Rotten apples or just apples and pears? Understanding patterns consistent with cheating in international test data

MARTIN GUSTAFSSON
CAROL NUGA DELIWE

Stellenbosch Economic Working Papers: WP17/2017

www.ekon.sun.ac.za/wpapers/2017/wp172017

December 2017

KEYWORDS: SACMEQ, TIMSS, assessment data, cheating, corruption, gender
JEL: C89, D73, I21

ReSEP (Research on Socio-Economic Policy)
http://resep.sun.ac.za

DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH

SOUTH AFRICA

UNIVERSITEIT STELLENBOSCH UNIVERSITY

BER
BUREAU FOR ECONOMIC RESEARCH

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

www.ekon.sun.ac.za/wpapers

Rotten apples or just apples and pears?

Understanding patterns consistent with cheating in international test data

MARTIN GUSTAFSSON AND CAROL NUGA DELIWE

JANUARY 2018

ABSTRACT

The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) programme has succeeded in generating valuable knowledge about the outcomes of schooling in the region, and in developing capacity to use data, including test data, in governments and amongst researchers. However, there is room for improvements in the programme. The current paper examines the extent to which patterns in the 2000 and 2007 test data suggest cheating occurred. The risk of cheating during the administration of the SACMEQ tests clearly exists, both because in-built controls can be subverted and because all pupils write exactly the same test, which is unlike the situation in a programme such as TIMSS, which employs a matrix sampling test design approach. Data analysis methods developed by Jacob and Levitt (2003) to detect cheating are adapted and then applied to the SACMEQ, but also TIMSS, data. It is concluded that whilst cheating does not substantially change the overall picture of performance derived from the 2000 and 2007 data, or country rankings, noteworthy patterns highly consistent with cheating can be found in some countries, and some regions within countries. Country-level indicators of cheating in SACMEQ correlate remarkably well with World Bank indicators of general corruption. An analysis of conditional correlations within the SACMEQ data reveals that schools serving more socio-economically disadvantaged pupils are more likely to cheat. In one country, having a male school principal is associated with a higher likelihood of cheating.

Martin Gustafsson
Department of Economics
University of Stellenbosch
Private bag X1, 7602
Matieland, South Africa
E-mail: mgustafsson@sun.ac.za

Carol Nuga Deliwe
Department of Basic Education
Private bag X895, 0001
Pretoria, South Africa
E-mail: nuga.c@dbe.gov.za

Both authors are based at the Department of Basic Education in Pretoria. We appreciate valuable exchanges and conversations with various people who have worked with SACMEQ data, including Qetelo Moloi, Servaas van der Berg, Toziba Masalila, Linda Zuze and Nic Spaull.

1 Introduction

The Sustainable Development Goals have accelerated an already growing interest in standardised testing, including international testing systems. There are compelling arguments for stakeholders to take more care in the generation of test data. In part, the challenge is to reduce the 'black box' effect whereby aggregate test scores are accepted uncritically, and without an appreciation of the inevitable risks in the data generation process.

We take up this challenge by focussing on one aspect of the data generation process: the risk of cheating during the administration of tests. Specifically, we adapt and then apply approaches developed by Jacob and Levitt (2003) aimed at detecting test data patterns which are suggestive of cheating. Their work focussed on the data of Chicago's public schools. We analyse data mainly from the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), for the years 2007 and 2000, but also Trends in International Mathematics and Science Study (TIMSS) data for comparison purposes. To our knowledge, analysis aimed at detecting patterns suggestive of cheating using these international datasets has not been undertaken previously.

Section 2 describes past attempts at using data analysis to detect the likely prevalence of cheating across schools. Jacob and Levitt's pioneering work, which informed their book *Freakonomics* (which in turn was used for a film with the same name), is discussed. Moreover, we discuss work undertaken in Italy to detect cheating, and how this has been used to bring about official score adjustments in that country's national testing system. We emphasise what has been emphasised in past attempts at this work: at best one can detect different probabilities of cheating having occurred. One is not dealing with an exact science. Given the nature of test data, it is impossible to draw an exact dividing line between cheating and what we refer to as innocent 'exceptional specialisation', a phenomenon which can lead to similar patterns in the data.

Section 3 describes the SACMEQ programme, an important programme which has deepened the understanding of educational quality in primary schools in eastern and southern Africa, but which still displays a number of design flaws. This section also introduces the SACMEQ data.

Section 4 presents the main body of our article. We begin by explaining a school-level indicator of possible cheating which we find particularly useful, and which draws from Jacob and Levitt's methods. We then calculate indicator values for schools in SACMEQ countries, using mathematics and reading scores separately, and a combination of the two, and examine differences across countries, and across regions within countries, in the distribution of these indicator values. We moreover use indicators calculated from TIMSS data, and explain why patterns from these data can be considered indicative of low levels of cheating. Noteworthy correlations between our indicator of possible cheating and widely used international indicators of general corruption (not corruption specific to education) are discussed. Our finding is that our indicator is useful and that patterns are highly suggestive of cheating in some countries and regions, but also suggestive of a 'clean' test administration process in certain countries, in particular Botswana, Swaziland and Mauritius.

Section 5 presents a method to adjust the SACMEQ data in order to counteract the effects of suspected cheating. Importantly, it is concluded that cheating did not distort country-level average scores to a large degree. Country rankings remain virtually the same after adjustments. Whilst cheating in SACMEQ warrants serious attention, the valuable picture of educational quality over time offered by SACMEQ remains valid.

Section 6 explores correlations between suspected cheating, on the one hand, and school and home background variables, on the other. A key finding is that within countries, schools

serving more socio-economically disadvantaged pupils are more likely to display patterns in the data consistent with cheating.

Section 7 concludes.

2 Past attempts at detecting suspicious patterns in the data

A seminal text on detecting patterns consistent with cheating in test data is that of Jacob and Levitt (2003). They used data from the city of Chicago's extensive testing system, the aim being largely to warn authorities of the dangers of judging schools on the basis of data which display clear signs of cheating. Data from tested grades 3 to 8 pupils spread across approximately 1,000 classes were used. Crucially, Jacob and Levitt had data from more than one point in time for the same pupil. They were also given permission to re-test selected pupils in a controlled environment. Two indicators were devised, one for 'unexpected test score fluctuations' and another for 'suspicious answer strings'. The first focussed on strange discrepancies in the aggregate score for the same pupil across different tests in the same subject. The second indicator drew from four 'measures' focussing on discrepancies at the level of items, or questions. The first three of these measures used as statistical controls past or future results for the same pupil. The fourth measure did not include these controls, and was furthermore meaningful in terms of the kind of analysis we wished to undertake. We explain 'measure 4', the only Jacob and Levitt statistic we replicate (with some adaptations), in section 3 below. Clearly, not having panel data with results for the same pupil in the same subject at different points in time severely limited the extent to which we could replicate Jacob and Levitt's work. Yet, as we hope to demonstrate, what we were able to replicate reveals important patterns.

Jacob and Levitt concluded that in at least 4 to 5 per cent of Chicago's schools cheating had occurred. The kinds of cheating they consider include illicit teacher access to the test before test administration, the sharing of correct answers whilst the test is being administered, and the manipulation of answer sheets by teachers after the test to raise test scores. Jacob and Levitt find that cheating worsened when testing was linked to high-stakes teacher accountability programmes. The policy implication is clearly that high-stakes accountability should not rely on tests which can be manipulated by teachers.

The Italian national testing programme has been the basis for further pioneering work on detecting data patterns consistent with cheating. This programme, Invalsi, includes its own official methodology for detecting cheating, which is linked to a scores adjustment approach intended to remove the effects of cheating. These adjustments, as one might expect, have been controversial as they carry the risk that schools which did not cheat will erroneously be classified as cheaters, and thus experience a downward adjustment. The effects of cheating in Invalsi are glaring. Poorer regions in the south of Italy have fared better than richer regions in the north, according to unadjusted Invalsi data, although the international testing programmes TIMSS and PISA have pointed to the opposite. This is because the international programmes offer fewer opportunities for cheating. Battistin, De Nadai and Vuri's (2017) use their own analysis of Invalsi data to conclude that in the south of Italy cheating occurs in at least 15% of schools, whilst in the north it is almost absent. Ferrer-Esteban (2013) uses Invalsi plus other data, including local tax avoidance data, to explore the sociological aspects of cheating involving teachers. He concludes that cheating is more common in homogenous and socially disadvantaged communities, where teachers will often prioritise assistance to pupils whom they identify closely with, above compliance with rules set by distant authorities. Ferrer-Esteban, like the current paper, draws from the methods of Jacob and Levitt (2003).

3 Background on the SACMEQ programme

The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) was started in 1991 to generate comparable learning outcomes data across southern and eastern Africa, and to develop capacity in testing and educational research generally in the region. The programme was largely designed by Unesco's International Institute for Educational Planning (IIEP), which worked closely with fifteen African governments. In preparation for the 2013 cycle of testing, management of SACMEQ moved from the IIEP in Paris to a SACMEQ office based at the University of Botswana. In the years 1995, 2000, 2007 and 2013 testing has occurred in samples of grade 6 pupils in fifteen countries, though not every country has participated in every year. Tanzania has been treated as two countries: mainland Tanzania and Zanzibar. The 2007 round of the programme, the most recent for which data are widely available, tested around 61,000 pupils in around 2,800 schools in all fifteen countries. Apart from testing pupils in mathematics and reading, background questionnaire data were collected from pupils, their teachers and school principals. Crucially, teachers have also been tested.

