Catching Cheaters in Hungary

estimating the ratio of suspicious classes on the National Assessment of Basic Competencies tests

Daniel Horn*
Institute of Economics, Research Centre for Economic and Regional Studies, Hungarian Academy of Sciences
and
Department of Economics, Eötvös Lóránd University, Budapest
horn@econ.core.hu

address: H-1112 Budapest, Budaörsi str. 45. room 814. Hungary

*Assistance by Ágnes Lak (Hungarian Education Authority) is highly appreciated. Comments from Zoltan Hermann, Gabor Kertesi, Gabor Kezdi and Julia Varga are thankfully acknowledged. This research was supported by a grant from the CERGE-EI Foundation under a program of the Global Development Network. All opinions expressed are those of the author and have not been endorsed by CERGE-EI or the GDN.

Catching Cheaters in Hungary - estimating the ratio of suspicious classes on the National Assessment of Basic Competencies tests

1 - Introduction

The 99§ of the LXXI/2006 amendment to the LXXIX/1993 law on education has introduced high stakes testing in the public education system of Hungary. The amendment states that if a school performs below a national minimum on the National Assessment of Basic Competencies (NABC) the education provider must investigate the reasons and the school must take appropriate measures. If the school does not improve in the following two years, the local government has to invite independent experts to assist school reform. The reform should mainly address teaching practices, but could also have effects on the leadership or on the staff composition. Besides these direct interventions reports on schools are annually publicized. Such transparency can be considered a serious accountability tool in a choice based school system, such as Hungary.

Although there is only scattered (anecdotic) evidence about the effects of this accountability system, it is likely that the NABC will play a larger role in the governance of the Hungarian education system in the near future. For that reason it is necessary to scrutinize how reliable the test score information is.

This paper builds largely on the procedure presented by Brian A. Jacob and Steven D. Levitt (2003a) in their seminal paper "Rotten Apples" that estimates the ratio of potentially cheating classes. Although their methods were tested on multiple choice data from Chicago, the procedure is general enough to apply it in a rather different setting, with notes of caution attached where necessary. As value-added, I complement their methodology by offering new ways of indicating suspicious classes for tests based on item response theory (IRT). The main goal of this paper is to estimate the ratio of suspicious-cheater classes in Hungary, but also to show that the Jacob and Levitt method can be used in a rather different setting, with some modifications.

2 – Cheating literature

There are three typical clusters of literature that investigates the prevalence of cheating on tests. The first use simple survey based methods. Asking students about their typical cheating behavior gives a picture on how prevalent cheating is and thus how acceptable are the figures that teachers, administrators or policy makers rely on (Bunn, Caudill, and Gropper 1992; Kerkvliet 1994). This is a simple, but not a very reliable way to estimate the size of the problem. These studies, for instance, highlight differences between sexes (Whitley, Nelson, and Jones 1999), or between countries. Grimes and Rezek (2005), for instance, argue that Eastern European students of economics are more likely to cheat on tests than their American peers; an observation that makes my current study even more interesting.

Another cluster of studies, written mainly by psychometricians, want to investigate the reliability of a given test or the reliability of a given indicator that is used to assess tests (e.g. Cizek 1999; Sotaridona and Meijer 2002; Wollack 1997, 2003). There are several indices that are invented and carefully assessed in order to produce the best results in detecting which student is likely to have copied the answers from the others nearby. These analyses usually rely on very detailed data that contain information on the seating order of the students, or use simulated databases to compare the validity of cheating indicators.

The third line of study looks at the perverse effects of the accountability systems. These papers usually assess the reliability of tests used for accountability purposes. They investigate the prevalence of perverse effects such as teaching to the test (Koretz 2002), the reallocation of resources within or between schools due to the high-stakes testing (Burgess et al. 2005; Jacob 2005), selection of teachers between schools (Ladd 2001; Murnane and Levy 2001), or the exemption rate of the worst students (Cullen and Reback 2006). Cheating on the tests is a strong form of teaching to the test (Koretz 2002): it is only a matter of degree whether teachers focus on a specific group of questions/topics (reallocation of resources), or whether they give students the answers before testing (coaching) or during testing (cheating). An important piece of work is Jacob and Levitt's (2003a) Chicago study that merge the two latter braches. They look at one specific test – the Iowa Test of Basic Skills – but still offer new ways to estimate the ratio of potential cheaters in order to assess the reliability of high stakes testing in Chicago.

Rotten Apples – method

Although Jacob and Levitt (2003a) describe their methods extensively in the appendix of their paper, in order to make this study more comprehensible I describe

their methodology briefly. Jacob and Levitt (2003a) suggest two indicators that might be adequate for catching cheaters on multiple choice literacy tests. They argue that if a class scores high on both of these indicators, the class is a candidate for cheating. The first indicator is more straightforward, but potentially biased, while the second is much more complex, but more likely to address the real cheater.

The first indicator simply looks at the unexpectedly *large test score fluctuation*. If a class of students perform unexpectedly well in one year as compared to the previous and to the next, it is considered a potential cheater. Although this measure is quite straightforward, obviously it easily contains both type I and type II errors. In other words large test score fluctuations might be caused by many other things besides cheating (the type II error): change in classroom composition, change in teacher composition, or change in classroom conditions, just to mention a few. Additionally, cheating can occur without apparent changes in tests scores as well (the type I error): for instance cheating badly, so that it does not increase test scores, or cheating well so that it counteracts other reasons that would have produced a drop in test scores.

