The impact of alternative formats for formative assessment on summative educational outcomes

Nicky Nicholls*

August 4, 2017

Incomplete draft version: Do not quote!!!

Abstract

We investigate the impact of formative assessment format (online versus in-class) on summative assessment performance (semester test/exam results). We conducted a controlled field experiment with students in a post-graduate micro-economics course in South Africa, where computer use is less commonplace for many students than it is in developed countries. Students were required to take formative assessments either online or onsite (in class) at different points in the course. We use difference-in-differences (DD) analysis to compare the impact of online versus onsite formative instruments on summative (test/exam) performance, where earlier test performance provides a pre-treatment baseline and final test performance gives the post-treatment result. Our results suggest that the formative assessment format does not have a significant impact on summative assessment results. Given the advantages of online instruments for increased teaching and discussion time as well as reduced grading time for faculty, a shift to greater use of online formative assessment is recommended.

Keywords: Higher education; Online assessment; Difference-in-differences
JEL Classification Numbers: A20, C90, I23

*Department of Economics, University of Pretoria, South Africa. E-mail: nicky.nicholls@up.ac.za

1 Background

Online assessments are being used with increasing frequency, particularly with the growth of MOOCs (Massive Open Online Courses) and other online/distance courses being offered by many insitutions. Many universities (including University of Pretoria, where the current research is conducted) are increasingly taking a blended/hybrid approach to education whereby face-to-face and online tools are used in tandem as important elements in a curriculum. The pedagogical advantages of such an approach are discussed in Bocconi and Trentin, 2014. In the South African context, practical considerations are added to pedagogical motivations for embracing online learning: wide scale student protests in 2016 led to South African universities having to adapt many courses to increase the share of online versus in-class assessment.

Recall that formative assessments aim to assist students with learning (these might take the form of homework exercises or problem sets, for example) while summative assessments aim to evaluate learning (for example, a test or examination on a section of course material). Adding online formative (or low mark summative) quizzes to face-to-face courses has generally positive (Angus & Watson, 2009; Dobson, 2008; see also Peng, 2007 and references therein) or at least neutral (Galizzi, 2010) outcomes in terms of overall (summative) course performance .

Further, online instruments of formative assessment offer some advantages over on-site (in-class) equivalents: the most obvious is that an online assessment taken outside of class time enables class time to be used more productively for more in-depth treatment of course content and/or transmission of the lecturer's personal expertise (Peng, 2007, Trentin, 2010). The online environment also offers opportunities for collaborative learning processes, enabling students to collaborate across time and location (Trentin, 2010).

Other advantages apply to online in comparison to both onsite and to pen and paper homework alternatives: online testing allows for immediate feedback, such that students can learn from errors rather than practicing incorrect approaches (Bonham, Deardorff & Beichner, 2003). The automated grading made possible in many online assessment tools reduces the grading burden on faculty, as well as reducing costs incurred in paying grading assistants (Bonham et al, 2003; Engelbrecht & Harding, 2004). Further, many platforms for online assessments (including Blackboard, the platform used by University of Pretoria) offer overviews of response accuracy and discrimination potential of questions, which are helpful to teachers in deciding which kinds of questions to reuse and which to discard. The ease of seeing which questions posed particular difficulty for students in online platforms is highlighted by Kennelly, Considine and Flannery (2011),

who note that subsequent lectures could focus on enhancing student understanding of challenging areas.

Other authors have considered the risk of increased cheating with online assessments (Hollister & Berenson, 2009; Kibble, 2007; Ladyshewsky, 2015; Peng, 2007).

It seems that online formative assessment offers two possible outcomes: collaboration made possible by unproctored online assessment (versus more formal in class assessment) might increase engagement with the course material through discussions and debates, enhancing student mastery of the course subject matter (measured in summative assessments). Alternatively, the lack of supervision in online assessment might increase copying of answers without engaging with course material, negatively affecting mastery of subject matter.