The SACMEQ data have been used by government and non-government researchers within SACMEQ countries to produce national reports, many of which are available on the SACMEQ website[1]. A few articles using SACMEQ data, some with a cross-country focus, have been published in peer-reviewed research journals. If one examines the work undertaken using SACMEQ data, and discusses SACMEQ with education planners in the region, it becomes clear that SACMEQ has played a crucial role in advancing our understanding of the cognitive skills of pupils and promoting the use of data in government and by researchers.

For the current article, a central question is how loopholes in the procedures for administering the SACMEQ tests and the structure of the tests might facilitate cheating. The 2000 administration process is better documented than that for other years, though discussions with people involved in the programme indicate that the process would be similar across all years. In 2000, 15 to 50 'data collectors' were trained per country. These people had the responsibility to administer the tests to pupils, a process for which a detailed manual for use in all countries was developed[2]. These data collectors were officials of the education authorities. Different countries employed different strategies to ensure that officials complied with the rules. Zambia's approach is noteworthy, and appears to have been particularly rigorous. In this country, officials were deployed to localities where they did not usually work, in order to reduce the risk that personal relationships between the officials and school staff would compromise the process.

Though it was emphasised in SACMEQ that the identify of schools would remain anonymous in the programme's reporting procedures, and that there would be no negative consequences for individual schools arising out of poor test results, it is understandable that school staff would be concerned that these assurances might not be upheld. It should be remembered that virtually all tested schools would have not experienced SACMEQ testing previously, and that examinations are routinely used to judge the effectiveness of individual schools in most countries.

Two SACMEQ countries, South Africa and Botswana, have participated in TIMSS. The differences between these two countries with respect to their TIMSS processes, and the differences between SACMEQ and TIMSS processes, are telling. In Botswana, tests were administered by teachers in the schools in which they normally worked, with extensive training and monitoring of teachers being the responsibility of external people. In South

[1] http://www.sacmeq.org.

[2] Available at catalog.ihsn.org/index.php/catalog/4574/download/78024 (accessed November 2017). Title is *Southern Africa Consortium for Monitoring Educational Quality: Manual for data collectors.*

Africa, on the other hand, test administration was the responsibility of employees of an independent agency, who had no past connection to the school. This approach was followed in South Africa to enhance the integrity of test administration. TIMSS allows countries to follow different approaches, as long as common minimum standards are adhered to. A key feature of TIMSS is the International Quality Assurance Program, which involves the appointment of a totally independent quality assurer, who visits fifteen tested schools per country[3]. This level of oversight seems absent in SACMEQ[4].

To sum up, with regard to test administrators, current practices suggest that there is nothing inherently wrong with test administrators who work in the tested school, or who are familiar with individual schools. What does seem of greater fundamental importance are quality assurance arrangements involving oversight by independent observers. The optimal approach in each country seems in part to depend on country-specific circumstances relating, for instance, to public trust in those working in the schooling system. What is optimal is not a one-dimensional matter. Higher levels of control will often come with greater costs, and could make the testing prohibitively expensive for specific countries. SACMEQ clearly faces difficult budget constraints. Even in TIMSS, it is accepted that in some countries deviations from minimum standards, due to budget constraints, must sometimes be permitted[5].

Perhaps the weakest link in the SACMEQ programme, in terms of opportunities to cheat, is the fact that SACMEQ tests do not follow a matrix sampling approach, as is the case in TIMSS. In SACMEQ, every pupil in the testing venue writes exactly the same test, whilst in TIMSS different pupils write different combinations of questions. This clearly makes it easier in SACMEQ for the person responsible for administering the test to assist pupils, or for pupils to assist each other. To illustrate the TIMSS matrix sampling approach, in South Africa in 2015, the grade 8 mathematics tests comprised 14 different versions, and on average only three pupils in a test venue with 42 pupils wrote exactly the same mathematics test. Matrix sampling is employed mainly to widen the range of questions covered in the programme as a whole, but an additional benefit would be a lower risk of cheating. Clearly, matrix sampling comes with increased complexity and this seems to be a key reason why this approach has been avoided in SACMEQ. Yet it is worth noting that another programme focussing on a group of developing countries, namely Latin American Laboratory for Assessment of the Quality of Education (LLECE, following the Spanish name), uses matrix sampling in the design of its tests[6].

Marking of SACMEQ tests occurred at a central venue in each SACMEQ country, meaning that cheating during marking by school teachers who know the pupils would be virtually impossible.

One serious drawback with SACMEQ is that technical documentation on the data collection methods in general, and the calculation of the widely publicised final SACMEQ scores, is far less available than it should be. This is especially so for the 2007 run of the programme, a run which was critical insofar as it was the first to employ anchor items to equate the test score scale of 2007 to the earlier 2000 scale. Even stakeholders intimately involved in the programme, such as SACMEQ organisers in national ministries of education, lack access to the kinds of technical documentation one would find published on the web in the case of TIMSS or even LLECE. One can conclude that critical elements of the data collection and generation processes were not sufficiently documented. This does not mean that these processes are necessarily faulty, but it does negatively impact on the credibility of SACMEQ,

[3] Martin, Mullis and Hooper, eds., 2016: 9.2.
[4] Ross *et al*, 2008.
[5] Martin, Mullis and Hooper, eds., 2016: 6.19.
[6] Solano-Flores and Bonk, 2008.

and makes interpreting the data more difficult. These are problems which should be resolved as SACMEQ evolves.

One strange feature of the SACMEQ scores is a one-to-one relationship between classical scores and the final SACMEQ scores calculated using item response theory (IRT) methods. Within a country (and frequently across many countries), pupils with the same classical score would be given the same final SACMEQ score. To illustrate, in 2007 mathematics, in all countries other than Tanzania (mainland) and Mozambique, any pupil who obtained 26 (out of 45) questions correct, was given a SACMEQ IRT score of 610. This type of one-to-one relationship was found in both mathematics and reading, in 2000 and 2007. This is strange because a key purpose of IRT scoring is to differentiate between pupils with the same classical score, depending on the level of difficulty of the questions they got right[7].

The following table sums up the dimensions of the data used for our paper. Items were excluded from the analysis if the available documentation or the data indicated that an item was excluded from the calculation of the IRT score. This would occur because those who processed the data decided that an item was not sufficiently consistent with the general patterns in the data. This is normal practice. For the purposes of our paper, an item would be dropped from the data of all countries, even if in terms of the calculation of the IRT score it had been dropped from just some countries. This was to enhance the across-country comparability of our statistics. Where cells indicating a pupil's score were blank, it was assumed that the value should be zero, meaning the pupil did not get the question right. There were considerably more blank cells in the 2000 data than the 2007 data. In the end, every cell could contain just zero or one. All SACMEQ test questions are multiple-choice questions, so one means correct, zero means incorrect.

Table 1: Item exclusions and missing values in the SACMEQ data

	Total items	Items after exclusions	% cells with missing data (after exclusions)	Total pupils	Total schools
2000 reading	83	79	2.3	41,686	2,294
2000 mathematics	63	60	2.4	41,686	2,294
2007 reading	55	55	0.1	61,396	2,779
2007 mathematics	49	42	0.2	61,396	2,779

4 Assessing the utility of an 'exceptional specialisation' indicator

4.1 The logic of the indicator

Jacob and Levitt's measure 4, introduced above, is calculated according to the two steps illustrated in the first two equations below. Firstly, the score q for item i in the case of student s in class c is considered. In Jacob and Levitt, and in our SACMEQ analysis, q is either zero, for not correct, or one, for correct (in our TIMSS analysis, there are a few cases where q can assume the values 0, 1 or 2, depending on level of correctness). From q one subtracts $\overline{q}$, the mean of q for many students. The superscript A denotes that in calculating the mean of q, only students who obtained the same overall score as student s are considered. Each student's overall classical score would be the student's sum of q. The squared difference of the two terms for item i is found. Then the squared differences for all items in the case of student s are summed. This produces a value Z for each student. This value Z will be greater the more the student's item scores q differ from those of other students performing at the same level. It is this difference which is the basis for assessing how strange or suspicious a student's results are.

[7] This aspect of the SACMEQ data is discussed in more depth in Crouch and Gustafsson (2017).

Turning to the second equation, $\overline{Z}$ is the mean of *Z* across all students in the same overall score category *A*. For each student, the difference between her value *Z* and the corresponding $\overline{Z}$ is found. The greater this difference, the larger the exceptionality of the student's responses. All differences within class *c* are summed, producing *M* for each class, the desired indicator of exceptionality for the class, or school. For the purposes of all the discussion that follows, class and school can be considered synonymous. Even if students from the same school but different classes, but of the same grade, are combined for a test session, we can think of the students as being from the same class in terms of our topic of possible cheating. If students were assisted, they would have been assisted during a single test session, and could therefore be expected to display similar suspicious patterns.