The second indicator comprises four different measures of *suspicious answer strings.* The first of these four measures (M1) looks at blocks of identical answers on consecutive questions. The idea behind M1 is that teachers are more likely to alter blocks of answers both across questions and across students. It is easier to erase items in one block than to randomly erase some items, while leaving others untouched. Or it is also easier to tell groups of students the results on consecutive items than to randomly choose some. Naturally, if this assumption is incorrect, their first measure fails to identify cheaters. Jacob and Levitt (2003a) predicts the probability of choosing a particular response for each student for each question based on their previous and future test scores and background characteristics using a multinomial logit estimation. They get the probability of choosing a response for each student for each item. Then they search over several combinations of items and students to find the least likely string of identical answers within each class. The smaller the probability of this string of answers within a class is the more likely cheating has occurred.

Their second measure (M2) of suspicious answer strings looks at each item and estimates how unexpected the answer was. They predict the probability of a response with the same multinomial logit as above, and calculate the error term for each response (1 minus the probability if the response was chosen, 0 minus the probability if it was not

chosen). Summing this across students will give some sort of a within class correlation for each response on each item. If there is no pattern in the answers this statistic should be around zero. Then they take the average sum of squares across the four responses and then sum the four means across students, and finally take the mean across questions. This measure captures something like a mean of within class correlation of student responses on a question. "This measure is high if students give the same answer across many questions, especially if the answers given are unexpected." (Jacob and Levitt 2003a:851)

The third measure (M3) is the variance as opposed to the mean of M2. If blocks of questions are altered within class and other blocks remain unchanged, the variance of within class correlation of responses is higher.

The fourth measure (M4) looks at the pattern of correct responses taking into account the "ability" of the student.[1] That is it calculates the ratio of correct responses within a class for each score level, the larger the differences between the class mean and the national mean across score levels, the likelier a class has cheated.

In order to generate the suspicious answer string indicator Jacob and Levitt (2003a) rank the classes along all four measures, take the sum of squared and rank the values again. The higher a class of students is on this rank the more suspicious the class is.

They use these two indicators to estimate the ratio of potential cheaters in Chicago. Using the 95% cutoff point they find that only 1.1% of all schools within one year and one subject are suspected of cheating. In other words, of the 5-5% schools, which score highly on either one of the two indicators described above, only 1.1% scores above the 95% cutoff point on both dimensions. When looking at more years, or more subjects, this ratio goes up to 3.5%. When relaxing the 95% cutoff point down to 80%, the ratio will go as far up as 5.6% (see Jacob and Levitt 2003a:859).

3 – Hungarian data and differences

The National Assessment of Basic Competencies (NABC)

As far as I know no similar study has been done for Hungary or for the Central Easter European region. This is of course unsurprising, since student literacy measures are not

[1] Note that the measures M1, M2 and M3 do not consider whether an answer is correct or not.

widely spread in this region. Of the new EU member states Hungary has the most developed testing system, the National Assessment of Basic Competencies (Balazsi 2012 forthcoming).

The NABC is a standard based assessment designed similarly to the OECD Programme for International Student Assessment (OECD 2001) survey, but conducted annually in May. It measures reading and mathematical literacy of the 6th, 8th and 10th grade students and it is standardized to a mean of 1500 with standard deviation of 200.[2] Mathematics and reading scores are standardized not only within but also across years. The average score of 6th grade students in 2008 is set to 1500 (both in math and reading) and each cohort and grade is measured to this 2008/6th grade cohort. For instance, if the average mathematics score of the 6th grade students in 2010 is higher than 1500 that means that this cohort's average mathematical literacy is higher than that of the two year older cohort. Similarly, one can compare scores across years.[3] Table 1 below shows when and who was measured within the NABC survey. There are several explicit goals of this assessment: first is to provide more detailed and more frequent feedback for educational policy than the international surveys. The second is to offer objective assessment for the local school providers and schools themselves. And the third goal is to set the grounds for an accountability system (this includes publicizing the data). In addition to all this, it offers invaluable data for researchers to address education related puzzles. Unfortunately, up until 2008 the database could only be analyzed on a cross sectional basis, since it contained no permanent student identification numbers. From 2008 onwards the biannual datasets are connected on the student level, thus from 2010 onwards more detailed analyses are possible. (Note that from 2012 three waves of the assessment could be linked.)

(table 1 around here)

[2] In earlier waves, before 2010, the mean was 500 with 100 standard deviation.
[3] See the description of the score generation procedure here (in Hungarian, accessed 07-01-2011): http://www.kir.hu/okmfit/files/Valtozasok_az_Orszagos_kompetenciameres_skalaiban_vegleges.pdf

Besides test scores the database contains variables from an extensive student background questionnaire and a school site level questionnaire. These questionnaires resemble that of the PISA survey.[4]

Differences between systems and tests

Although the National Assessment of Basic Competencies has been conducted since 2001, and school reports are publicized since 2006, no systematic study of cheaters has been done. To fill this gap I adopt the introduced Jacob and Levitt (2003a) method to estimate the ratio of suspicious classes, with notes of caution where necessary.

There are important differences between the education testing of the United States (or Illinois, Chicago in this case) and Hungary. One of the two main differences between the Iowa Tests of Basic Skills and the Hungarian NABC is that the American test contains only multiple choice items, while a large part of the NABC is not pre-coded (e.g. open-end questions, or exercises with one correct but not pre-defined answers). Also there are complex questions that worth more than one points. The other major difference is that in US teachers collect and "correct" test sheets after testing (i.e. teachers supposed to erase unnecessary marks from the paper, emphasize vague marks… etc.), while Hungarian teachers collect the sheets but cannot alter any information on them (students use pen instead of pencil).