Although a number of studies (see Ladyshewsky, 2015 for an overview) have compared results from online and onsite (in-class) assessments, fewer have considered the impact of different formative assessment formats (online versus onsite) on summative results. Those that have done this have introduced other variables that make a direct comparison difficult. Mitra and Barua (2015) found higher formative scores for online tests, but no significant difference in mean summative results. Their study, however compared a single in-class formative assessment to three online formative assessments, where participation in the formative assessments was voluntary. Maclean and McKeown (2013) compared formative online quizzes to take-home assignments, finding no difference in overall summative (exam) performance. In their study the weighting in the final course mark differed (10% weight for the online assignments versus 30% weight for the written ones) and the online assessment tool was used 5 years after the written one, such that 2001-3 data for written assignments was compared to 2008-2011 data for online ones. Further, multiple attempts were allowed on the online quizzes, where the written assignments had no such option. Angus and Watson (2009) introduced regular low-mark online formative assessments, and note positive impacts on summative results. They do not, however, include a control alternative to the online assessments. As they note in their paper, it is possible that similar results would have been achieved had the formative assessments been on paper.

Other studies compare the impact of pen-and-paper versus online homework assignments on final results: Palocsay and Stevens (2008) found no significant difference in exam results; Bonham et al. (2003) found no significant difference in course outcomes between students who were given homework assignents on pen and paper and those who completed homework assignments online. Bonham et al., however, allowed multiple attempts for the online group, and partial marks for the pen-and-paper group, making direct comparisons difficult. Lee, Courtney and Balassi (2010) compare performance

of students using online versus pen and paper homework assignments in different years (where one year had pen and paper homework, and another had online homework), and find no significant difference in summative performance. Kennelly et al. (2011) use an interesting design where students are grouped to write some assignments online and others on paper, with the medium used for the assignment being linked to performance on the associated portion of the summative assessment. They note that the assignments counted for a high proportion of the course grade (25%), making the assignments more of a summative instrument than a formative one. They too find no significant difference in exam outcomes. Most of these authors who found no significant differences in summative outcomes note that even if online assessment doesn't significantly improve summative outcomes, the fact that these outcomes are maintained is a positive finding, given the significant cost and time savings inherent in online homework.

Our paper attempts a more controlled comparison between summative outcomes with online versus onsite formative assessments. Specifically, we randomly assign students in a post-graduate class to either an online or onsite formative assessment group. The groups are structured such that students who are in the online group in the first half of the course write onsite assessments in the second half, and vice versa. In order to control for any pre-existing differences between the groups, we use a difference-in-differences (DD) approach, whereby we compare the change in each group's performance between a pre-formative assessment instrument (previous course grade/summative test) and a post-formative assessment instrument (summative test). By having each student participate in online assessments for part of the course and onsite assessments in another part, and by varying this across students, we are able to control for both the possibility that one part of the course is easier than another; and for the possibility that one group of students performs better than another. By making the formative assessments compulsory for course completion, we ensured that all students participated in these formative assessments.

The South African context with respect to online course materials is somewhat different from that in countries such as the USA and the UK, where much of the existing research is based. Specifically, high levels of inequality in South Africa and particularly in government funding for different South African schools make access to and familiarity with computers and the internet more variable for South African students than might be the case for students in developed countries. Internet penetration in South Africa is reported to be 52%, versus 89% for the USA and 93% for the UK (internetlivestats.com, 2016). Our study therefore also asks whether the common finding of no significant difference in summative outcomes with online versus onsite formative assessment applies in this context.

Our findings align well with the existing research: we do not find significant differences in summative results based on whether formative assessments were conducted onsite or online. Given the many benefits of online assessment discussed previously, a shift from onsite to online formative (and/or low grade summative) assessments would seem justified. Since our sample size was fairly small, and since we did note some patterns of cheating that had not been seen in previous research, we plan to repeat our experiment with a bigger sample and with some appropriate controls in place for cheating. In this way we will be able to confirm whether our findings are robust to such variations.

The remainder of our paper is structured as follows: Section 2 discusses our study design; Section 3 considers our main findings; and Section 4 concludes.

2 Study Design

Students in the 2017 microeconomics honours course (first year of post-graduate work) at University of Pretoria were randomly divided into two groups. Our final sample size (after excluding students who did not consistently write assessments in the allocated online/in-class set-up described below) was 26 for Group 1 and 24 for Group 2. To comply with the ethics requirements of the University of Pretoria and to give all students equal opportunities with online and onsite assessment, each group wrote two formative assessments online and two onsite (in class). In order to incentivise all students to put appropriate effort into the assessments, students were told that the assessments would count towards their final grade. In line with other studies (Angus & Watson, 2009; Kibble, 2007) each assessment counted ~2% of the final grade for the course. Since these assessments did form a (small) part of students' course grade, they are, strictly speaking, summative instruments. However, since the goal of these assessments was formative in nature, that is, to assist students in evaluating their own learning and to prepare them for the primary summative assessments, we treat them as formative instruments.