Jacob and Levitt's measure 4, or *M* below, produces a mean of zero across all classes analysed, by design. If there are suspicious patterns, exceptionally high positive values for *M* would be indicative of institutions where cheating is likely to be occurring.

$$Z_{sc} = \sum_{i} \left(q_{isc} - \overline{q}_i^A \right)^2 \tag{1}$$

$$M_c = \sum_{s} \left(Z_{sc} - \overline{Z}^A \right) \tag{2}$$

Jacob and Levitt use the above method in order to rank classes according to the suspiciousness of their data within a specific test. They are not interested in the absolute value of *M*. For our analysis, however, we do want some indicator whose absolute values we can compare, for instance across different testing systems. The above equations clearly indicate that *M* will be inflated the more items there are in a test, and the more students there are in a class. In order to obtain an indicator that was not sensitive to these two variables, quantity of *i* and quantity of *s*, we adapted the above method to produce the following:

$$Z_{sc} = \frac{\sum_{i} \left(q_{isc} - \overline{q}_i^A \right)^2}{ni} \tag{3}$$

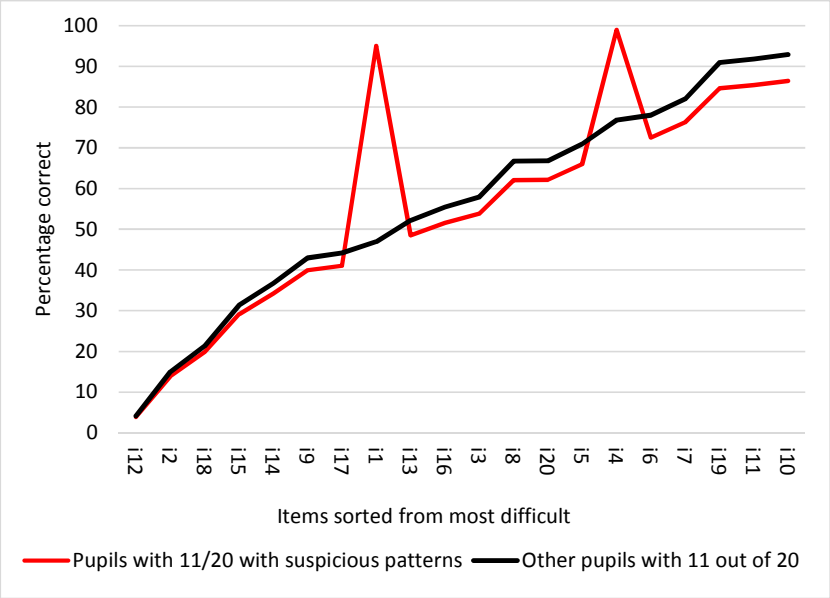
$$M_c = \frac{\sum_{s} \left(Z_{sc} - \overline{Z}^A \right)}{ns} \tag{4}$$

Equation (1) is adapted by dividing the sum of the differences by *ni*, being the number of items in the test. Similarly, equation (2) is adapted by dividing the right-hand side by *ns*, the number of students. The adapted indicator *M* also produces a mean of zero and the adapted values for *M* correlate perfectly with the original Jacob and Levitt *M* values if one uses a dataset where the number of students per class is exactly the same. Having more or fewer students, or more or fewer test items, will of itself not change the adjusted *M* values. This is the key advantage of the adjusted *M* values for our purposes. The Stata code used to implement our version of measure 4 is attached as an appendix.

Figure 1 below provides a stylised representation, using fictional data, of the meaning of our measure 4. It does not precisely depict the two equations, but it conveys the essential concepts. The red curve represents students from a school where cheating may have occurred. The students represented are those who obtained 11 correct out of 20 questions. If we compare these students to other students in our dataset who also obtained 11 out of 20, then it is clear that the 'red' students performed exceptionally well in items 1 and 4. The gap between the two curves indicates the exceptionality of the red students. We can refer to this

exceptionalism as 'exceptional specialisation'. The red students may have been assisted by their teacher during the test. Alternatively, the students may have recently focussed on the topics in items 1 and 4, meaning the exceptional specialisation is not a case of cheating.

Figure 1: Stylised representation of measure 4



Importantly, the fact that the 'red students' performed worse than 'black students' in items other than items 1 and 4 is a part of the suspiciousness of the patterns. We would not expect pupils with 11 out of 20 correct to perform so poorly in these other items. Hence our equations focus on both the positive and negative vertical differences between the two curves.

If within the same class, some students had displayed 'exceptional specialisation' in items 1 and 4, whilst other students had displayed this in, say, items 5 and 8, the value of *M* for the class is also likely to be elevated. This points to an important feature of the equations. They are not concerned with whether students in a class are displaying 'exceptional specialisation' in the *same* items. Practically, what this means is that the value of *M* could be raised by students helping each other, even if the teacher does not actively assist the whole class. Of course this would still be an indication of undesirable practices on the part of the teacher, as it is the teacher's responsibility to prevent cheating amongst students.

Much of the analysis that follows is aimed at establishing how much of the exceptional specialisation can be considered the result of honest specialisation, as opposed to cheating. Put differently, we consider the degree to which we are seeing the 'rotten apples' of Jacob and Levitt, and to what extent we are seeing innocent differences between 'apples and pears', where the difference is unrelated to cheating.

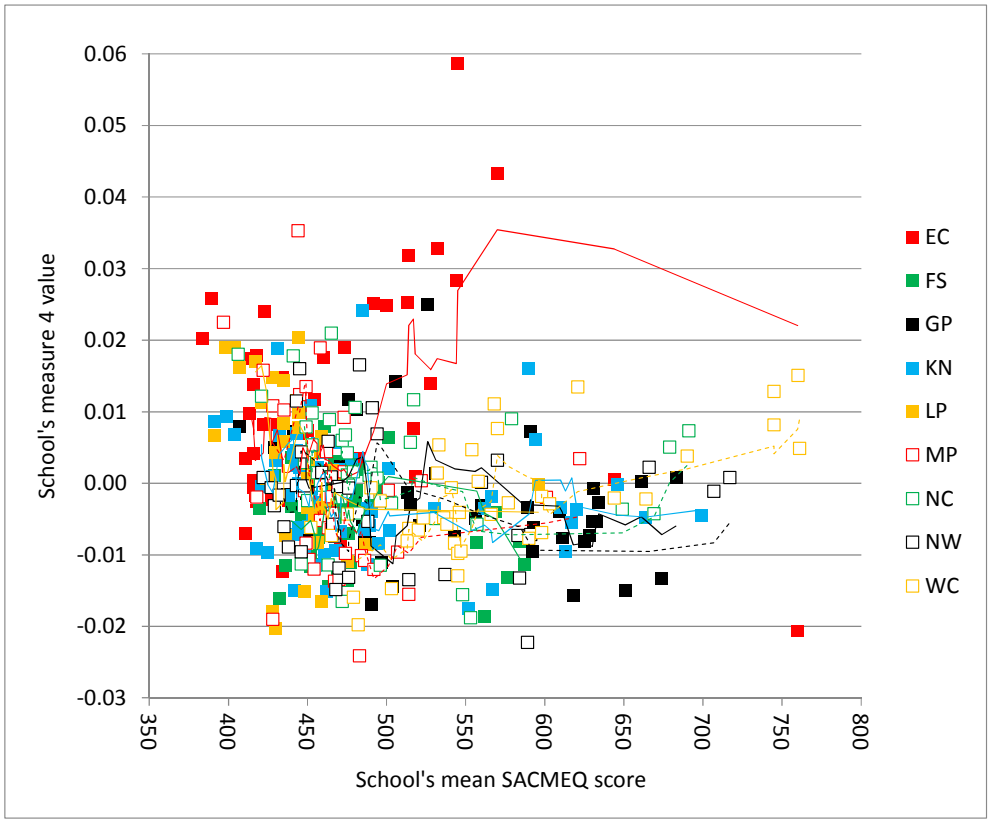
4.2 Indicator values from the SACMEQ (and TIMSS) data

Curricula and pedagogical traditions in SACMEQ countries tend to be specific to individual countries. One would expect the innocent and non-cheating version of the 'exceptional specialisation' discussed above to be driven by national dynamics. Some countries may promote more flexibility with respect to curricular focus than others, meaning one would expect higher levels of 'exceptional specialisation' in the form of specific classes having mastered specific topics exceptionally well. Comparing our measure 4 indicator values across regions within countries thus seemed like a good point of departure. One would expect regions across a country to be fairly uniform with respect to curriculum-related policies and traditions, yet regions in the same country might differ in terms of their acceptance of

cheating (as we have seen is the case in Italy). Even if policies are national, practices could differ by region because of differing governance trajectories. Moreover, regions within a country are likely to be accustomed to competing with each other, because national authorities emphasise inter-regional disparities. Competition might make regions accepting of illicit score inflation. Regional manipulation of enrolment data, to obtain a larger share of national funds, is fairly well documented[8], and some of this culture may permeate testing systems.

Focussing on across-region differences in our measure 4 values in fact emerged as a rather relevant point of departure. Figure 2 illustrates the measure 4 values of South Africa's schools using the 2007 SACMEQ mathematics data. Importantly, measure 4 values per school were calculated using only South Africa's data (an approach followed subsequently for all countries). The graphs use values from 392 schools, of which 50 are from Eastern Cape (EC). Clearly there is something different about the measure 4 values of this province. For several schools they are substantially above the values found in the school sample in general, and Eastern Cape is the only province where the moving average curve exceeds 0.03. These patterns would quite easily raise suspicions about cheating in Eastern Cape amongst those familiar with the South African schooling system. This province is often considered particularly poorly governed[9].

Figure 2: Measure 4 values in South Africa mathematics 2007 (I)



Note: Trendlines are moving averages across 5 points. Solid lines correspond to solidly coloured markers of the same colour, whilst dotted lines correspond to hollow markers.

Apart from examining measure 4 values for both SACMEQ subjects and both years, we also calculated a stricter measure 4, which was simply the minimum measure 4 value across the two subjects for the same school and the same year. This measure would be stricter insofar as

[8] Gustafsson, 2015.

[9] See for instance Wills, Shepherd and Kotze, 2016. Gustafsson (2015) found particularly clear signs of manipulation of Eastern Cape's enrolment data.

suspiciously high values would have to exist in both subjects for the school to stand out. A high value for both subjects would point to the possibility of a general culture of cheating in the school, as opposed to a single teacher assisting pupils during the test.