These two are small but important differences. The Jacob-Levitt method is specifically designed for multiple choice questions. Their core indicator of *suspicious answer strings* considers two equally wrong but different strings of answers as not similar. However I had to transform all answers to points due to the large number of non multiple choice items. Thus I can only test "point-strings", i.e. look at whether students received the same points on consecutive items.[5] This probably overstates my estimate of cheating classes in the suspicious answer string indicator.[6]

My analysis is more likely to identify cheating classes, than cheating teachers, because teachers are unlikely to modify anything after the tests have been collected. Or more precisely, the Hungarian cheating are more likely to occur during testing (teachers

[4] The questionnaires, the school reports and all related documents can be downloaded in Hungarian from the http://www.kir.hu/okmfit/ website.
[5] Note that most of the items are 1 point items. Hence when looking at "point-strings" I practically look at strings of correct/incorrect answers. There are five items for 2 and one item for 3 points out of the 144 items.
[6] In other words, I consider two wrong answers as similar, even if students answered different – but wrong – answers to these.

helping students or students copying from each other) and thus affect one class of students, while the American cheating is assumed to be done after testing and thus affect a whole grade.[7] I believe this difference between the cases is a lesser problem, because teachers in Hungary are also more likely to help more students than one, and also intuitively more likely to answer consecutive questions than to randomly answer some.

There are more serious problems with Jacob and Levitt's *large test score fluctuation* indicator when I want to adopt it to the Hungarian case. The data used by the American analysis are much longer, as it covers seven years (1993-2000) and five grades (3rd to 7th), all organized in an individual level panel. The NABC data however is still a two wave panel. The *large test score fluctuation* indicator looks at the "middle" year: it looks at each year as compared to the previous and to the next and signals if it is significantly greater than both. Such indicator cannot be calculated from the NABC data. Therefore I will rely on a unique feature on the NABC dataset, which allows me to substitute the Jacob-Levitt indicator with an alternative (see section *"Indicator 1 – unusually small standard error"* below), and only use the change in individual test scores for a robustness check.

Additional difference between the two tests is that there are two groups of test sheets in the NABC tests. In order to limit cheating, students sitting next to each other receive two types of test sheets, in which the order of the blocks of questions is altered.[8]

In order to make the most out of the differences between the two tests and the two systems I will use the 2008/8th grade 2010/10th grade panel, analyzing the 8th grade items to estimate the ratio of cheaters.

Students in Chicago do not change tracks during the observed period, but all Hungarian students leave general school after 8th grade and enter a secondary track afterwards. Thus, it is less likely that students in 8th grade have cheated, as compared to 6th or 10th grade, because they already know the secondary school where they have been accepted at the time of testing. So they would not be in the same school a year later,

[7] *Class* in the American literature usually refers to a whole grade, while in Hungary a class is a group of students (approximately 20-30 students) attending the same lessons through the whole year. There are usually parallel classes within one grade.

[8] The first group of students fill out the first block of questions, while the second groups deals with the second block and then they switch. Both groups receive the same questions.

when results arrive. Hence testing for 8th grade students are especially low stakes. But it is not for the teachers and the school. Therefore I expect my estimates to be modest.

Differences between the implementation of two the tests suggest that the ratio of cheaters should be smaller in Hungary than in the US: open-ended questions, cheating can only occur during testing, two types of test sheets, low stakes for students. On the other hand examining point strings instead of answer strings overestimates cheating. Moreover difference in cultures also suggests a larger rate of cheating in Hungary. Grimes and Rezek (2005), for instance, argue that Eastern European students are more likely to cheat on tests than their American peers. Thus I am unable to tell whether estimates should be higher or lower in Hungary compared to the US. I believe that the method presented by Jacob and Levitt complemented with a new indicator gives a decent approximation of the ratio of cheaters in Hungary.

4- Cheating in the Hungarian classrooms

Indicator 1 – unusually small standard error

The Hungarian NABC data does not allow for the estimation of Jacob and Levitt's (2003a) *large test score fluctuation* indicator. Fortunately the NABC test is based on Item Response Theory (IRT - see DuToit 2003), which provides an opportunity to estimate the probability of a correct answer for each student on each item. In the NABC test each item is characterized by the parameters β - slope or discrimination, γ - difficulty and δ - step difficulty (for 2 or 3 point questions). Using these parameters the predicted probability of correct answer (i.e. the predicted point) can be calculated for every student for each item. This probability (P) is calculated as

$$P_{isc} = \frac{\exp\left[\sum_{v=0}^{k} 1{,}7\beta(\theta - \gamma + \delta_v)\right]}{\sum_{\mathrm{d}=0}^{\mathrm{m}} \exp\left[\sum_{v=0}^{d} 1{,}7\beta(\theta - \gamma + \delta_v)\right]}$$

for item i and student s in class c. k=1, 2, … m, where m is the maximum number of points of an item. θ is the ability of the child and $\theta - \gamma - \delta_0 \equiv 0$ and $c_0 \equiv 0$ (see DuToit 2003:556).