We did not explicitly discuss the possibility of collaborative work in the online assessments with students, precisely because we were interested to see whether or not online assessment would result in more collaboration. Indeed 53% of students in an anonymous survey following the assessment treatment claimed to have done some work in groups for the online assessments. Bocconi and Trentin (2014) differentiate between learning spaces (onsite versus online) and learning processes (individual versus collaborative), and suggest that well constructed blending learning would allow for the intersection at different times of all combinations of these dimensions. Our design aimed at merging

the online space with a choice of learning process; while the onsite space only allowed for individual learning. To allow more opportunity for collaboration, and to leverage the benefit of flexibility with respect to time, the online group had the option to take the quiz at any point in a one week window. The onsite (in class) group had the full week to prepare for the assessment, but actually answered the assessment in class. In order to ensure that no differences existed in the assessment standard and in the questions included in this formative assessment, we used the same assessment questions for all students. This approach, of course, risked sharing of answers by online group students with onsite students. Having seen no obvious evidence of this in the first 2 formative assessments (indeed, anonymous feedback saw only 4 people from the first onsite group claiming to have been involved in answer sharing of this kind), we continued with this approach. Both online and onsite students received marks only for the correct answer to the MCQ problems, and neither group had the option of multiple attempts.

The formative assessments were structured such that the first group had two online formative assessments followed by the mid-term summative assessment; and then two in-class formative assessments followed by the final summative assessment. The second group started with two in-class formative assessments followed by the mid-term summative assessment; and then two online formative assessments followed by the final summative assessment. Each formative assessment covered approximately two weeks of content, while each summative assessment covered half of the course content. All students wrote the same summative assessments, and both summative assessments took place as formal proctored tests.

This set-up allowed for a difference-in-differences (DD) analysis of the two parts of the course. The DD analysis was conducted as follows: the course was split into two halves, such that each student had a pre- and post-formative assessment score in each half. The previous (undergraduate) microeconomics course mark was used as the pre-treatment score to compare against the mid-term post-treatment score for the first half of the course; and the mid-term score was used as the pre-treatment score to compare against the final post-treatment score for the second half of the course. We then introduced a "time" variable to capture the difference between pre- (time = 0) and post- (time=1) formative assessment scores; and a "treated" variable taking the value 1 where students had online formative assessments in a given part of the course, and 0 where students had onsite formative assessments. The DD variable itself ("did" in our results) captures the interaction of time and treatment, allowing us to compare the changes in summative outcomes with online versus onsite formative assessment. Since most students found the material covered in the second part of the course more challenging than that in the first part, we also introduced a "sem" variable, taking the value of 0 for the first half of

the semester and 1 for the second half. Finally, we included dummy variables for race, gender and whether or not students worked during their studies.

That is, we estimated the following regression:

$$Outcome = \beta_0 + \beta_1 Time + \beta_2 Treated + \beta_3 Did + \beta_4 White + \beta_5 Female + \beta_6 StudyFT + \beta_7 Sem$$

3 Findings

We start by confirming that our random assignment indeed resulted in similar groups: the test scores from previous microeconomics courses showed no significant differences between the group means (Mann-Whitney rank-sum test between scores for the two groups gives $z = 1.46, p = 0.14$). Any differences between the groups should, however, be accounted for in the DD approach used.

To ensure that our findings were robust to different econometric specifications, we considered three different DD regression approaches: OLS, with standard errors clustered on individuals; and panel regressions with fixed and random effects. Table 1 below shows the results from student performance on summative assessment measures. The difference-in-differences ("did") measure comparing expected mean changes in results pre-formative assessment to post-formative assessment was not significant. This indicates that the shift to online formative assessment did not alter students' performance in the summative assessments. Variation in summative results is explained by the "time" variable (most students found the honours course more challenging than their ingoing undergraduate course; and found topics in the mid-term exam easier to grasp than those in the final exam. As a result, the post-formative assessment performance was significantly weaker on average, as captured in the significant negative coefficient on the "time" variable. Recall that the "sem" variable was used as a dummy to isolate students assessed between mid-term and final exam (as opposed to between previous microeconomics performance and mid-term exam). The significant coefficient on this variable captures the increased difficulty of the course material in the second half of the course. We also see that white students achieved significantly better on average than non-white students. More than 20 years after apartheid policies were removed in South Africa, educational disparities persist between races, largely related to white South Africans continuing to have better access to historically better funded schooling prior to university.