Table 2 presents a summary of the patterns found. Across all six rows Eastern Cape displays a considerably higher measure 4 value, at the 90th percentile, compared to the other eight provinces. The gap is particularly large in the case of mathematics. One might expect more cheating during a mathematics test, relative to a reading test, as correct answers are likely to be more easily identifiable, and easier to communicate in the test venue. What is noteworthy is that even against the stricter 'minimum measure 4' Eastern Cape's values emerge as suspicious. In fact, the 2007 figures suggest that in almost all Eastern Cape schools with high measure 4 values for reading, there were also high values in mathematics. The two tend to co-exist in the same school.

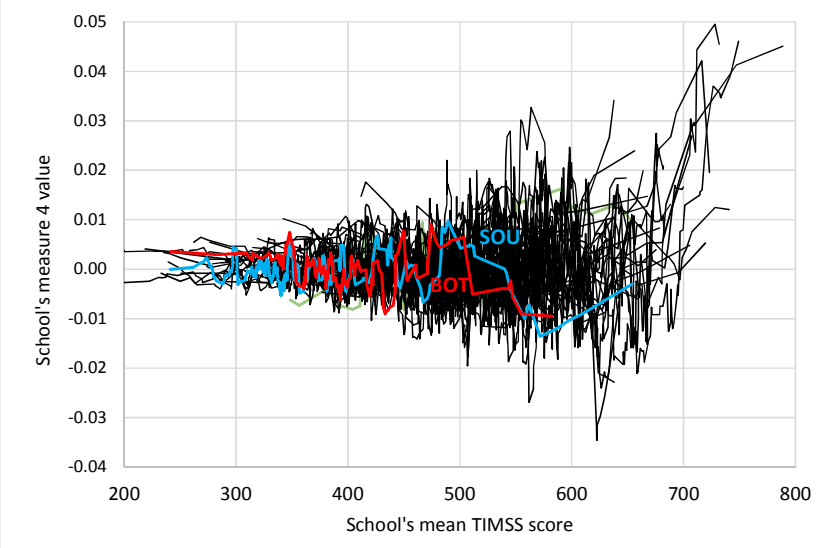
Table 2: Degrees of Eastern Cape unusualness

	Measure 4 90th p'tile		
	Eastern Cape	Other provinces	Difference
2007 mathematics	0.027	0.011	0.016
2007 reading	0.015	0.009	0.006
2007 minimum measure 4	0.014	0.005	0.009
2000 mathematics	0.032	0.012	0.020
2000 reading	0.016	0.007	0.009
2000 minimum measure 4	0.008	0.003	0.005

A key question is whether it is possible to find absolute benchmarks for 'clean' schools. Are the 'Other provinces' figures in Table 2 indicative of low levels of cheating? We used TIMSS data to explore this matter. One would expect TIMSS to be virtually free of cheating due to the programme's strong quality control mechanisms and its matrix sampling approach to test design[10]. Figure 3 below presents what we found using 2015 grade 8 TIMSS mathematics data. Each school's measure 4 value was calculated using just data from its own country. We analysed just students whose test was a combination of booklets 4 and 5 (each student had a mathematics test containing two standard booklets). This focus meant we analysed on average 4.8 students per school and 47 test items in total. The graph indicates that virtually no countries had schools with measure 4 values as high as those of Eastern Cape, although there were a few, especially at the high end of the performance spectrum.

[10] The only reference we found to the possibility of cheating in TIMSS is that described by Papanastasiou and Zembylas (2006), who conclude that the alleged cheating in the case of Cyprus is unlikely to have really occurred.

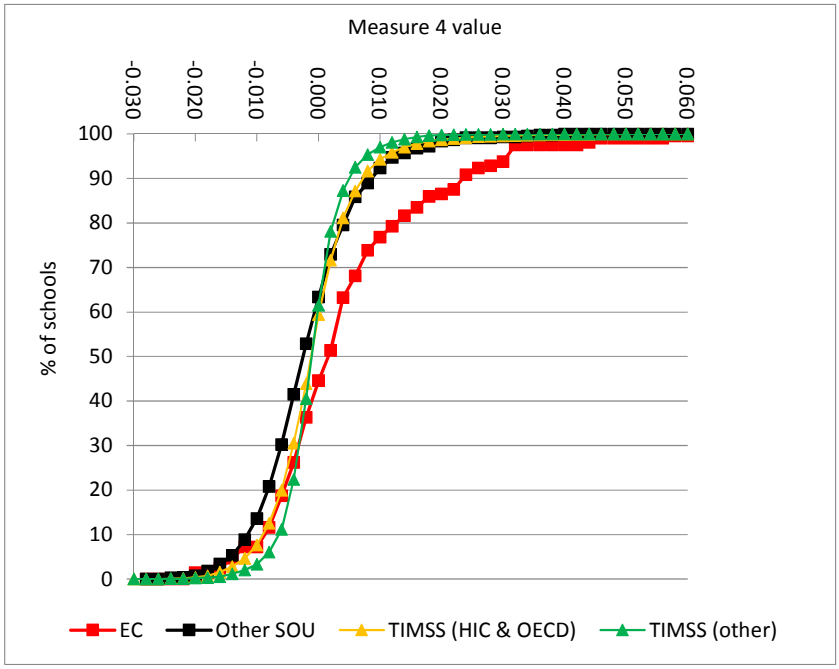
Figure 3: Measure 4 distribution in TIMSS 2015 Grade 8 mathematics



Source: TIMSS microdata from https://timssandpirls.bc.edu.
Note: Each country's curve is a moving average curve using averages across five schools (as in Figure 2). Red and blue curves are those of Botswana and South Africa.

The following graph is more useful in terms of establishing benchmarks. In Figure 4 the cumulative percentage of schools (not weighted by pupils) reaching specific measure 4 multiples of 0.002 is shown. Median schools have measure 4 values close to zero, something one would expect given that measure 4 always produces a mean of zero. However, there are large differences to the right of zero between Eastern Cape, on the one hand, and the other three curves. Indeed, it appears that South Africa's other eight provinces are relatively 'clean' insofar as they produce a curve very close to the TIMSS curves. Two TIMSS curves were produced, one for high income countries which are also OECD members, and one for other TIMSS countries, which we can think of as developing countries. This was done to check whether developing countries in TIMSS, which may not have the resources to implement all the required quality controls correctly, displayed higher measure 4 values. Clearly they do not. In fact, developing countries in TIMSS produce the 'cleanest' of the four curves, with the lowest prevalence of high measure 4 values.

Figure 4: South Africa SACMEQ and TIMSS international compared



The above graph can be used to propose a few absolute benchmarks. In looking at across-region disparities in all SACMEQ countries, we decided to consider a region as displaying suspicious patterns if (1) an individual subject carried a 90th percentile measure 4 value above 0.024 or the minimum for both subjects exceeded 0.012 *and* (2) if the 90th percentile measure for a region was higher than the corresponding value for other regions by a margin of at least 0.005.

The results of our analysis of across-region differences are presented in the next two tables. Table 3 sums up the details of Table 4. In Table 3, a specific region could be counted in one column, and again in another column (in data columns 3 to 8). The totals are thus not regions with high values, but region-specific instances of high values. It is noteworthy that three countries present no instances of regions with high values: Botswana, Mauritius and Swaziland. The probability of cheating concentrated in specific regions in these countries thus appears to be low. Seychelles also features no instances, but it should be noted that this country's 'sample' is very different to those of the other countries. In 2007 all 24 primary schools in the country were tested, compared to samples in all the other countries ranging from 139 in Malawi to 392 in South Africa (for all other countries the mean sample size was 197). The exceptional smallness of the Seychelles 'sample', and of Seychelles as a country, is likely to produce a different context with regard to cheating, and a lower probability of high measure 4 values.

Five countries display regions with a relatively high likelihood of cheating: Lesotho, Mozambique, South Africa, Tanzania and Uganda. These are the only countries where the total instances of high regional values comes to 0.4 or more of the total number of regions.

It is furthermore noteworthy that the measure 4 situation was roughly similar in 2000 and 2007, and that higher values in mathematics, relative to reading, is common.

Table 3: SACMEQ regions with unusual values

		Regions		2007			2000			
Abbr.	Country	2007	2000	Math.	Read.	Both	Math.	Read.	Both	Total
BOT	Botswana	7	7							0
KEN	Kenya	8	8				1		1	2
LES	Lesotho	10	10	2	1	1				4
MAL	Malawi	6	6				1			1
MAU	Mauritius	7	5							0
MOZ	Mozambique	11	11	2		3	1		1	7
NAM	Namibia	13	13	1						1
SEY	Seychelles	6	6							0
SOU	South Africa	9	9	1		2	2			5
SWA	Swaziland	4	4							0
TAN	Tanzania	11	11	2			4			6
UGA	Uganda	4	5	2			2	1	1	6
ZAM	Zambia	9	9	1		1				2
ZAN	Zanzibar	5	5				1			1
ZIM	Zimbabwe	10		1						1
Total		120	109	12	1	7	12	1	3	36

Note: Abbreviations for countries are those appearing in the SACMEQ data.

In Table 4 below, there are 20 regions listed, of which seven display instances of high measure 4 values in both years. Had high measure 4 values been distributed randomly across regions, only around 1.7 regions would emerge in both 2000 and 2007. This raises the likelihood that there is something systematically different about the listed regions when it comes to the way testing occurs.

The criterion that a noteworthy region should display a value 0.005 higher than that for other regions combined plays a relatively small role. If one removes this criterion, the overall total of 36 in Table 3 rises only slightly to 38.

