Does the repetitive use of the same test in consecutive examination sessions facilitate cheating?

András István Kun

Department of Organization Sciences, Faculty of Economics and Business, University of Debrecen, Debrecen, Hungary

Correspondence details

E-mail: andras.kun@econ.unideb.hu; mobile: +36(20)5610912; postal address: University of Debrecen Faculty of Economics and Business, 138 Böszörményi út, Debrecen, Hungary, H-4028.

Notes on contributor

András István Kun is associate professor in the Department of Organization Sciences at the University of Debrecen. His research interests include economics of education; labour economics; and human resource management.

Does the repetitive use of the same test in consecutive examination sessions facilitate cheating?

This paper introduces an empirical study testing three kinds of bias in higher education student assessment. All of them are connected to the repetitive use of the same test questions which may facilitate academic cheating. The 'same tests effect' may appear if two or more groups of students are writing the same test one after the other and, as a result, a statistically significant improvement is detectable in the test scores of the second student group. The 'revealed sameness effect' is the impact of informing the students in some way that the test questions will be repeated. The 'self selection effect' arises when the students choose their examination turn themselves and this boosts their measured performance. The present study examines the three effects with independent *t*-tests and linear regression models on samples of 1221, 235 and 201 students (in this order), from 4 business courses in 6 academic semesters. The results do not support the 'same test effect', but support the 'revealed sameness effect' and the 'self selection effect'.

Keywords: educational assessment; assessment bias; higher education; academic cheating

Introduction

When a course is attended by many students, it may happen – because the number of seats or the capacity of the staff is low – that the students will sit a given examination in two or more consecutive groups (turns). On these occasions a question arises: is it acceptable for both (or all) turns to take the same test? The greatest advantages of using the same test are that it saves time for the test makers and avoids the measurement bias caused by test differences. The strongest counter-argument, however, is that it facilitates cheating, through the possible information flow from the predecessor turn students to their friends in the succeeding turn. Cheating should be taken seriously in business higher education for the reason that some empirical research has found its occurrence the most frequent here among the various academic areas (Carauna and Ewing 2000; McCabe, Butterfield, and Trevino, 2006).

Although not every study supports this phenomenon (Klein et al. 2007; Simkin and McLeod 2010), the higher acceptance of some forms of cheating is present even in some of the latter studies (Klein et al. 2007). If there is a possibility to obtain and to use information from the preceding turn, the succeeding turn's mean score will be systematically and significantly higher, which leads to a measurement bias.

With appropriate security measures (e.g. students leaving the first turn cannot meet students waiting for the second, a higher number of examination supervisors, severe punishment for any cheats caught) the threat of this type of cheating can be eliminated or highly reduced, since the role of situational factors in motivation to cheat was found significant by many previous studies (Corcoran and Rotter 1987; McCabe et al. 2001b; Abdolmohammadi and Baker 2008). Nevertheless, security has costs (management costs, lost time, inconvenience caused to students etc.) that should be kept as low as possible, until the point where it does not increase the possibility of cheating significantly. Summing up, it is reasonable to ask whether the usage of the same tests in two or more consecutive turns of examinees induces – when accompanied by only minimal cost security measures – a significant increase (a bias) in the subsequent turn's average test score. In other words, does it lead to a significant increase in cheating?

This question is interesting mainly from a practical, classroom or examination management point of view: is it possible for an examiner to spare time and energy – and thus to save costs for the institution – and to avoid the bias from different test sheets through the repetitive use of the same test in multiple turns without inducing cheating (which leads to another type of assessment bias) in the succeeding turn?

In this paper three main hypotheses will be tested. The first one is as follows:

H1: If the same test is given to two consecutive student groups within a short time range, and if only minimal security measures are adopted (assigning students into turns in name order, ensuring they have no information on the similarity of the tests, ensuring that only a short time elapses between the examination turns, but allowing students finishing earlier to leave the site), a statistically significant increase will be manifest in the succeeding group's test scores.