These results were robust to the inclusion of other variables, such as a dummy variable to identify likely cheaters, whether students preferred online or onsite assessment, and whether or not students were assessed using their preferred assessment method (onsite or online).

Table 1: Difference-in-differences regression results

	OLS	Panel FE	Panel RE
time	-9.779***	-9.495***	-9.581***
	(2.114)	(2.072)	(2.092)
treated	1.021	1.305	1.219
	(2.076)	(2.029)	(2.050)
did	-2.616	-3.184	-3.012
	(3.565)	(3.525)	(3.546)
white	8.369**	0.000	8.264**
	(3.391)	(.)	(3.391)
female	-1.963	0.000	-2.159
	(2.925)	(.)	(2.939)
studyful~e	1.638	0.000	1.769
	(1.895)	(.)	(1.877)
sem	-11.105***	-11.105***	-11.105***
	(0.870)	(0.863)	(0.870)
_cons	65.692***	67.951***	65.691***
	(2.177)	(1.475)	(2.156)
N	198	198	198
adj. R-sq	0.311	0.530	

Standard errors in parentheses
* p<.10, ** p<.05, *** p<.01

3.1 Cheating

At the end of the module, students were asked to provide anonymous feedback on whether they had collaborated with others and whether they had cheated in any way on the formative assessments. Previous research has considered differences in test scores for unsupervised online tests versus supervised in-class tests. Some authors found no significant differences in scores (Ladyshewsky, 2015; Peng, 2007); while others found higher performance on the supervised tests than on the online tests (Fask, Englander & Wang, 2014). Hollister and Berenson (2009) compare supervised online to unsupervised online tests, and find similar overall performance with no evidence of cheating behaviour in the unsupervised environment. Kibble (2007) finds more cheating with higher incentives (higher share of final grades for online assessments).

Our anonymous reports showed that, while collaborative work was common in both online groups (~53% overall), very few members of the first onsite group tried to obtain questions from the online group. However, 57% of the second onsite group admitted to obtaining questions from the online group ahead of the assessment.

Students' claims to have cheated in this way are supported by comparisons between the formative assessment results for onsite and online assessment. We compared assessment scores for the online and onsite groups for each of the four formative assessments. Where the group with only four people reporting receiving assessment questions before the onsite assessment were writing the assessment onsite, we saw no significant difference between online and onsite performance on either of the assessments (two sample Wilcoxon rank-sum tests: $p > 0.37$). Conversely, where the group with high reported cheating on the onsite assessment were being assessed onsite, we saw significantly higher results in the onsite (cheating) group for both assessments (two sample Wilcoxon rank-sum tests: $p < 0.001$).

We do not have a conclusive reason for this difference between the groups. Given the relatively small sample sizes, it is possible that a higher propensity to cheat where possible was a pre-existing characteristic in this group. An alternative explanation is that this group had onsite assessments closer to the end of the course, where students were more focused on their final grade and on ensuring that this was high enough to pass the course. Perhaps these students were more willing to use a shortcut to ensure they had a passing grade.

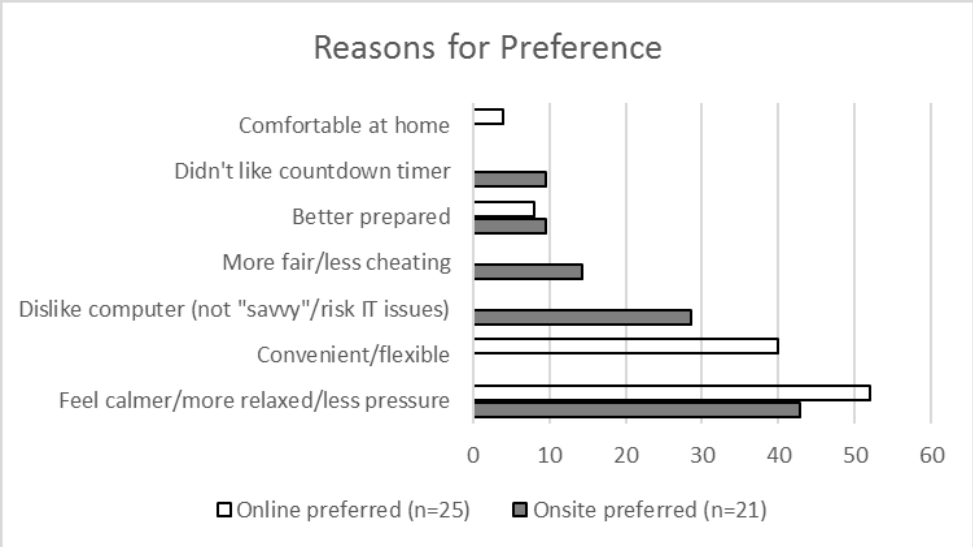
Although the anonymous feedback did not allow us to identify the specific students who had cheated in this way, we used a combination of above average results in the onsite formative assessment and below average results in both the mid-term and final exams as a proxy for likely cheaters, and reran our regressions including a "cheat" dummy variable (taking the value 1 for suspected cheaters). The results are included in appendix A. The overall results are in line with those reported in this paper[1]. Of note, this was not an ideal proxy: although thirteen students in this group reported cheating in this way, we were only able to isolate seven likely cheaters.