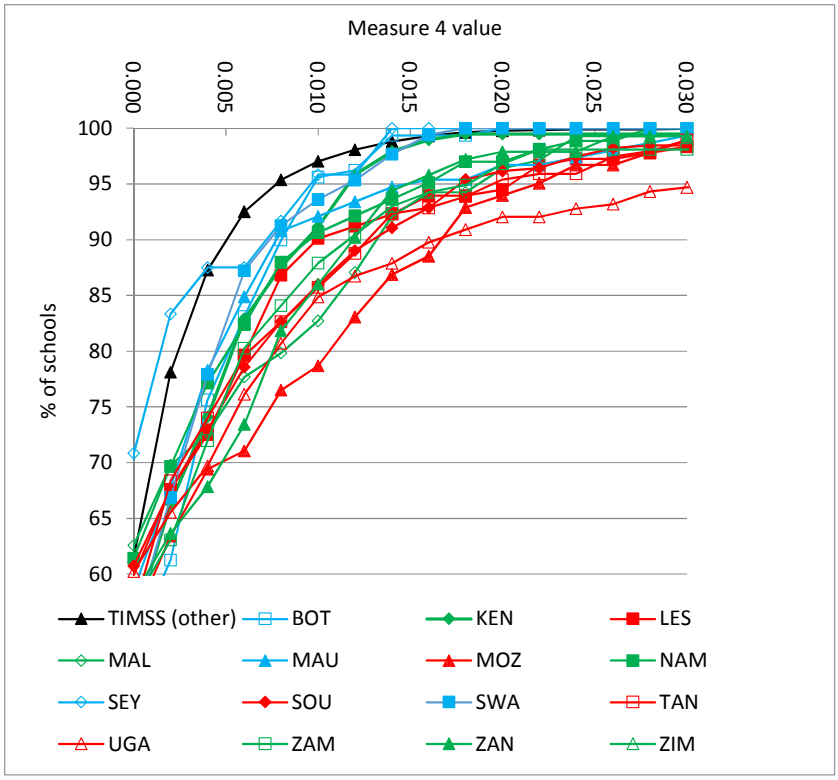
Table 4: SACMEQ regions and 90th percentile measure 4 values

Country and region abbr. in SACMEQ and derived full name of region			2007 90th percentile				2000 90th percentile			
			Schools	Math.	Read.	Both	Schools	Math.	Read.	Both
KEN	NOR	North Eastern					15	0.029		0.014
LES	BUT	Butha-Buthe	16	0.028	0.030	0.023				
LES	QNK	Qacha's Nek	15	0.031						
MAL	SEA	South East					21	0.045		
MOZ	CAB	Cabo Delgado	16	0.028		0.014				
MOZ	NIA	Niassa	16	0.028		0.013	15	0.025		
MOZ	TET	Tete	16			0.012	15			0.013
NAM	OMA	Omaheke	16	0.025						
SOU	ECA	Eastern Cape	50	0.027		0.014	27	0.032		
SOU	LMP/NPR	Limpopo	40			0.014	24	0.029		
TAN	KIL	Kilimandscharo	16	0.026			15	0.036		
TAN	NOE/NEA	North East	20	0.030			15	0.027		
TAN	NOR	North					20	0.038		
TAN	WES	West					20	0.028		
UGA	EAS	East	83	0.025			45	0.033	0.024	0.013
UGA	NOR	North	58	0.034						
UGA	WES	West					24	0.029		
ZAM	NOR	North	20	0.024		0.012				
ZAN	SOP	South Pemba					33	0.024		
ZIM	MVG	Masvingo	16	0.024						

Note: Abbreviations for regions are those appearing in the SACMEQ data. Full names of regions were derived by consulting separate lists of each country's subdivisions and comparing these to the around 120 region codes in SACMEQ. Where two three-letter codes for a region appear, this is the code in the 2007 data followed by the code in the 2000 data.

We now move to country-level distributions of the measure 4 values we have calculated. Figure 5 confirms what one might expect, namely that countries with outlier regions are also those countries with the greatest general within-country variation in their measure 4 values. Countries which emerged as relatively 'clean' in Table 3 above have blue curves in Figure 5, whilst red curves are used for regions with many instances of outlier regions. Clearly, the blue curves lie closer to the TIMSS developing country curve (which is reproduced from Figure 4).

Figure 5: Countries' distributions of measure 4 for 2007 mathematics



One finding which strengthens the argument that high measure 4 values reflect more cheating is a relatively high correlation, at the country-level, between cheating in schools and more general corruption. To obtain a measure of the latter, we made use of the World Bank's Worldwide Governance Indicators (WGI). These indicators are calculated using 'the views on the quality of governance provided by a large number of enterprise, citizen and expert survey respondents'[11]. The indicators cover six dimensions. We used only values from the dimension 'control of corruption' as this seemed the most pertinent. The other five dimensions refer to: voice and accountability; political stability and absence of violence; government effectiveness; regulatory quality; rule of law. We should emphasise that we selected the 'control of corruption' because of its pertinence, not because it yielded the highest correlation. We wanted to avoid a 'data fishing expedition'. In fact, we did not examine the data from the other five dimensions at all.

In calculating a country-level cheating-in-schools indicator from our measure 4 values we applied some Bayesian logic by combining new evidence and prior beliefs. We assumed that cheating in schools was a manifestation of a general culture of corruption and then asked ourselves what measure 4 threshold would result in the highest correlation with the 'control of corruption' values, for which we have both 2000 and 2007 values. Using ordinary least squares, we estimated b in equation (5). Here m is the percentage of schools in country i exceeding a threshold with respect to measure 4. We used different thresholds in the range

[11] http://info.worldbank.org/governance/wgi, accessed October 2017.

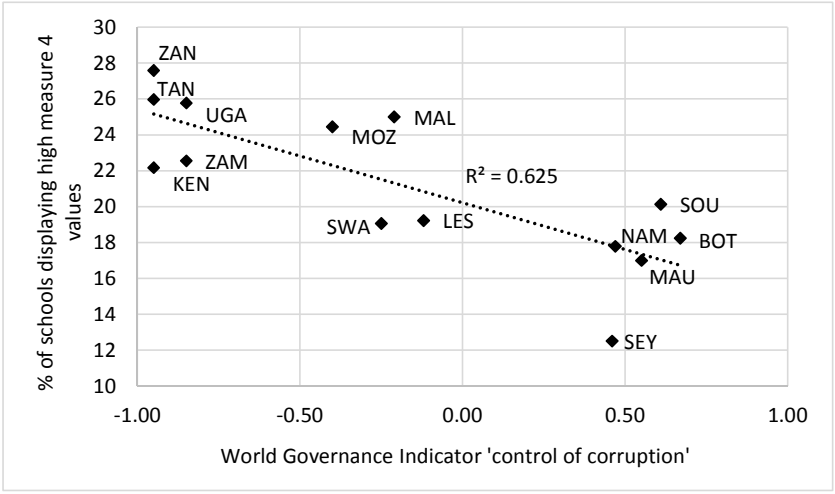
0.001 to 0.040 and selected the threshold which gave the lowest *p* value for *b*. In other words, we maximised the statistical significance of *b*. This would be permissible 'fishing' in the sense that were are not searching different data sources to approximate our hypothesis, but identifying a threshold within one set of data.

$$m_i = \hat{a} + \hat{b}W_i \tag{5}$$

The measure 4 thresholds we arrived at were rather low: 0.006 for mathematics in 2000 and 0.008 for mathematics in 2007. To illustrate the meaning of this, in the relatively 'clean' TIMSS developing country data of Figure 5, just 5% of schools displayed measure 4 values exceeding 0.008. As will be seen below, the percentages of schools in the SACMEQ data exceeding these thresholds are much higher than 5%. Obviously one reason why a relatively low threshold would be best is that at higher thresholds, beyond around 0.014 using the SACMEQ data, countries start dropping out of one's dataset as they have no schools exceeding the threshold.

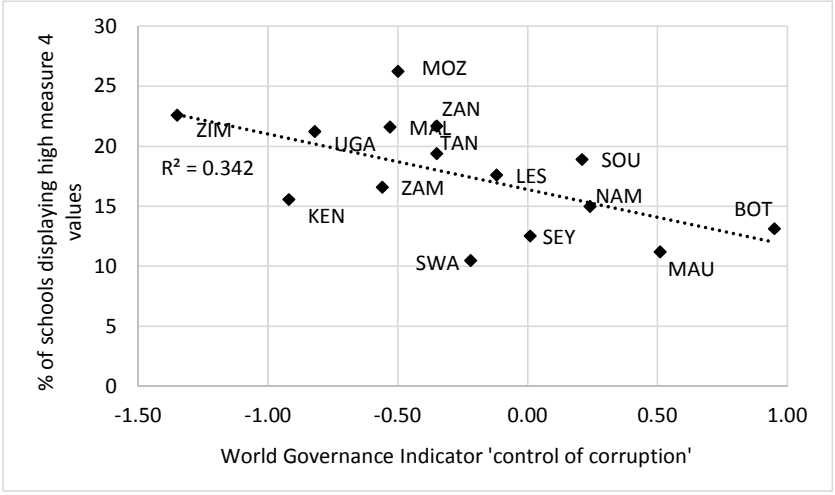
Figure 6 displays the correlation for mathematics in 2000. The R squared value is 0.625, and the Pearson correlation would be as high as 0.79. A less robust association is found when one uses the 2007 SACMEQ and WGI data (see Figure 7).

Figure 6: Mathematics measure 4 and control of corruption in 2000



Note: The 'high measure 4' threshold is 0.006 for this graph.