The increase in the test scores of the second turn originating from the identical nature of the tests will be referred to as 'same test effect'.

Since the sample provides the possibility to estimate the effect of violating some of the minimal security measures in H1, two additional statements are formulated for examination.

H2: If the students are informed about the possibility of repeating the test questions, than the 'same test effect' will be stronger, or will arise even if it did not exist otherwise.

This will be called the 'revealed sameness effect' in the study.

H3: In cases in which students are assigned to examination turns not randomly or not in the alphabetic order of their names, but are free to select the session they prefer, it is expected that the subsequent turns will significantly outperform preceding turns.

The effect mentioned in H3 will be referred to as the 'self selection effect', and this is a special case of the general self selection bias (Heckman 1979, 153-154). Many other effects can interfere in the self selection effect besides extra information, such as better chances for collaborative cheating (cheating in teams), differences in motivation (those willingly choosing a prior turn may prefer finishing the examination earlier to getting a better mark), the fact that the duration of the preceding turn serves as extra time to study for the following turn, or, if the better students like the company of other good performers and by chance they are grouped together in the succeeding turns, this will also increase the turn's average grade.

The following section provides a brief overview of the related literature. The third section introduces the sample and the research methods, while the fourth describes the results of the empirical research. The paper closes with a Conclusions section.

Background

The most important purposes of any student assessment are to diagnose students' strengths and weaknesses, monitor students' progress, assign grades representing the students' knowledge or skills, evaluate teachers and instructional methods, and influence public perceptions about educational effectiveness (see e.g. Popham 2003). If test results are not reliable (the test instrument is not consistent or not stable, and does not provide similar results under similar conditions) or not valid (the test instrument is measuring something other than it was designed for), none of these aims can be achieved (for a more detailed explanation of validity and reliability see e.g. Kumar (2005, 152-159)). The present study focuses only on a special area of the validity problem(s): if student assessment tests are repeated in multiple examination turns, do the test scores measure the true performance or the test outcome modified by belonging to one of the turns?

Validity weakening of a test might be traced back to individual, random sources (mistakes), or to contextual, systematic, non-random sources (errors). The latter are somehow connected to the selected testing method or arrangements. Systematic errors are termed biases. Assessment bias can be any element of the assessment process that hinders a student or a group of students in reaching a test score reflecting their true abilities, skills and knowledge (or which helps them get a too high score compared to their abilities). More accurately, in the words of Reynolds and Suzuki (2003, 87-88), test bias is a deviation from examinees' real level of performance. Some authors (e.g. Popham 2012) mistakenly emphasise socio-economic status, race, gender, and religion as the main, or only, sources of such biases, although many other factors can have a strong negative or positive impact on students' performance. This is likely because these authors confuse bias with fairness and offensiveness, even though – unlike bias – these two particular attributes of an examination are not statistically estimated characteristics of two or multiple groups but are rather moral or legal issues based on values and opinions (Reynolds and Suzuki 2003, 87).

From the point of view of the methodology followed by this paper, the taxonomy of Van de Vijver and Tanzer (2004, 124-127) is a more useful way of classifying the important sources of assessment biases. They distinguished between three kinds of biases, with related sources: (1) construct bias (e.g. overlapping definitions and poor sampling of relevant factors associated with the construct), (2) method bias (e.g. insufficient sample, instrument, or administration), and (3) item bias (e.g. poor translation, low familiarity of the item content, cultural context that distorts perception). A subgroup of method biases is called 'situational bias' by Jensen (1980), who defines it as 'influences in the test situation, but independent of the test itself, that may bias test scores' (Jensen 1980, 377). Characteristics of test settings, behaviour and characteristics of the examiners can play a role in this bias.