The ability of the student (θ) is her/his aggregate test score point.[9] This variable is likely to be endogenous: cheating students are likely to achieve higher scores, thus gain higher predicted probability. Looking simply at the difference between the actual scores and the predicted probability as an indicator for cheating would lead to biased results.
Therefore instead of taking the difference between the means I looked at the difference between the standard errors of the two statistics. I expect that classes have lower standard errors on items where they cheated. I assume that if teachers cheat on an item, they do not help individual students, but help all. This is a strong assumption. But it follows that the standard error of the predicted probability of an item should not be affected by cheating. Let's assume alternatively that teachers help only the less bright ones within the class. This would decrease the standard error of the actual points on an item, but it would also decrease the distribution of θ a little, which in turn decreases the standard error of the predicted probability of all items. In this case the standard error of the actual points on an item is changed only if cheating occurs, but the standard error of the predicted probability of all items is decreased. Hence, if teachers help only the lower ability students indicator1 (defined below) underestimates the number of cheaters. If teachers help the brighter ones it overestimates cheating. If teachers help all students equally the estimate is unbiased.

To estimate indicator1 (*unusually small standard error*) I calculated the standard error of the mean of the average number of points (*H*) given for an item for each class:

$$se(H_{ic}) = se(\frac{\sum_{s=1}^{n} h_{isc}}{N})$$

where *h* is the actual point given for an item and N is the number of students in the class. The standard error of the predicted probability is also calculated.[10] The mean of the difference between the two statistics is *indicator 1:*

$$indicator1_c = \frac{\sum_{i=1}^{n}[se(P_{ic}) - se(H_{ic})]}{z},$$

$$where\ \ indicator1 = 0\ \ if\ indicator1 < 0$$

[9]I recalculated the test score points for each child using a Maximum Likelihood estimator.
[10] The standard error of the mean of the predicted probability is calculated using a normal distribution, while the standard error of the mean of the actual test score point is calculated using a binomial distribution, except for the 2 and 3 point items.

where z is the number of items on the test.[11] I minimized *indicator1* in 0, because I want to focus on those suspicious classes that have a lower actual error compared to the predicted error. In this way the differences between the suspicious classes and not the difference between the non-suspicious classes drive the robustness checks below.

Indicator 2 – suspicious answer strings

There are some minor, but important modifications that I had to make while adapting the Jacob-Levitt (2003a) *suspicious answer string* measures.

M1, as presented in the Jacob-Levitt appendix (2003a), does not take into account the size of the class.[12] However, the larger the size of the class is, the more likely it is that some students will by chance choose the same responses on consecutive questions. US classes tend to be large, because usually the whole grade is considered a class, however in Hungary classes are small groups of students studying together for the whole year in every subject, thus controlling for class size is more important in the Hungarian context. In order to adjust for class size I multiplied all probabilities with the size of the given class divided by the number of students answering similarly for given string of items.[13] I also drop 285 of the 5202 classes from the whole analysis that have less than 10 students for the same reason. Within these very small classes the estimation of any statistic is especially unstable (Kane and Staiger 2002). Also due to the small number of these small classes the change should not affect the result of this paper – the estimation of the cheating ratio – greatly.

M2 and M3 can both be easily calculated the same way as it is described in the Jacob and Levitt (2003a) appendix. No modifications were necessary, besides the already mentioned fact, that not all Hungarian test items are multiple choice questions, thus all items are transformed into points (mostly 0 and 1).

M4 can also be straightforwardly adopted to suit the NABC data. However, because Jacob and Levitt do not specify how they understand ability levels I use the

[11] There are 65 items in the math test and 79 items in the reading test.

[12] Note that Jacob and Levitt do not specify how they adjust for the different class size, but in the working paper version of their paper they note that they do, without sharing the details (Jacob and Levitt 2003b).

[13] That is $p_{sc}^{jl} = (\prod_{i=j}^{l} p_{isc}) * \frac{N}{n}$, where p is probability, *i* is the student, *s* is school, *c* is class, N is the class size and n is the number of students giving the same responses for the j-l string.

deciles of the average test score to classify students, and calculate the average ratio of correct answers within the deciles. Although the NABC provides test score levels similarly to the PISA data, the number of observation in the upper and lower levels were relatively low, so I opted for deciles instead of the pre-defined levels.

In order to generate *indicator2* Jacob and Levitt rank each measure on the class level and calculate the sum of squared.

$$Indicator2 = rankM1_c^2 + rankM2_c^2 + rankM3_c^2 + rankM4_c^2$$

In their empirical work, they use alternative cutoff points to indicate possible cheating.

Ratio of potential cheaters

TableTable 2 below shows the percent of suspicious classes on the 2008 8th grade mathematics and reading test, calculated similarly to Jacob and Levitt (2003a:859). Instead of their *large test score fluctuation* indicator I use the *unusually small standard error* indicator (indicator1).

Apparently, using this method, cheating in the Hungarian classrooms is a bit more prevalent than in the US. Using the 95th percentile cutoff, around 1,5-1,8% of the classes are suspected of cheating in any of the two tests. This is slightly higher than the Chicago figure of 1,1%, but practically is not different.

(table 2 around here)

For following robustness checks I use the 95% cutoff point on both dimensions to indicate potentially cheating classes. That is, I assume that there are 89 classes (1,83% of all classes) in math, and 74 classes (1,52% of all classes) in reading that are potentially cheaters, that is they are *suspicious*.[14]

Figure 1 below shows whether the 95% point is a sensible cutoff point at all. Apparently the fraction of classes on both the *unusually small standard error* and the *suspicious answer strings* indicators decline continuously, without large gaps in the data. Hence the

[14] The 89 classes are in 85 schools, and the 74 classes are in 70 schools. 36 single class schools are suspicious on both tests.

choice of cutoff points is somewhat arbitrary. Nevertheless in the next section I will show that these suspicious classes are indeed different from the others.