Figure 7: Mathematics measure 4 and control of corruption in 2007



Note: The 'high measure 4' threshold is 0.008 for this graph.

Of course the analysis cannot completely rule out possibilities other than a systematic link between general corruption and high measure 4 values. For instance, 'exceptional specialisation' caused by factors other than cheating could be more common in less developed countries, and less developed countries tend to be the countries with the worst WGI values. What does strengthen the evidence of a direct link between the two is the fact that when the country's average mathematics score, *S* in equation (6), is entered as an explanatory variable, the *p* value for *b* remains very low, 0.001 for 2000 and 0.040 for 2007. In other words, even if one controls for each country's level of mathematics performance, the association between what we suspect is corruption in education and perceived society-wide corruption remains statistically significant.

$$m_i = \hat{a} + \hat{b}W_i + \hat{c}S_i \qquad (6)$$

Given the thresholds of 0.006 and 0.008 we found in our comparisons with WGI data, we decided to use a threshold of 0.007 in calculating the statistics in Table 5 below. Higher percentages can be considered indicative of a higher probability of cheating, though given how low the threshold is, the percentages should not be considered exact indicators of how many schools are cheating. For instance, it seems counter-intuitive, given what has been shown above, to conclude that 17% of Botswana's schools cheated in 2007 mathematics. However, it seems one can say that the difference between Botswana's 17% and South Africa's higher 21% is suggestive of more cheating in the latter than the former.

Table 5: Percentage of schools with measure 4 exceeding 0.007

	2007					2000				
	Avg. score		% above .007			Avg. score		% above .007		
Coun-try	Math.	Read.	Math.	Read.	Both	Math.	Read.	Math.	Read.	Both
BOT	521	535	17	9	3	513	521	14	6	0
KEN	557	543	17	9	1	563	547	19	10	2
LES	477	468	20	18	8	447	451	16	18	7
MAL	447	434	22	20	12	433	429	23	13	7
MAU	623	574	15	9	4	585	536	15	12	2
MOZ	484	476	29	22	9	530	517	21	16	3
NAM	471	497	18	10	5	431	449	16	10	4
SEY	551	575	13	0	0	554	582	13	4	0
SOU	495	495	21	14	8	486	492	18	13	7
SWA	541	549	13	6	0	517	530	16	13	2
TAN	553	578	20	11	5	522	546	25	11	5
UGA	482	479	24	17	8	506	482	23	26	10
ZAM	435	434	20	13	7	435	440	20	9	4
ZAN	486	534	27	13	5	478	478	26	16	7
ZIM	520	508	25	15	6					
Average across countries			20	12	5			19	13	4
Correlation across years			0.75	0.65	0.69					
Correlation across subjects			0.84					0.54		
Correlation with avg. score			-0.53	-0.76				-0.22	-0.37	

The last three rows of Table 5 provide a few correlations across the country values. The correlation across years is high. Both in absolute terms, and in terms of country rankings, the percentage of schools above the 0.007 threshold did not change much. Whatever is being measured, which we believe is the probability of cheating, possibly combined with other factors contributing to 'exceptional specialisation', is rather stable over time. Moreover, countries with higher measure 4 values in one subject are likely to also have this in the other subject. Lastly, it is those countries with lower SACMEQ scores which tend to have more schools with high measure 4 values – this relationship is considerably stronger in 2007 than 2000.

Table 6 below attempts to sum up where the SACMEQ countries stand with respect to the probability of cheating. A solidly coloured marker indicates a relatively high probability, whilst a hollow marker indicates that the probability is low. For the within-country analysis, a solid marker appears if the instances of regional outliers across three year-specific columns in Table 3 was 0.2 or more of the total number of regions in the country. A hollow marker appears if there were no instances at all. For the comparison across countries, a solid marker appears if all three percentages in Table 5 were above the overall average, whilst a hollow marker appears if all three percentages were below this average.

Table 6: Summary of country performance against measure 4

		Within-country comparison across regions (Table 3)		Comparison across countries (Table 5)	
Abbr.	Country	2007	2000	2007	2000
BOT	Botswana	○	○	○	○
KEN	Kenya	○	●	○	
LES	Lesotho	●	○	●	
MAL	Malawi	○		●	●
MAU	Mauritius	○	○	○	○
MOZ	Mozambique	●		●	
NAM	Namibia		○	○	○
SEY	Seychelles	○	○	○	○
SOU	South Africa	●	●	●	
SWA	Swaziland	○	○	○	○
TAN	Tanzania		●		
UGA	Uganda	●	●	●	●
ZAM	Zambia	●	○		
ZAN	Zanzibar	○	●		●
ZIM	Zimbabwe		n/a	●	n/a

Note: ○ indicates patterns in the data indicative of low levels of cheating. ● indicates that patterns compatible with cheating were found.

Emerging as most problematic in this table is Uganda, followed by South Africa. Least problematic are Botswana, Mauritius and Swaziland (and Seychelles, though the provisos relating to the smallness of this country need to be kept in mind).

5 Adjustments to factor out cheating

An obvious question is whether it is possible to adjust the results of schools in the SACMEQ datasets in a manner that removes possible distortions in the results associated with high measure 4 values. A further question would be if such adjustments change substantially the patterns of performance which users of SACMEQ statistics have become familiar with over many years.

A simple approach was devised to replace the 2007 mathematics scores of pupils with imputed values in schools with high measure 4 values. Different thresholds for 'high' were employed. Imputation occurred using the reading scores of pupils, if they were considered sufficiently 'clean', and the pupil's socio-economic status (SES) as reflected by an SES indicator constructed by the producers of the SACMEQ datasets, using parent education, possessions in the home and the type of dwelling. Two stages were followed. First the regression of equation (7) was run, separately for each country, using only observations where *both* measure 4 values, for mathematics and reading, did not exceed the threshold. The dependent variable Y is SACMEQ's IRT score for pupil i in school s. For the reading score, R, and socio-economic status, E, the untransformed, squared and school-level mean values were entered in order to take into account the possibility of non-linear relationships and elements of school-level peer effects.

$$Y_{is} = \hat{a} + \hat{b}E_{is} + \hat{c}E_{is}^2 + \hat{d}\overline{E}_s + \hat{e}R_{is} + \hat{f}R_{is}^2 + \hat{f}\overline{R}_s + \varepsilon_{is} \quad (7)$$

The coefficients *a* to *f* were then used to calculate imputed mathematics scores for pupils in schools where the measure 4 value for mathematics, but not for reading, exceeded the threshold. Thereafter, the following regression, without the reading score variables, was run, using any observations where the measure 4 value for mathematics did not exceed the threshold (the measure 4 value for reading could exceed the threshold).

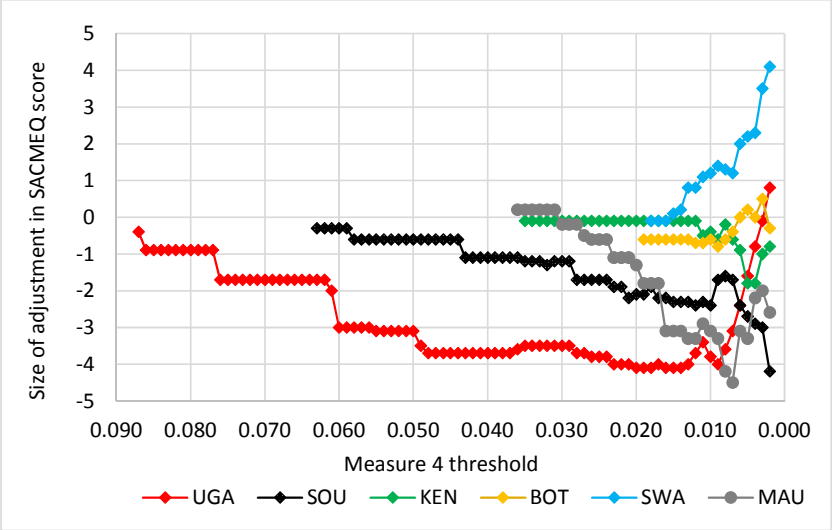
$$Y_{is} = \hat{a} + \hat{b}E_{is} + \hat{c}E_{is}^2 + \hat{d}\overline{E}_s + \varepsilon_{is} \quad (8)$$

Coefficients *a* to *d* were then used to calculate imputed mathematics scores for pupils in schools where the measure 4 values for *both* mathematics and reading exceeded the threshold.

Figure 8 below indicates by how much the 2007 mathematics averages for selected countries are adjusted, using different measure 4 thresholds. Uganda and South Africa have emerged above as countries with probabilities of cheating which seem high, whilst this probability has seemed low for Botswana, Swaziland and Mauritius. Kenya is included in the graph as a country roughly in the middle of the suspected cheating continuum. The reason why curves do not all begin at the same left-hand horizontal point in the graph is that countries with 'cleaner' data experience no adjustments at high thresholds because no schools have such high measure 4 values.