According to Van de Vijver and Tanzer's (2004) typology cited above, all three hypotheses in this study (H1, H2, H3) focus on effects that count as method biases, and also fulfil Jensen's (1980) definition of situational biases; however, the situations are different in each of the three examined effects. The bias caused by the 'same test effect' has its source in the potential informational asymmetry between the two student groups created by a specific type of cheating: the preceding group has zero information about the test questions, while the succeeding group is potentially better informed (by the first test writers). The source of the 'revealed sameness effect' bias can be both informational (if there are three consecutive turns, experience from the first two can indicate the similarity of the test papers) and/or connected to insufficient administration (students somehow acquire information on the similarity directly from the examiner). The 'self selection effect' leads to insufficient samples and also enables cooperative cheating within the turns, caused by non-randomness in the students' assignment to turns.

It should be emphasized that the supposed information flow from the first turn to the second is only potential, and as such, it need not be present in the case of every participant in order to support our hypotheses. The above mentioned three effects concern the 'representative student' or, in other words, the total populations of each turn. Moreover, because the three effects discussed are rooted in a special method of cheating – maybe its most important feature is that the subsequent turns have the chance to increase their test results, but this automatically decreases the relative test outcome of the predecessor turns –, they are highly dependent on educational and social incentives (Gino et al. 2009; Dee and Jacob 2012), on interpersonal connections and organizational (sub)culture (McCabe et al. 2001a; Paccagnella and Sestita 2014) or on national differences (Magnus et al. 2002). Whilst this paper focuses on the existence of the three hypotheses within one institution only, motivational and cultural background will not be analysed. However, future studies with samples from more than one institution must take these factors into consideration.

Sample and method

This paper analyses 13 examinations – real, regular examinations – from 4 different business subjects over 6 university semesters (the total number of the tests is 1497) at the University of Debrecen, Hungary, where students took their examinations in a minimum of 2 (in one case 3) consecutive turns. The test papers were identical in all turns of the same examination and there was a chance to exchange information between the students from different turns for at least some minutes; meanwhile they were not informed about the identical nature of the tests. Students were assigned to turns according to alphabetical name order, in such a way that in every examination the same proportion of the students were assigned to each turn (one half in each if there were two turns, one third if there were three); thus the group at the beginning of the alphabet was always assigned to the first turn and the group at the end to the last turn.

The sample only includes examinations where the number of the participants could be analysed and the turn numbers were registered. The test questions were multiple choice and calculations, thus the scoring is measured on a ratio scale and fairly objective (compared to essays or oral examinations). Calculation problems consist of 1 to 4 separate exercises (type and number vary from examination to examination) scored independently; however, for the current analysis only the total of these scores is used. Even if an examination contained both multiple choice and calculation type exercises, it was only the total test score which affected the course grade. Since mid-term and end-term examinations from multiple majors and years (on both bachelor and master levels) are involved, standardised scores – by examinations and test versions – are used in the analyses. Description of the sample by examinations is presented in Table 1.

[Table 1 near here]

Analysing the H1 and H2 hypotheses, only those cases were involved where the students took the examination in the turn they were assigned to: the number is shown in Table 1 column *N* in parentheses. Only the data from the first two turns are included. The data necessary to examine the three main hypotheses are available from every examination, but the additional data for every single analysis is not, thus the sample size differs analysis by analysis. Descriptive statistics are presented for all the 13 examinations, but *t*-tests and regression analyses will be conducted on the totalized sample only (not examination by examination). Variables used are shown in Table 2. The dependent is always the total test scores standardised by examinations and by test versions to avoid bias arising from the different tests. Since in the standardisation process all students' results in the same examination are used, including those not taking the examination in the turn originally assigned to them (to enable a comparison to be made with those not changing their turn) in some analyses the mean of the STDSCORE may differ from zero.

[Table 2 near here]

Results

Testing the same test effect

This section examines the H1 hypothesis through descriptive statistics, independent samples *t*-tests and linear regression analysis on the sample of those cases where the students took their examination in the assigned turn. Table 3 presents the descriptive statistics by examinations and by sex.