(figure 1 around here)

Figure 2 reproduces Jacob and Levitt's Figure II (2003a:854), which shows how the identification strategy works empirically. The horizontal axis is the rank value of *indicator2*, while the vertical axes shows the probability of a class being above 95th percentile on *indicator1.*

Until about the 60th percentile on the horizontal axis the relationship between the two indicators is not very strong. Most likely, this part of the distribution contains very few cheating classes. On the far right end of the distribution, the correlation of the two indicators rises sharply. The probability of being in the top 5% on *indicator1* rises to a rather high 40-60% as a class approaches the high end of the *indicator2* distribution. This observation is comparable to that of the Jacob and Levitt study.

Black dots in figure 2 represent averages of suspicious classes (classes above the 95th percentile on both dimensions). This shows that these classes are indeed outliers in these two dimensions, but also that here are other classes (grey dots near the black dots) that are on the margin, and due to the arbitrariness of the cutoff point they fall outside the suspicious class category.

(figure 2 around here)

5- Robustness checks

In this section I run several robustness checks to see whether the two indicators are in line with other intuitive tests of cheating. I show three robustness checks. The first tests whether schools that are under the national minimum, and thus are likely to be under the education providers closer attention, are more likely to cheat or not. The second looks at schools where the 2008/8th grade average test scores were drastically higher than the average test scores of 8th graders in 2007 and 2009. And the final test looks at students in 10th grade, and how their test scores have changed from 8th grade.

Schools under the national minimum

The §7 of the 3/2002. (II. 15.) decree of the Ministry of Education sets the national minimum in both subject areas. According to the LXXI/2006 amendment to the LXXIX/1993 law on education schools that are under this minimum must provide an

administrative plan for the education provider that outlines what the school will do to jump over this threshold. While this is not a very strong measure, intuitively the pressure on these schools should be high.

The national minimum is set on a really low level: more than half of the students must be above level 1 in reading and three-quarters of the students must be above level 1 in mathematics. Students on or under level 1 are practically considered as illiterate.[15] I assume that schools that fall into this category are more likely to use disallowed methods to improve the average test score.

Note the difference between the thresholds of the two subject areas. While the threshold in reading is very low (50% of students have to be under level 1) the threshold in math is not that low (only 25% of students have to be under level one). Of all schools 27% in 2007 and 24% in 2006 fell under the minimum in math, but only 4% in 2007 and 3% in 2006 have fallen under the minimum in reading. These are typically very small schools.

Kane and Staiger (2002) show that small schools are more likely to be either at the low or at the high end of the quality distribution. Due simply to the fact that few good or bad students can greatly affect the quality of the school I have to control for the size of the class. In all estimations below I use the fifth order polynomial of school size as control.

Table 3 shows that schools under the national minimum in 2007 are just as likely to fall into the suspicious-cheating class category as the others. The results do not support the intuition that these schools are more likely to cheat.

Note however that looking only at the bottom half of the distribution – classes under the average mean test score in 2007 – the relation in reading becomes slightly significant. This observed non-relationship on the total population might be due to cheating classes that are on the top of the distribution: classes that have cheated last year and still cheat this year. In order to see whether this is really the case, one must look at more than one year of tests.

(table 3 around here)

[15] Levels of the NABC data are similar to that of the OECD PISA study. The average test score (and standard deviation) in 2008/8th grade was 1652 (207) in reading and 1617 (201) in math. Level 1 students, on average, scored 1154 in reading and 1253 in math.

Outlier schools

Schools that jump out in one year and then fall back are rather suspicious. Unfortunately I cannot investigate classes through time, but I can compare different cohorts of the same school. Even if comparing different children might not be as helpful in detecting cheating, it is still suspicious if a school performs very well in one year but not before or after. I consider a school an "outlier" if the average score in 2008 is higher than the 90th percentile of the same schools in 2007 *and* in 2009, or if the 10th percentile in 2008 is higher than the average in 2007 *and* the average in 2009. Only 39 classes in 29 schools are considered as outliers in math, and 19 classes in 16 schools in reading (there are 2428 schools with 5175 classes with non missing data). That is 0.75% of the classes are outliers in math and 0.37% in reading.

Table 4 below shows the relationship between outliers and cheating suspicion. Classes in schools that score better in 2008 than before or after are significantly more likely to be considered suspicious. The size of the effect is highly significant and very large. An outlier school is about 8% more likely to fall into the suspicious category in both math and reading. The average ratio of suspicious classes in the total population is around 1,5%, while the average of the suspicious classes among the outlier schools is 28,5% in math and 26,3% in reading.

(table 4 around here)

Value added

A strong indicator of cheating is the large test score fluctuation of students. A little less robust, but still important result would be if the test scores of students in cheating classes have fallen back significantly in the next year. I can compare the 10th grade test score of students in suspicious classes with students in non-suspicious classes using the panel feature of the NABC data. Each student enters secondary level schooling at the end of the 8th grade. Most students, but the ones in early selective tracks, change school, and start to attend a different class with different peers. I expect the test score points of students in suspicious classes to drop significantly, as compared to other students with similar family status and sex in the same secondary level class. Table 5 below shows a regression with secondary level class fixed effects. Apparently those, who attended suspicious primary classes, drop 85 points in math and 50 points in reading. Note that the average standard deviation is around 200, so the effects are not only significant but quite sizeable as well.