3.2 Perceptions of online assessment

At the end of the semester, students completed a questionnaire where they were asked about their preferences for online versus onsite assessment. They were asked to give reasons for their preference as open-ended feedback, which was subsequently coded into categories. These responses are shown in Figure 1.

[1]Limiting our analysis to only the first half of the semester, such that the data with cheating in the onsite formative assessment is not included, does not alter our main findings.

Figure 1: Reported reasons for assessment medium preference (percent)



Four students reported indifference between the two approaches, with the rest being divided between online (25 preferred) and onsite (21 preferred), as seen in Figure 1. While studies in developed countries reported very positive perceptions of online homework (Lee et al., 2010; Galizzi, 2010), our results show only a slightly higher proportion of students preferring online assessment. While each group reported feeling less pressure or better prepared with their preferred assessment instrument, those favouring onsite assessment frequently reported disliking aspects of the computer set-up (not feeling comfortable with computers or not being "computer savvy"). The responses pertaining to less cheating in class show that not all students felt comfortable collaborating in the online assessments. This is also reflected in the anonymous feedback on cheating, where uncertainty regarding the acceptability of collaboration was evident in responses to the question, "did you 'cheat' on this test in some way?": one student answered, "if working together constitutes cheating, then yes".

Despite the difference in preferences, students being assessed in their less preferred medium (online versus onsite) perform no worse than those being assessed in their preferred medium (two sample Wilcoxon rank-sum tests). This is true of both summative assessments ($p = 0.26$) and formative assessments ($p \geqslant 0.8$).

4 Discussion

Our findings align well with existing research, which largely points to no significant differences in summative outcomes between online and onsite formative assessments:

Students completing online assessments did not do better or worse on their mid-term or final exam than those completing onsite (in-class) assessments.

We do, however, see mixed responses to the online assessments from South African students: close to half of our sample expressed a preference for onsite over online assessments, partly due to not feeling comfortable with computer-based work. These preferences do not translate into better or worse performance when students are assessed in their preferred medium.

By having the same students take online and onsite formative assessments at different points in the course, and by employing a difference-in-differences methodology to control for ingoing differences between groups, our work provides a robustness check for some of the existing studies, where fewer such controls were in place. Further, our work demonstrates that findings on the impact of formative assessment medium on summative results from developed countries are also applicable to a developing country.

We therefore join other authors in recommending a shift towards increased online formative assessments in lieu of onsite ones. Although online formative assessments do not improve summative outcomes when compared to onsite assessments, these do as well as onsite assessments, while also offering many benefits (immediate feedback, time savings for faculty, allowing for class time to be better used to further educational goals, among others).

We plan to address a few points in follow-on research: first, we will remove the cheating opportunity where one group gains answers from another group. This can be done without removing the opportunities for collaboration in the online environment, by requiring the online group to take the online assessment at the same time as the onsite group (this time limit is easily enforced on the Blackboard platform used by University of Pretoria). Second, while some students had no qualms about collaborating on online assignments, others expressed uncertainty about whether this was acceptable. In future research we plan to explicitly allow for collaboration in the online context in order to better understand whether permitting collaboration improves summative performance (by improving subject matter mastery through discussion and engagement with the material in a group context) or whether summative performance decreases on average (allowing collaboration might simply increase "free-riding": some students doing the work while others simply copy solutions).

References

Angus, S. D., & Watson, J. (2009). Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology*, **40(2)**, 255-272.

Bocconi, S., & Trentin, G. (2014). Modelling blended solutions for higher education: teaching, learning, and assessment in the network and mobile technology era. *Educational Research and Evaluation*, **20(7-8)**, 516-535.