[Table 3 near here]

As Table 3 suggests there is only a very slight difference (lower than 1.75% of the standard deviation) between the two turns in their average score; however, this is far from convincing: there are many examinations where one or the other sex performed better in the first turn (according to both the median and the mean). In the total sample the mean of the first turn is better for women (with 0.0556) and worse for men (with 0.1485). For the two genders together there were 6 examinations where the median was higher in the first turn and 3 where it was higher in the second turn. For the mean the same ratio is 6 to 7.

Following this, the significance of the mean differences of the STDSCORE for the total of the examinations is examined by independent samples *t*-tests. Table 4 presents the output of the *t*-tests for both sexes and the total sample. According to these results the equality of the STDSCOREs of the two turns cannot be rejected at the 10% significance level. The 'same test effect' is not supported.

[Table 4 near here]

Linear regression models are built – putting all the available additional information (major, year, examination trial) to use – to deepen our understanding. Table 5 presents the descriptives of the variables used in the linear regression models.

[Table 5 near here]

Table 6 contains a 'maximal' model (Model1), where all available independent variables are present and a 'minimal' (Model2) where only those that are significant at at least 5% plus TURN2. TURN2 is entered the model after all the other variables, to measure the R^2 change it causes.

[Table 6 near here]

The conclusion of the regression analysis echoes those from the *t*-tests: belonging to the second turn has no significant explanatory power over the (standardised) test scores. The models themselves have very low R^2 values and the distributions of their residuals are not normal (the value of the Shapiro–Wilk statistic is 0.9919 for Model 1 and 0.9920 for Model 2). Yet, these results are enough to state that the 'same test effect' is not supported even when all available data are taken into consideration. Supplementary findings are that female, higher year students can expect better scores, those who have already tried to pass many times are likely get worse results, and there is a significant difference between some majors (IBE, BIOLOGY, ML, VHE_BAM, BIBSC, GEOGRAPHY, ENVIRONMENT, EARTH) and the (referential) BAM students.

Testing the 'revealed sameness effect'

The H2 hypothesis is examined in this section with independent samples *t*-tests and linear regression analysis. The sample is constructed from the three turns of the fifth examination (see Table 1). As previously, only those cases join the sample where the students took their examination at the originally assigned turn. The existence of the third turn provides the possibility to test what happens if the students find out – after the second turn – that the tests are likely to be identical in the subsequent turns. Table 7 contains the means and standard deviations of the measured variables by turns (genders and other demographic groups are not separated).

[Table 7 near here]

As can be seen from the data of Table 7, the average test score of the third turn is significantly higher than the averages in the first or second turns, or their mean. This significance was tested by independent samples *t*-tests and later by linear regression. Results of the *t*-test are presented in Tables 8-9-10.

[Table 8 near here]

[Table 9 near here]

[Table 10 near here]

Standardised scores are significantly higher in the third turn in every comparison (at least at the 5% level); however, this is not true for the male subsample. As in the case of H1, the effect was tested in the presence of other independent variables, too. Table 11 contains a 'maximal' Model1, where every variable is included and a 'minimal' Model2 where only significant ones were (at a 5%level).

[Table 11 near here]

Both regression models, similarly to the *t*-tests, support the H2 hypothesis. Since the third turn where the students have the chance to acquire information on the identical nature of the test in different turns significantly outperforms the predecessor turns, but the second turn – where there was no information on the identicality – does not, the 'revealed sameness effect' cannot be rejected. Additionally, introducing TURN3 into the models also significantly increases their explanatory power, even if it is still low, and the residuals do not have normal distributions (the value of the Shapiro–Wilk statistic is 0.9467 for Model 1 and 0.9460 for Model 2). As a secondary result it was found that the students' sex and major affect their performance, but the number of examination attempts does not.

Testing the 'self selection effect'

In the examination of the last hypothesis the sample was provided by those 201 students who did not take their examination in the turn originally assigned to them according to the name order (the reason for this deviation was not recorded). The distribution of these students by examination is shown in Table 2, and means and standard deviations by variables are presented in Table 12.