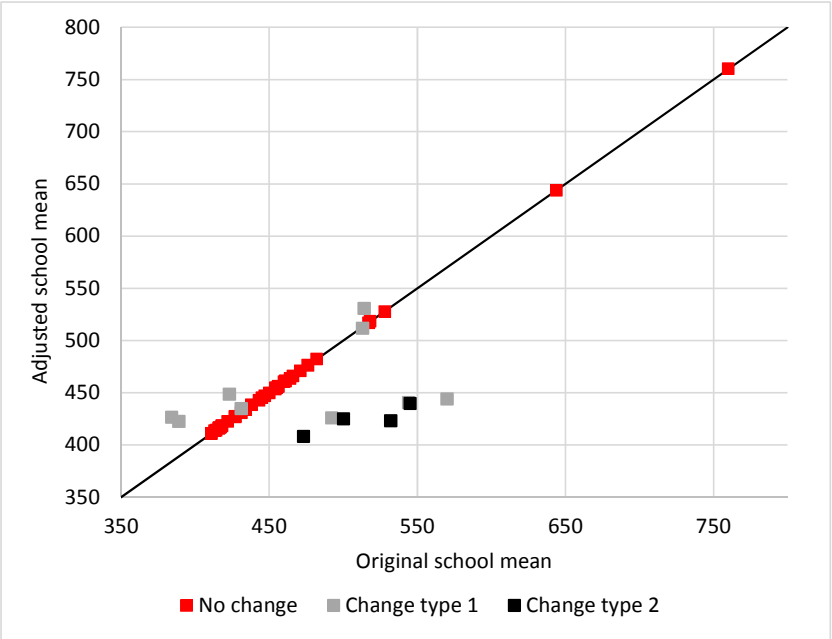
Clearly, below a threshold of 0.015 large adjustments occur which cannot be dealing with cheating. Swaziland experiences large upward adjustments and Botswana large downward adjustments in this part of the graph (the far right, as the thresholds are sorted from highest to lowest). A useful threshold would have to lie above (to the left of) 0.015. In the cases of Uganda and South Africa, it appears as if most of the adjusting that is required has been achieved by the time the threshold has moved down to around 0.020. Mauritius is interesting. Though its percentage of schools with high measure 4 values in the 2007 mathematics data is slightly better (lower) than Kenya's (see Table 5), according to Figure 8 Mauritius requires substantial downward adjustments in its average score, whilst Kenya requires virtually no adjustment. This is because of a few schools in Mauritius with fairly high measure 4 values which, according to our calculations, require very large downward adjustments. In fact, at the point at which Mauritius requires a downward adjustment of three SACMEQ points, only eight schools are driving the adjustment.

Figure 8: Measure 4 threshold and 2007 mathematics score adjustments



The following graph illustrates the adjustments for just Eastern Cape, to further clarify the methodology. Here a threshold 0.018 was used. 'Change type 1' means the first of the two regressions was used, whilst for 'Change type 2' the second regression was used. What is noteworthy is that in the adjustment process some schools with high measure 4 values are adjusted *upwards*. This would occur where reading scores and socio-economic status predict that an upward adjustment should happen. This is not inconsistent with our general framework. For example, schools may cheat in a manner that produces incorrect responses, and an overall level of performance below that which would be possible without cheating. Of course non-cheating factors could also explain why some schools experience upward adjustments. The overall pattern, however, is for downward adjustments to occur. The overall mean score for Eastern Cape before the adjustments was 468. After the adjustments it is 454.

Figure 9: Mathematics 2007 before and after adjustments in Eastern Cape



Based on the patterns seen in Figure 8, we decided to use 0.018 as the measure 4 threshold applicable to the adjustments seen in Table 7 below. We believe a threshold of 0.018 is neither so high that it misses many schools needing an adjustment, nor so low that it brings about adjustments which clearly have little to do with cheating.

Table 7: Mathematics 2007 adjustments using a threshold of 0.018

Abbr.	Country	Orig. mean	Total schools	Schools adjust 1	Schools adjust 2	% adjusted schools	New mean	Change in mean	Rank before	Rank after
BOT	Botswana	521	160	1	0	1	520.4	-0.6	6	6
KEN	Kenya	557	193	2	0	1	556.9	-0.1	2	2
LES	Lesotho	477	182	8	3	6	475.5	-1.5	12	12
MAL	Malawi	447	139	7	1	6	445.1	-1.9	14	14
MAU	Mauritius	623	152	7	0	5	621.2	-1.8	1	1
MOZ	Mozambique	484	183	15	1	9	483.5	-0.5	10	10
NAM	Namibia	471	267	11	1	4	470.6	-0.4	13	13
SEY	Seychelles	551	24	0	0	0	551.0	0.0	4	3
SOU	South Africa	495	392	18	6	6	493.1	-1.9	8	8
SWA	Swaziland	541	172	1	0	1	540.9	-0.1	5	5
TAN	Tanzania	553	196	13	0	7	550.7	-2.3	3	4
UGA	Uganda	482	264	19	5	9	477.9	-4.1	11	11
ZAM	Zambia	435	157	7	1	5	435.0	0.0	15	15
ZAN	Zanzibar	486	143	4	0	3	485.3	-0.7	9	9
ZIM	Zimbabwe	520	155	9	0	6	519.1	-0.9	7	7

What should reassure researchers and education planners who have become accustomed to using SACMEQ country scores, is that after new scores have been imputed for schools with relatively high measure 4 values, the overall mean values for countries change only slightly. Country rankings remain virtually unchanged, though Tanzania and Seychelles, which were close to each other to begin with, exchange places. The percentage of schools requiring adjustments never exceeds 9%. In all countries far more schools are adjusted using the more reliable equation (7), which includes the reading score as an explanatory variable, than equation (8), which relies entirely on the pupil's SES.

Given that mathematics produces higher measure 4 values than reading, in both 2000 and 2007, and given that the measure 4 situation for mathematics appears similar across 2000 and 2007, it seems we can conclude that in neither of the years and in neither of the subjects would these type of cheating-related adjustments substantially change country rankings.

6 School and home background factors associated with suspected cheating

What follows is a rudimentary analysis of background factors associated with high measure 4 values. The analysis is rudimentary in the sense that all relationships are assumed to be linear, a limited set of possible explanatory variables were explored and a deeper examination to establish causation was not attempted. With regard to the latter, it seems unlikely that the available data allow for firm conclusions around cause and effect.

Our analysis is rudimentary as explaining why certain schools in a country display higher probabilities of cheating was not a primary aim the paper. Our primary aim was to identify a useful indicator of possible cheating, and to explore the prevalence of cheating in SACMEQ. Yet even our rudimentary analysis appearing below seems to throw new light on how schools work, at least within the SACMEQ countries.

The dependent variable in our analysis was the degree to which a school's measure 4 value for mathematics in 2007 exceeded a reasonable threshold. The threshold we used was 0.018

(which we have already used above). A school which did not exceed this threshold was assigned a value of zero. A school with a measure 4 value of, say, 0.021 was considered to exceed the threshold by a margin of 0.003. This was then multiplied by 1000 to facilitate the reporting of our regression results, meaning 0.003111 would become 3.111.

We excluded from our analysis five countries which had fewer than seven schools exceeding the threshold of 0.018. This resulted in the group of ten countries listed in Table 8.

The nine independent, or explanatory, variables are indicated by the column headings in the next three tables. Table 8 provides the mean values across schools in each country, for the dependent and nine explanatory variables. Division by 100 or 10 occurred to facilitate the reporting of results. Small discrepancies between the score values in the second column of Table 8 and the scores seen in Table 7 above are due to the fact that pupil weights were not used in the analysis that follows. We were interested in examining correlations in the sample, not in arriving at nationally representative coefficients, though the coefficients we obtain are unlikely to differ much from estimates for the population. The teacher test cell for Mauritius in Table 8 is blank as this country did not participate in this part of the programme in 2007.

We decided to pay special attention to gender, given the widespread interest in the literature on the relationship between gender and corruption, and compelling empirical evidence, such as that of Dong and Torgler (2013), suggesting that a stronger presence of female leaders leads to less corruption. The percentage of school principals who are male ranges from 21% in Lesotho to 86% in Malawi. Similar differences are found across countries with respect to the gender of the mathematics teacher. We constructed a zero-one dummy variable indicating whether both the principal and mathematics teacher were male. The mean for the principal's age ranges from 41 (Mozambique) to 56 (Mauritius). SACMEQ's measure of socio-economic status was transformed into a z-score within each country, and at the pupil level, giving a mean of zero and a standard deviation of one (across pupils in a country, not necessarily across the school averages of a country). Finally, schools described by the school principal as being 'In or near a small town' or 'In or near a large town or city' were considered urban, whilst schools described as 'Isolated' or 'Rural' were not. This produces a proportion of urban schools ranging from 23% (Malawi) to 61% (Mozambique)[12].

Table 8: Mean of variables 2007

	Meas-ure 4 excess × 1000	Math. score / 100	Teach-er math score / 100	Princ-ipal is male	Math. teach-er is male	Both are male	Princ-ipal age / 10	Math. Teach-er age / 10	SES (z-score)	School is urban
LES	0.7	4.7	7.4	.21	.35	.10	5.1	3.8	0.0	.31
MAL	0.5	4.5	7.6	.86	.75	.70	4.5	3.7	0.0	.23
MAU	0.4	6.1		.58	.61	.36	5.6	4.3	0.0	.45
MOZ	0.6	4.8	7.4	.80	.71	.58	4.1	3.1	-0.1	.61
NAM	0.2	4.7	7.7	.61	.58	.35	4.7	3.9	0.0	.43
SOU	0.6	5.0	7.7	.67	.46	.31	4.9	4.1	0.0	.54
TAN	0.6	5.5	8.3	.81	.83	.68	4.2	3.6	0.0	.30
UGA	2.2	4.8	8.3	.79	.91	.73	4.4	3.4	0.0	.26
ZAM	0.6	4.4	7.4	.71	.50	.42	4.9	3.2	0.0	.32
ZIM	0.7	5.2	8.5	.71	.70	.51	4.8	3.8	0.0	.30

Table 9 below reports on the results of 90 bivariate regressions (10 countries by nine explanatory variables), as an introduction to the results of ten country-specific multivariate

[12] The Mozambique figure seems unexpectedly high and underlines the problem of asking school principals to provide information on the urban-rural classification, especially where different languages are used.

regressions in Table 10. A value appears if the coefficient for the explanatory variable was statistically significant at least at the 10% level.