(table 5 around here)

A non-parametric estimation shows similar results. Using propensity score (nearest neighbor) matching I compare the 10th grade test scores of the treatment (suspicious class) and the control (not suspicious class) groups (table 6). Propensities are calculated using 8th grade test scores, family status, sex and the 10th grade school of the student. Apparently the effects are not much different from the linear effects (table 5). Students in suspicious 8th grade classes perform much worse in 10th grade.

(table 6 around here)

6- Conclusion

This study adapts the Jacob and Levitt (2003a) method of cheating detection. Although their method is useful for detecting suspicious-cheating classes for most student-testing, the lack of long panel data in Hungary forced me to invent new measures of cheating indicators. Using the IRT feature of the Hungarian Assessment of Basic Competencies data a new indicator of cheating is the unusually small standard error. This indicator compares the variance of the real data with the variance of the simulated data within classes. Using this indicator and the suspicious answer string indicator of Jacob and Levitt (2003a) I show that the rate of cheating in the Hungarian classes are rather similar to the ratio of cheating in Chicago.

Using three separate robustness checks I show that the combination of the two measures, the unusually small standard error and the suspicious answer strings, indeed finds the suspicious classes.

References

Balazsi, Ildiko. 2012. "OECD Reviews on Evaluation and Assessment in Education - Hungary."

Bunn, Douglas N., Steven B. Caudill, and Daniel M. Gropper. 1992. "Crime in the Classroom: An Economic Analysis of Undergraduate Student Cheating Behavior." *The Journal of Economic Education* 23(3):197–207.

Burgess, Simon, Carol Propper, Helen Slater, and Deborah Wilson. 2005. *Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools*. Department of Economics, University of Bristol, UK.

Cizek, Gregory J. 1999. *Cheating on tests: how to do it, detect it, and prevent it*. Routledge.

Cullen, Julie Berry, and Randall Reback. 2006. *Tinkering Toward Accolades: School Gaming Under a Performance Accountability System*. National Bureau of Economic Research, Inc.

DuToit, Mathilda, ed. 2003. *IRT from SSI*. Scientific Software International, Inc.

Grimes, Paul W., and Jon P. Rezek. 2005. "The Determinants of Cheating by High School Economics Students." *International Review of Economics Education* 4(2):23–45.

Jacob, Brian 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, Brian A., and Steven D. Levitt. 2003a. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *The Quarterly Journal of Economics* 118(3):843 –877. Retrieved July 18, 2011.

Jacob, Brian A., and Steven D. Levitt. 2003b. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *National Bureau of Economic Research Working Paper Series* No. 9413. Retrieved July 18, 2011.

Kane, Thomas J., and Douglas O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *The Journal of Economic Perspectives* 16:91–114.

Kerkvliet, Joe. 1994. "Cheating by Economics Students: A Comparison of Survey Results." *The Journal of Economic Education* 25(2):121–133.

Koretz, Daniel M. 2002. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity." *The Journal of Human Resources* 37(4):752–777.

Ladd, Helen F. 2001. "School-based Educational Accountability Systems: The Promise and the Pitfalls." *National Tax Journal* 54(2):385–400.

Murnane, Richard J., and Frank Levy. 2001. "Will Standards-based Reforms Improve the

Education of Students of Color?" *National Tax Journal* 54(2):401–15.

OECD. 2001. "Knowledge and Skills for life - first results from PISA 2000."

Sotaridona, Leonardo S, and Rob R Meijer. 2002. "Statistical Properties of the K-Index for Detecting Answer Copying." *Journal of Educational Measurement* 39(2):115–132.

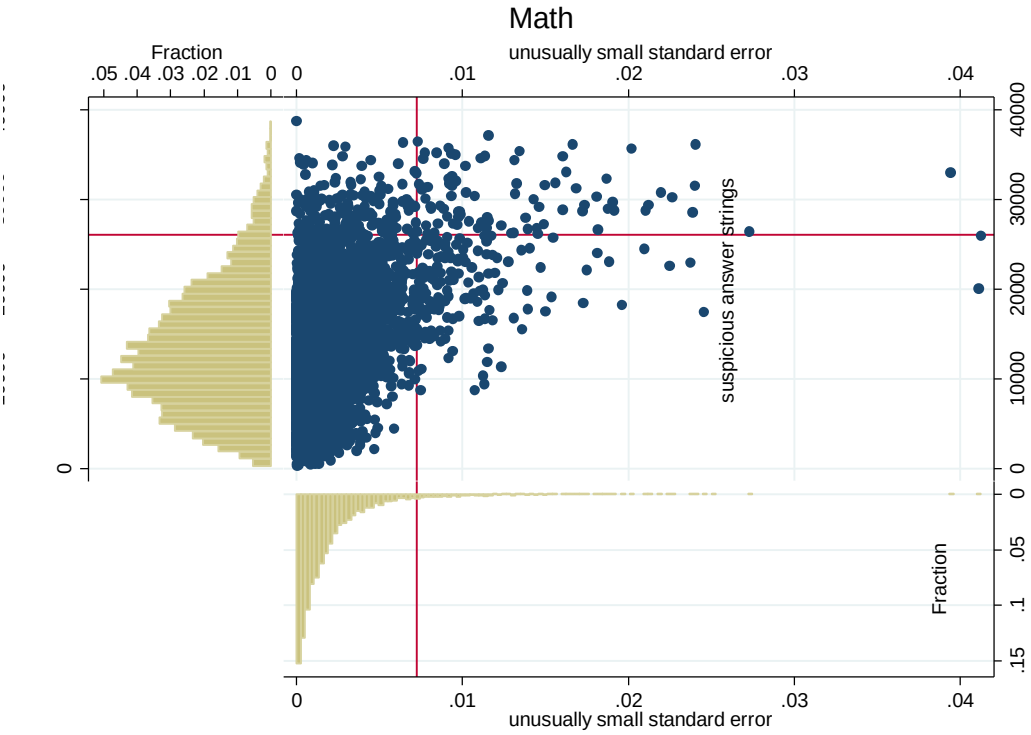
Whitley, Bernard E., Amanda Bichlmeier Nelson, and Curtis J. Jones. 1999. "Gender Differences in Cheating Attitudes and Classroom Cheating Behavior: A Meta-Analysis." *Sex Roles* 41(9):657–680.