[Table 12 near here]

Testing methods were independent samples *t*-test and linear regression. Table 13 presents the *t*-test statistics that show that the standardised test scores in the second turn are significantly higher (at the 1% level) in the total sample and in the male subsample. In the female subsample the difference is not significant, although its direction is the same. Implicitly, the 'same test effect' arises if the students select (or modify) their turns themselves, unlike in the case of those students who do not, which should be the case if H3 were true.

[Table 13 near here]

Obviously this result may have multiple causes, as was already discussed in the introductory section. For a better understanding of these causes it is useful to know which turn-changing students perform better compared to their non-changing fellows: those moving from the second turn to the first, losing potential information (first type changers), or those moving from the first to the second, potentially gaining extra information (second type). If the second type outperforms the first type of turn-changing students, this supports H1. However, to support H3, too, the second type should also be better than those students who are originally assigned to the second turn, and take the examination there (assuming that they are better at making use of the potential extra information, and this is the reason for the change in turn), while the first type of changers should not be better than those originally assigned to the first turn who stay there, or, if there is a difference, it should be significantly smaller than in the second turn (because in this case the better performance could not have derived from the extra information available). Table 14 contains the statistics of the independent samples *t*-tests where the turn-changing students (both the ones changing to the second turn instead of the first and the others moving in the other direction).

[Table 14 near here]

We can see that if a student changed his/her turn from first to second, this increases the expected (standardized, totalized) test score; however, this result is significant only at the 10% level for males. At the same time, students changing their turns from second to first can be expected to perform worse (at a 5% significance level) than those originally assigned to this turn. These results support H3, based on the logic described above.

As in the case of the first two hypotheses, regression analyses are used to test the 'self selection effect' in the presence of other available independent variables. Similarly to the previous hypotheses, two models are presented in Table 15 (Model 1 contains every available explanatory variable, while Model 2 retains only those that are significant at a 10% level). TURN2 is entered into both models.

[Table 15 near here]

As Table 15 suggests, we can reject the idea that TURN2 has no effect on the STDSCORE (at a significance level of 5% for Model 1 and 10% for Model 2), which means that the 'self selection effect' seems to exist and H3 has to be accepted. It does not affect the overall conclusion, but it must be noted that Model 1 is not significant, the R^2 is low and the residuals do not have a normal distribution in any of the models (the value of the Shapiro–Wilk statistic is 0.9778 for Model 1 and 0.9722 for Model 2). This is not connected to H3 but it is interesting that in the sample of turn-changer students, sex has no significant effect on performance, in contrast to the previous samples.

Conclusions

This paper attempted to test three hypotheses about the effect on the test results of using the same tests for two or more sequential examination turns. All hypotheses concentrate on a single aspect of the possible bias arising from the repeated use of the tests in cases when the time interval between the consecutive turns is relatively small. All the three examined aspects of the possible bias can be connected to the increased opportunity to cheat – obtaining information from the preceding turns – for students in the later turns. The first hypothesis (the same test effect, H1) formed a statement on the pure form of the expected phenomenon: the second (and third etc.) turn will get a higher test score if only minimal security measures are adapted (students are assigned to a turn according to the alphabetic order of their names and they have no information about the repetition of the test). This hypothesis was examined with independent samples *t*-tests and linear regression models and was rejected. The conclusion we can draw is twofold: either 1) the above mentioned security measures are enough to avoid an increase in cheating, and thus in the bias of the examination results, or 2) the 'same test effect' does not exist, even if there are no security measures at all. The following two hypotheses help to choose between these two possible answers.