Table 11 indicates that the attrition of observations from the multivariate regressions due to missing values was not worryingly high. However, the overall explanatory power of the regressions, as indicated by R squared, was in many cases very low. Yet the fact that 17% of the variation in our indicator of 'exceptional specialisation' could be explained by the nine variables in the case of Malawi and Uganda, and 29% in the case of Zimbabwe, is remarkable, and underlines the non-randomness of the phenomenon we are measuring.

Table 9: Coefficients from bivariate regressions

	Math. score / 100	Teach-er math score / 100	Princ-ipal is male	Math. teach-er is male	Both are male	Princ-ipal age / 10	Math. Teach-er age / 10	SES (z-score)	School is urban
LES	2.0								
MAL	2.9					0.9	0.9		
MAU					0.7				-0.7
MOZ								-0.6	
NAM				0.3					
SOU								-0.7	
TAN									
UGA	2.8							-1.8	-2.5
ZAM		0.6							
ZIM	1.0			-1.6					

The co-existence of positive coefficients for the mathematics score and negative coefficients for the SES variable in Table 10 is probably indicative of the fact that cheating tends to lead to higher scores. It is more socio-economically disadvantaged schools which appear to engage in more cheating, a pattern which would be consistent with the notion that teachers in these schools are caught between the home background disadvantage of their pupils and an administration that wants results, a situation which leads to desperate measures to improve scores. In the case of Uganda, the Table 10 results moreover suggest that teachers turn to cheating to compensate for their own lack of mathematics knowledge.

The gender relationships in three countries are noteworthy. In both Uganda and Zimbabwe, the worst measure 4 situation (so the highest measure 4 values) are associated with having both a female principal and a female mathematics teacher (the high positive coefficient for 'both are male' in the case of Uganda is more than offset by the high negative coefficients for the two separate positions). In the case of Lesotho, however, having a male principal is associated a higher measure 4 value, which would be consistent with the existing evidence on a negative correlation between the presence of female leaders and levels of corruption.

Table 10: Coefficients from multivariate regressions

	Math. score / 100	Teach-er math score / 100	Princ-ipal is male	Math. teach-er is male	Both are male	Princ-ipal age / 10	Math. Teach-er age / 10	SES (z-score)	School is urban
LES	2.6		1.6						
MAL	3.0						0.8		
MAU									
MOZ								-1.3	1.2
NAM									
SOU	1.6							-1.9	
TAN									
UGA	7.1	-1.0	-15.0	-11.3	14.5			-3.8	
ZAM									
ZIM	4.2		-2.4	-3.9				-3.0	

Table 11: Additional multivariate regression details

	N (number of schools/ observations)	% of observations lost due to missing data	R squared
LES	182	4	0.075
MAL	139	1	0.166
MAU	152	0	0.068
MOZ	183	3	0.072
NAM	267	6	0.022
SOU	392	8	0.084
TAN	196	4	0.039
UGA	264	11	0.167
ZAM	157	11	0.046
ZIM	155	10	0.293

7 Conclusion

The paper has considered existing literature on the detection of patterns in test data compatible with cheating, and drawn from the seminal work by Jacob and Levitt in this area to arrive at a method we could apply to cross-sectional SACMEQ data. Crucially, SACMEQ has not introduced a matrix sampling approach to its test design, an approach that carries costs but is widely considered to be best practice. This makes SACMEQ relatively susceptible to the risk of cheating during test administration.

We find patterns indicative of 'exceptional specialisation', or unusual excellence in certain test items, which could point to cheating. Whether these patterns represent corrupt practices or not cannot be determined exactly, but three critical checks that we run suggest strongly that higher values for our school-level indicator of suspicious patterns do point to a higher probability of cheating.

Firstly, we find that within countries high indicator values tend to be concentrated in certain regions, a fact which weakens counter-arguments such as that some countries display more 'exceptional specialisation' because their national curricula encourage more flexibility in schools regarding what topics are covered. In the case of South Africa, patterns compatible with cheating are concentrated in a province which has also displayed evidence of corrupt practices in other research.

Secondly, we used data from TIMSS, a testing programme which does use matrix sampling, in order to establish what our indicator values would look like in a system we assume is not affected by cheating. We find that certain countries and regions in SACMEQ do attain

TIMSS-like patterns, confirming that there is something suspicious about schools with indicator values falling outside a reasonable range.

Thirdly, when we aggregate our indicator of suspected cheating to the country level, we find a strong correlation, of 0.79 in the case of 2000 mathematics, between our values and a relevant World Bank indicator of general country-level corruption.

We find that Uganda and South Africa are particularly noteworthy cases of countries with cheating issues, whilst Botswana, Swaziland and Mauritius emerge as relatively 'clean'.

Importantly, when we impute 'clean' average test scores for schools with suspicious patterns, the overall country scores in SACMEQ shift very little, and rankings remain virtually unchanged. The picture of school performance offered by SACMEQ remains intact. Yet the presence of cheating in SACMEQ should be taken into account. Up to 9% of schools (in the case of Uganda and Mozambique) needed score adjustments to deal with cheating, according to our calculations. These findings strengthen the argument for introducing matrix sampling into the design of SACMEQ's tests. A reduced risk of cheating would be one of many advantages associated with this design change in SACMEQ, a programme which has already contributed immensely to better research and planning, but which can be improved in a number of ways.

References

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

Crouch, L. & Gustafsson, M. (2017). *Worldwide inequality and poverty in cognitive results: Cross-sectional evidence and time-based trends*. Research Triangle Park: RTI International. [Forthcoming].

Dong, B & Torgler, B. (2013). Cases of corruption: Evidence from China. *China Economic Review,* 26: 152-169.

Ferrer-Esteban, G. (2013). *Rationale and incentives for cheating in the standardised tests of the Italian assessment system*. Torino: Fondazione Giovanni Agnelli. Available from: <http://www.fga.it> [Accessed October 2017].

Gustafsson, M. (2015). Enrolment ratios and related puzzles in developing countries: Approaches for interrogating the data drawing from the case of South Africa. *International Journal of Educational Development,* 42: 63-72.

Jacob, B.A. & Levitt, S.D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics,* 118(3): 843-877.

Makuwa, D.K. (2010). Mixed results in achievement. *IIEP Newsletter,* XXVIII(3). Available from: <http://www.iiep.unesco.org> [Accessed October 2010].

Martin, M.O., Mullis, I.V.S. & Hooper, M., eds. (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill: IEA. Available from: <https://timssandpirls.bc.edu/publications/timss/2015-methods.html> [Accessed November 2017].

Papanastasiou, E.C. & Zembylas, M. (2006). Did the Cypriot students really cheat on TIMSS? *Research in Comparative & International Education,* 1(2): 120-125.

Ross, K.N., Saito, M., Dolata, S. & Ikeda, M. (2008). *Chapter 2: The conduct of the SACMEQ II project*. Paris: IIEP. Available from: <http://microdata.worldbank.org/index.php/catalog/1245/download/22682> [Accessed July 2017].

Solano-Flores, G. & Bonk, W. (2008). *Evaluation of the Latin American Laboratory for the Evaluation of Educational Quality (LLECE)*. Paris: UNESCO. Available from: <http://unesdoc.unesco.org/images/0016/001626/162674e.pdf> [Accessed November 2006].

Wills, G., Shepherd, D. & Kotze, J. (2016). *Interrogating a paradox of performance in the WCED: A provincial and regional comparison of student learning*. Stellenbosch: University of Stellenbosch. Available from: <https://ideas.repec.org/p/sza/wpaper/wpapers245.html> [Accessed October 2016].

Appendix: Stata (12) code for adjusted Measure 4

```
* CODE FOR AN ADAPTED JACOB AND LEVITT (2003) MEASURE 4

* What is needed open is a table of classical scores by item, and by pupil. Each observation is a
* pupil. Pupil ID should be excluded, though there should be an ID for country (or region) and school
* (or class in a school). These variables should be called 'IDCountry' and 'IDSchool'. They should
* both be numeric. If everyone is from the same country (or region), the first variable should carry
* just one value. Item names should constitute all the remaining variable names. They should be named
* 'i1, 'i2', i3, and so on, for instance up to 'i45'. Each should contain a numeric value indicating
* the score of the pupil, and no values should be missing. The code produces an ADJUSTED Measure 4
* value for every school (or class), within the variable 'M4'.

set more off
egen totscore = rowtotal(i*)
summ totscore, det // Might be useful to observe.
local maxscore = r(max)
foreach i of varlist(i*) {
  gen qdiff2`i' = .
}
forvalues s = 0 / `maxscore' {
  display "s is " `s'
  foreach i of varlist(i*) {
    egen tempmax = max(`i')
      gen tempperc = `i' / tempmax
    quietly by IDCountry, sort: egen tempmean = mean(tempperc) if totscore==`s'
    quietly replace qdiff2`i' = (tempperc - tempmean) ^ 2 if totscore==`s'
      drop temp*
  }
}
egen Z = rowmean(qdiff2*)
gen Zmean = .
forvalues s = 0 / `maxscore' {
  quietly by IDCountry, sort: egen temp = mean(Z) if totscore==`s'
  quietly replace Zmean = temp if totscore==`s'
  drop temp
}
gen M4sum = Z - Zmean
gen s = 1
collapse (sum) M4sum s, by(IDCountry IDSchool)
gen M4 = M4sum / s
drop M4sum s
```