	0.158	-2.101	-0.057	-0.445	0.058
			(1.091)	(1.113)	(1.081)*	(0.904)	(1.033)	(0.922)
CBT -1	0.029	-1.818	-2.493	0.823	-5.183	-0.341	-3.082	-0.5181
	(1.490)	(1.520)	(1.168)**	(1.153)	(1.583)***	(1.211)	(1.088)***	(1.054)
CBT 0	-10.177	-3.085	-7.813	-0.773	-9.415	-1.898	-8.900	-1.978
	(1.256)***	(1.624)*	(1.748)***	(1.253)	(2.015)***	(1.265)	(1.595)***	(1.185)*
CBT 1	-4.826	-0.040	-6.968	-0.497			-6.296	-0.253
	(1.899)**	(1.743)	(1.516)***	(1.319)			(1.580)***	(1.442)
CBT 2	-3.890	0.837					-3.890	0.837
	(1.800)**	(1.835)					(1.800)**	(1.835)
CBT -3 × No Computers					-0.901	-0.824	-0.901	-0.824
					(1.078)	(0.479)*	(1.078)*	(0.479)
CBT -2 × No Computers			-0.749	-0.223	0.323	-0.989	-0.225	-0.582
			(0.816)	(0.434)	(1.004)	(0.471)**	(0.806)	(0.363)
CBT -1 × No Computers	-2.732	1.240	0.807	-0.338	1.972	-1.216	0.598	0.007
	(1.135)**	(0.645)*	(0.862)	(0.445)	(1.035)*	(0.623)*	(0.668)	(0.364)
CBT 0 × No Computers	-2.844	1.075	-3.141	-1.824	0.012	-1.234	-1.835	-0.551
	(0.663)***	(0.594)*	(0.805)***	(0.647)***	(1.415)	(0.435)***	(0.801)**	(0.377)
CBT 1 × No Computers	-3.004	0.907	-3.238	-1.701			-3.164	-0.305
	(0.680)***	(0.553)	(0.806)***	(0.623)***			(0.748)***	(0.453)
CBT 2 × No Computers	-2.470	0.864					-2.470	0.864
	(0.667)***	(0.563)					(0.667)***	(0.563)
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	26,185	53,225	33,205	49,490	32,650	46,515	92,040	149,230
Number of Schools	5237	10,645	6641	9898	6530	9303	18,408	29,846
R^2	0.609	0.226	0.550	0.167	0.473	0.122		
Control Mean	59.2	53.2	54.5	48.4	54.4	49.2	55.1	50.1
CBT 0 (weighted)	(13.0)	(10.4)	(12.5)	(10.5)	(12.5)	(10.0)	(12.7)	(10.5)
Control Mean	53.8	50.7	55.1	49.5			54.9	50.0
CBT 1 (weighted)	(12.0)	(11.2)	(12.9)	(10.1)			(12.8)	(10.6)
Control Mean	54.2	51.5					54.2	51.5
CBT 2 (weighted)	(12.8)	(10.7)					(12.8)	(10.7)

Note: Table only includes schools for which the integrity index and computer information is available in 2015. Standard errors between parentheses and corrected for clustering that the district level. The 'combined' columns show the sample-weighted average effect across cohorts. * p<0.10 ** p<0.05 *** p<0.01

Table A7: Impact Estimation Result for Exam Score Standard Deviation Within Schools

Dependent Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
S.D. Within School	2017 Cohort		2018 Cohort		2019 Cohort		Combined	
CBT -3					0.048	0.048	0.048	0.048
					(0.076)	(0.076)	(0.076)	(0.076)
CBT -2			0.001	0.002	-0.092	-0.067	-0.04	-0.29
			(0.111)	(0.112)	(0.092)	(0.095)	(0.097)	(0.099)
CBT -1	0.305	0.308	-0.174	-0.128	-0.065	0.055	-0.021	0.040
	(0.226)	(0.231)	(0.129)	(0.144)	(0.079)	(0.077)	(0.100)	(0.113)
CBT 0	0.657	0.648	0.462	0.351			0.533	0.460
	(0.247)***	(0.236)***	(0.134)***	(0.148)**			(0.135)***	(0.134)***
CBT 1	0.191	0.160					0.191	0.160
	(0.204)	(0.265)					(0.204)	(0.265)
Fraction CBT in District		-0.144		-0.517		-0.906		-0.557
× (1-T)		(0.717)		(0.295)*		(0.332)***		(0.350)**
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	80,328	80,328	107,552	107,552	94,672	94,672	282,552	282,552
Number of Schools	20,082	20,082	26,888	26,888	23,668	23,668	70,638	70,638
R^2	0.018	0.018	0.047	0.047	0.027	0.030		
Control Mean	5.6		5.9				5.8	
CBT 0 (weighted)	(2.4)		(2.1)				(2.2)	
Control Mean	6.1						6.1	
CBT 1 (weighted)	(2.2)						(2.2)	

Note: Standard errors between parentheses and corrected for clustering that the district level. The 'combined' columns show the sample-weighted average effect across cohorts. * p<0.10 ** p<0.05 *** p<0.01