The H2 hypothesis (the revealed sameness effect) states that if students can figure out that the test will be repeated in the following turn, that turn's results will be upwardly biased (even compared to the impact of the 'same test effect'), thus 'revealed sameness' triggers or intensifies the 'same test effect'. This was tested in a sample where 3 turns existed; thus after the second, the students could find out that the test had not been changed. Hence the second turn's results were not upwardly biased, but the third's were, and so H2 should be accepted. The statistical methodology used was the same as for H1. The conclusion is evident, showing that test repetition is not acceptable because it leads to increased bias (through cheating) in the test results. It must be noted that the current paper examined H3 only in the case in which information was revealed only after the second turn (by experience); thus it is not necessary that the phenomenon is the same if the information about the non-changing tests came after the first turn or was already available at the beginning.

The final hypothesis, H3 (the self selection effect), focuses on the importance of the other security measure. If the students can decide on the turn in which they take their examination, this will lead to a bias in the results, to the second turn's advantage. To test this, a sample was formed from a group of students (from many different examinations) who changed their originally assigned turn of their own accord. With similar statistical methods as before, it was found that H3 should be accepted, as it also enhances the 'same test effect'. Those students who changed their original (first) turn to a later (second) one were able to significantly increase their results compared to those who moved to the first turn from the second, and, in the case of male students, even to those who were originally assigned to the second turn. We can state that the practice of self selecting examination turns should be also restricted.

Since H2 and H3 are supported, we can also judge the first conclusion of H1 to be the more probable. Summarizing the results, it can be stated that using the same test multiple times is not likely to lead to an increased bias or cheating if 1) students are not informed about the repetition and 2) students are assigned to the examination turns by a rule that excludes students' self selection (this can be achieved randomly, or through the use of alphabetic order etc.). However, if these minimal requirements are not met, the 'same test effect' will occur, biasing the later turns' test scores upward.

Given that the analyses are composed of tests written at only one university, the generalizability of the results and the conclusions are limited. Another important limitation is the limited variety of the accessible explanatory variables: personal (e.g. intelligence, emotional intelligence, personality traits), socio-economic (e.g. family background, income, wealth, organisational and national culture) or situational factors may modify the three examined effects.

Furthermore, the non-randomness of the alphabetic rule used to assign students to examination turns might have biased the results by concealing some or all the impacts of the 'same test effect' and/or the 'revealed sameness effect' as well as 'self selection effects'. Some studies have found evidence for the effect of better labour market outcomes resulting from the first letter of one's surname coming closer to the beginning of the alphabet (e.g. Einav and Yariv 2006; Jurajda and Munich 2014). Nevertheless, both Regéczy (2011) and Jurajda and Munich (2014) have identified that individuals with surnames whose initial letter is closer to the end of the alphabet tend to perform better in schools. If the positive correlation between the alphabetical order of the surnames and learning success exists in higher education, too, then the alphabetical bias should not conceal but, on the contrary strengthen all the three effects in our sample (by increasing the test points of the students in the second and third turns).

However, the relatively large (at least in the case of H1) sample consisting of multiple subsamples (13 examinations, 4 subjects and 6 semesters) may provide an appropriate heterogeneity, allowing us to assume that these outcomes are still not highly specific, and they are reliable at least for the business subjects at the given university.

The above mentioned limits enable us to point to some further research directions, too. It may be interesting to test the hypotheses of this paper at different organisations with different cultural and institutional backgrounds in order to answer the question on generalizability. The involvement of additional personal, socio-economic and situational factors may reveal the real explanatory power of the three analysed effects.

Acknowledgements

The author would like to thank the help given by those who contributed to the data collection and to the discussion of the research problems and results. In surname order: Edit Barizsné Hadházi, Éva Farkas, Anna Kapitány, and Ágnes Kotsis.

Notes on contributor

András István Kun is associate professor in the Department of Organization Sciences at the University of Debrecen. His research interests include economics of education; labour economics; and human resource management.

References

Abdolmohammadi, M. J. and C. R. Baker. 2008. "Moral Reasoning and Questionable Behavior." *The CPA Journal* 78 (11): 58–60.

Carauna, A. and M. T. Ewing. 2000. "The Effect of Anomie on Academic Dishonesty of University Students." *The International Journal of Educational Management* 14 (1): 23–29.