Wollack, James A. 1997. "A Nominal Response Model Approach for Detecting Answer Copying." *Applied Psychological Measurement* 21(4):307–320.

Wollack, James A. 2003. "Comparison of Answer Copying Indices with Real Data." *Journal of Educational Measurement* 40(3):189–205.

Figures

Figure 1 – Distribution of *unusually small standard error* and *suspicious answer strings*



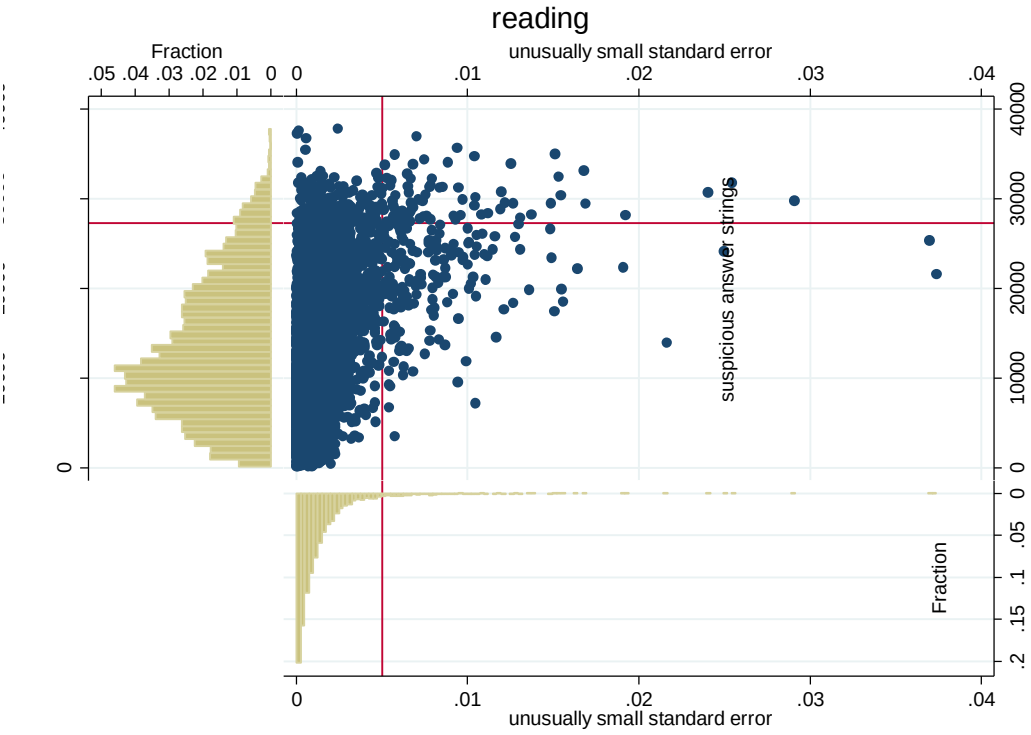
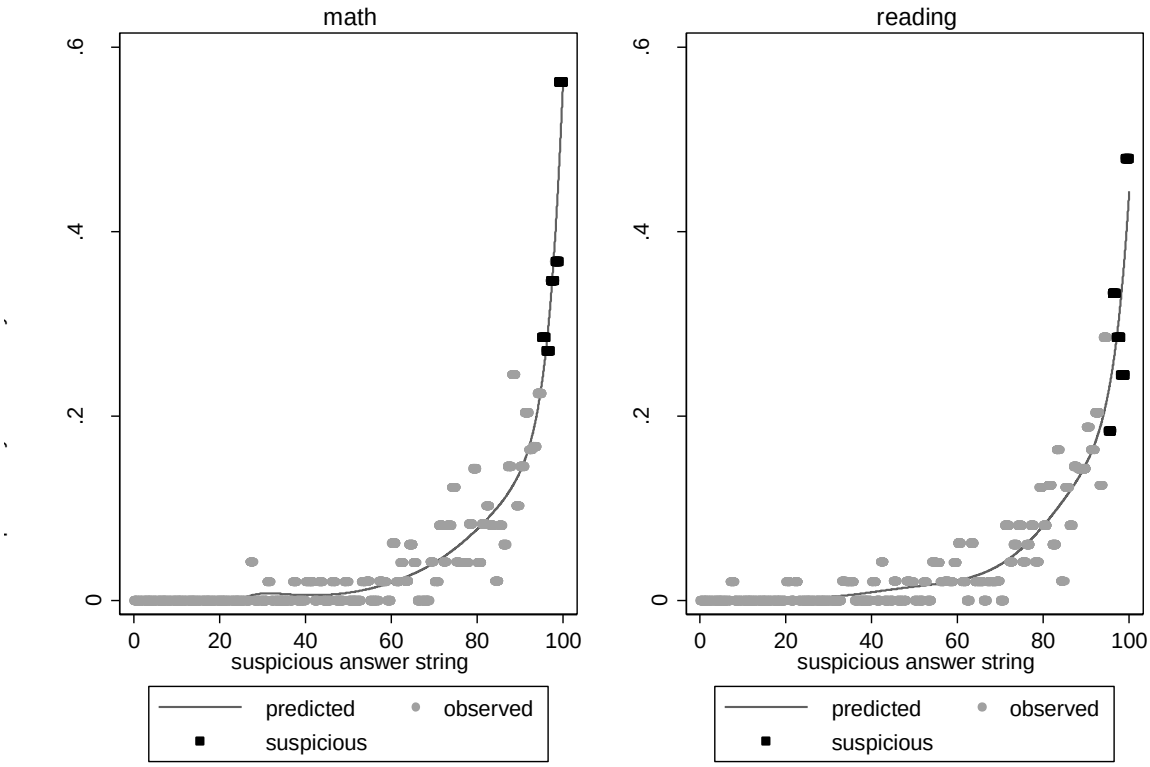