Bonham, S. W., Deardorff, D. L., & Beichner, R. J. (2003). Comparison of student performance using web and paper-based homework in college-level physics. *Journal of Research in Science Teaching*, **40(10)**, 1050-1071.

Dobson, J. L. (2008). The use of formative online quizzes to enhance class preparation and scores on summative exams. *Advances in Physiology Education,* **32(4)**, 297-302.

Engelbrecht, J., & Harding, A. (2004). Combining online and paper assessment in a web-based course in undergraduate mathematics. *Journal of Computers in Mathematics and Science Teaching,* **23**, 217-232.

Fask, A., Englander, F., & Wang, Z. (2014). Do online exams facilitate cheating? An experiment designed to separate possible cheating from the effect of the online test taking environment. *Journal of Academic Ethics,* **12(2)**, 101-112.

Galizzi, M. (2010). An assessment of the impact of online quizzes and textbook resources on students' learning. *International Review of Economics Education*, **9(1),** 31-43.

Hollister, K. K., & Berenson, M. L. (2009). Proctored versus unproctored online exams: Studying the impact of exam environment on student performance. *Decision Sciences Journal of Innovative Education*, **7(1)**, 271-294.

Kennelly, B., Considine, J., & Flannery, D. (2011). Online assignments in economics: A test of their effectiveness. *The Journal of Economic Education*, **42(2)**, 136-146.

Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. *Advances in Physiology Education*, **31(3)**, 253-260.

Ladyshewsky, R. K. (2015). Post-graduate student performance in 'supervised in-class' vs.'unsupervised online'multiple choice tests: implications for cheating and test security. *Assessment & Evaluation in Higher Education,* **40(7)**, 883-897.

Lee, W., Courtney, R. H., & Balassi, S. J. (2010). Do online homework tools improve student results in principles of microeconomics courses?. *The American Economic Review*, **100(2)**, 283-286.

Maclean, G., & McKeown, P. (2013). Comparing online quizzes and take-home assignments as formative assessments in a 100-level economics course. *New Zealand Economic Papers,* **47(3)**, 245-256.

Mitra, N. K., & Barua, A. (2015). Effect of online formative assessment on summative performance in integrated musculoskeletal system module. *BMC medical education*, **15(1)**, 29.

Palocsay, S. W., & Stevens, S. P. (2008). A Study of the Effectiveness of Web-Based Homework in Teaching Undergraduate Business Statistics. *Decision Sciences Journal of Innovative Education*, **6(2)**, 213-232.

Peng, Z. (2007). Giving Online Quizzes in Corporate Finance and Investments for a Better Use of Seat Time. *Journal of Educators Online,* **4(2)**, n2.

Trentin, G. (2010). *Networked Collaborative Learning: social interaction and active learning.* Elsevier.

http://www.internetlivestats.com/internet-users-by-country/ Accessed on 20 July 2017

Appendix A:

Table 2 reports the three DD regression approaches including a dummy variable for likely cheaters: those students who achieved above average results in the onsite formative assessment but below average results in both summative assessments. The cheating variable is significant, suggesting a negative link between cheating and final results. Recall, however, that the cheating variable was defined as low performance on summative assessments with high performance on formative assessments. As a result, those students with a value of 1 for the cheat dummy will by definition have lower summative results. We therefore consider the impact of adding the cheat dummy on the other coefficients: the time, treated and DD coefficients remain similar to before, suggesting no change to our overall findings.

Table 2: Difference-in-differences regressions, controlling for likely cheating

	OLS	Panel FE	Panel RE
time	-9.769***	-9.495***	-9.592***
	(2.115)	(2.072)	(2.097)
treated	1.031	1.305	1.208
	(2.076)	(2.029)	(2.055)
did	-2.637	-3.184	-2.990
	(3.567)	(3.525)	(3.554)
white	7.408**	0.000	7.312**
	(3.234)	(.)	(3.234)
female	0.585	0.000	0.417
	(2.906)	(.)	(2.915)
studyful~e	-3.039	0.000	-2.934
	(3.326)	(.)	(3.395)
cheat	-12.931***	0.000	-12.964***
	(2.639)	(.)	(2.636)
sem	-11.105***	-11.105***	-11.105***
	(0.872)	(0.863)	(0.872)
_cons	70.878***	67.951***	70.891***
	(3.577)	(1.475)	(3.644)
N	198	198	198
adj. R-sq	0.390	0.530	

Standard errors in parentheses
* p<.10, ** p<.05, *** p<.01