Corcoran, K. J. and J. B. Rotter. 1987. "Morality-Conscience Guilt Scale as a Predictor of Ethical Behavior in a Cheating Situation among College Females." *Journal of General Psychology* 114 (2): 117–123.

Dee, T. S. and B. A. Jacob. 2012. "Rational Ignorance in Education: A Field Experiment in Student Plagiarism." *Journal of Human Resources* 47 (2): 397–434.

Einav, L. and L. Yariv. 2006. "What's in a Surname? The Effects of Surname Initials on Academic Success." *The Journal of Economic Perspectives* 20(1): 175–188.

Gino, F., S. Ayal, and D. Ariely. 2009. "Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel." *Psychological Science* 20 (3): 393–398.

Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–162.

Jensen, A. R. 1980. *Bias in Mental Testing*. New York, NY: Free Press.

Jurajda, S. and D. Munich. 2014. “Alphabetical Order Effects in School Admissions.” *CERGE-EI Working Paper Series No. 509*. Available at SSRN: http://ssrn.com/abstract=2399686.

Klein, H. A., N. M. Levenburg, Marie McKendall, and William Mothersell. 2007. “Cheating During the College Years: How do Business School Students Compare?” *Journal of Business Ethics* 72 (2): 197–206.

Kumar, R. 2005. *Research Methodology*. London: SAGE Publications.

Magnus, J. R., V. M. Polterovich, D. L. Danilov, and A. V. Savvateev. 2002. “Tolerance of Cheating: An Analysis Across Countries.” *The Journal of Economic Education* 33 (2): 125–135.

McCabe, D., L. K. Treviño, and K. Butterfield. 2001a. “Cheating in Academic Institutions: A Decade of Research.” *Ethics and Behavior* 11 (3): 219–232.

McCabe, D. L., L. K. Treviño, and K. D. Butterfield. 2001b. “Dishonesty in Academic Environments: The Influence of Peer Reporting Requirements.” *Journal of Higher Education* 72 (1): 29–45.

McCabe, D. L., K. D. Butterfield, and L. K. Treviño. 2006. “Academic Dishonesty in Graduate Business Programs: Prevalence, Causes, and Proposed Action.” *Academy of Management Learning & Education* 5 (3): 294–305.

Paccagnella, M., and P. Sestita. 2014. “School Cheating and Social Capital.” *Education Economics* 22 (4): 367–388.

Popham, J. W. 2003. *What Every Teacher Should Know About Educational Assessment*. Boston, MA: Pearson Education.

Popham, J. W. 2012. *Assessment Bias: How to Banish It?* 2nd ed. Boston, MA: Pearson Education.

Regéczy, M. 2011. *What's in a Name? Effect of Alphabetical Ordering on Students' Achievement: Example from Czech Elementary and Lower Secondary Schools*. Prague: Faculty of Economics and Public Administration.

Reynolds, C. R., and L. A. Suzuki. 2003. "Bias in Psychological Assessment. An Empirical Review and Recommendations." Chapter 4 In *Handbook of Psychology. Volume 10. Assessment Psychology,* edited by Irving B. Weiner, Donald K. Freedheim, and John R. Graham, 82–113, Hobokes, NJ: Wiles & Sons.

Simkin, M. G., and A. McLeod. 2010. "Why Do College Students Cheat?" *Journal of Business Ethics* 94 (3): 441–453.

Van de Vijver, F., and N. K. Tanzer. 2004. "Bias and Equivalence in Cross-Cultural Assessment: An Overview." *European Review of Applied Psychology* 54 (2): 119–135.