Figure 2 - Relationship between unusually small standard error and suspicious answer strings



Reproduction of Jacob and Levitt (2003a) figure II on the Hungarian NABC data: The horizontal axis reflects a classroom's percentile rank in the distribution of suspicious answer strings (*indicator2*) with zero representing the least suspicious classroom and one representing the most suspicious classroom. The vertical axis is the probability that a classroom will be above the 95th percentile on the measure of unusually small standard error (*indicator1*). The circles in the figure represent averages from 100 equally spaced cells along the x-axis. The predicted line is based on a probit model estimated with seventh order polynomials in the suspicious string measure.

Tables

Table 1 – The official NABC database

	4th grade	6th grade	8th grade	10th grade
2003	0	20 students from every school	0	20 students from each track from each school
2004	0	20 students from every school	20 students from every school	20 students from each track from each school
2006	full cohort	every student from a sample of 195 schools	full cohort	30 students from each track from each teaching site
2007	full cohort	every student from a sample of 200 schools	full cohort	30 students from each track from each teaching site
2008*	every student from a sample of 200 schools	full cohort	full cohort	full cohort
2009*	every student from a sample of 200 schools	full cohort	full cohort	full cohort
2010*	every student from a sample of 200 schools	full cohort	full cohort	full cohort
2011*	every student from a sample of 200 schools	full cohort	full cohort	full cohort

* Permanent individual identification numbers are available

Note: arrow shows the cohort being studied in this paper

Table 2 – Percent cheating on mathematics test 2008/8th grade – Jacob-Levitt method

Cutoff for suspicious answer strings (indicator 2)	**Cutoff for unusually small standard error (indicator 1)**		
Math	**90th percentile**	**95th percentile**	**99th percentile**
90th percentile	4,31%	2,73%	0,76%
95th percentile	2,59%	1,83%	0,62%
99th percentile	0,64%	0,55%	0,16%
Reading	**90th percentile**	**95th percentile**	**99th percentile**
90th percentile	3,92%	2,49%	0,60%
95th percentile	2,26%	1,52%	0,39%
99th percentile	0,55%	0,47%	0,10%

Table3 – Relationship of schools under the national minimum and cheating-suspicion.

		Suspicious class			
		All classes		Classes under the average mean score	
		Math	Read	Math	Read
Under the national	Math 2007	-0.00140		0.00017	
		(0.66)		(0.09)	
	Read 2007		0.00536		0.01386*
			(1.29)		(1.83)
	Observations	4439	4439	2708	2450
	Fifth order polynomial of class size is controlled for.				
	Marginal effects are shown for a class over the national minimum with class size 20				
	Absolute value of z statistics, clustered on school level in parentheses				
	* significant at 10%; ** significant at 5%; *** significant at 1%				

Table 4 – Relationship between outlier schools and cheating-suspicion.

	Suspicious class	
	Math	Read
Outlier in 2008	0.08104***	0.07327***
	(4.93)	(4.44)
Socio-economic status (mean)	-0.00423***	-0.00187***
	(4.65)	(3.78)
Observations	4878	4878
Fifth order polynomial of class size is controlled for.		
Marginal effects from a probit regression are shown for a non outlier, average SES class with class size 20		
Absolute value of z statistics, clustered on school level in parentheses		
* significant at 10%; ** significant at 5%; *** significant at 1%		

Table 5 – Relationship between outlier schools and cheating-suspicion. Linear regression.

	Test score point 2010	
	Math	**Reading**
Suspicious primary class	-85.089***	-50.001***
	(13.15)	(6.50)
Math test score, 2008	0.525***	0.143***
	(101.74)	(39.24)
Reading test score, 2008	0.155***	0.501***
	(39.16)	(124.70)
socio-economic status	9.602***	6.645***
	(8.75)	(6.28)
sex	-68.945***	24.005***
	(63.63)	(22.80)
Secondary level class FE	y	y
Constant	570.613***	597.961***
	(73.35)	(92.33)
Observations	79962	79972
R-squared	0.72	0.73
Robust t statistics, clustered on the 10th grade class level in parentheses		
* significant at 10%; ** significant at 5%; *** significant at 1%		

Table 6 – Relationship between outlier schools and cheating-suspicion. Propensity score matching.

	Average Treatment Effect	**# of controls (students in non-suspicious classes)**	**# of treatment (students in suspicious classes)**	**t-stat**	**std. err.**
Math	-69.17	54652	1239	-5.82	11.87
Read	-42.82	60291	975	-3.52	12.16