Table 1. Description of tests involved in the research as samples in chronological order

Exam-ination	Course title	Type[a]	Date	N[b]	Max. pts.[c]	TbT[d]	Test versions[e]	MC ratio[f]	CA ratio[g]
1	Human Resource Management	mid	2/11/2010	157 (155)	40	28	4	1/1	0/1
2	Human Resource Management	end	15/12/2010	146 (143)	30	no data	2	1/1	0/1
3	Production and Process Management	mid	3/11/2010	71 (69)	35	app. 5	2	10/35	24.5/35
4	Production and Process Management	end	15/12/2010	63 (60)	35	app. 13	2	10/35	24.5/35
5	Introduction to Economics and Management	end	14/12/2011	168+75[h] (160+75)	60	app. 10	1	1/1	0/1
6	Management of Value Creating Processes	mid	30/5/2012	93 (93)	100	14	3	4/10	6/10
7	Management of Value Creating Processes	mid	25/3/2013	37 (37)	40	app. 10	2	4/10	6/10
8	Human Resource Management	mid	30/10/2013	112 (112)	30	app. 15	2	2/3	1/3
9	Human Resource Management	end	11/12/2013	86 (86)	30	13 & 26[i]	2	2/3	1/3
10	Human Resource Management	mid & end[j]	18/12/2013	72 (70)	30	app. 15	4	2/3	1/3
11	Introduction to Economics and Management	end	16/12/2013	250 (80)	60	0	2	1/1	0/1
12	Management of Value Creating Processes	mid	2/4/2014	84 (78)	80	app. 10	2	3/8	5/8
13	Management of Value Creating Processes	mid	21/5/2014	83 (78)	80	app. 10	4	3/8	5/8
Total of the 1st and 2nd turns				1422 (1221)					
Total of all 3 turns				1497 (1296)					

[a] mid = mid-term test; end = end-term test
[b] Total number of students (capita); number of those writing the test in their assigned turn in parentheses
[c] Maximum test score (point)
[d] Time between the end of the first and the beginning of the second turn (minutes)
[e] Number of different test versions used with the students within a turn
[f] Maximum score for multiple choice questions per maximum score for the whole test.
[g] Maximum score for calculation exercises per maximum score for the whole test.
[h] 168 is the number of students in the 1st and 2nd turns only. There was also a third turn with 75 (75) students.
[i] In this subsample, in the second turn 30 students entered the room after 13 minutes, and 17 after 26 minutes, due to the lack of test papers, because of an incorrect prediction of the number of test takers
[j] Some of the students retook the mid-term test from 30/10/2013, others retook the end-term test from 11/12/2013.

Table 2. Explanation of variables

Variable name	Explanation
SCORE	The student's unstandardized, totalized test score.
STDSCORE	The student's totalized test score standardized by tests and test versions.
SEX	1 if the student is female, 0 if male.
YEAR	University year of the student.
TRIAL	Number of times the student tried to pass the examination.
TURN2	If the student takes the examination in the first turn it is 0, if in the second it is 1.
TURN3	If the student takes the examination in one of the first two turns it is 0, if in the third it is 1.
TURN13	If the student takes the examination in the first turn it is 0, if in the third it is 1.
TURN23	If the student takes the examination in the second turn it is 0, if in the third it is 1.
GROUP	If the student wrote test version 'A' it is 0, for test version 'B' it is 1, for test version 'C' it is 2, and for test version 'D' it is 3.
BAM	1 if the student's major is 'BA in Business Administration and Management', 0 if not.
IBE	1 if the student's major is 'BA in International Business Economics', 0 otherwise.
EBA	1 if the student's major is 'MSc in Economics and Business Administration', 0 otherwise. This is a pre-Bologna process, 5 year university major.
VHE_BAM	1 if the student's major is 'Vocational Higher Education in Business Administration and Management', 0 if not.
ML	1 if the student's major is 'MSc in Management and Leadership', 0 otherwise.
BIBSC	1 if the student's major is 'BSc in Business Informatics', 0 otherwise.
BIMSC	1 if the student's major is 'MSc in Business Informatics', 0 otherwise.
BIOLOGY	1 if the student's major is 'BSc in Biology', 0 otherwise.
PHYSICS	1 if the student's major is 'BSc in Physics', 0 otherwise.
GEOGRAPHY	1 if the student's major is 'BSc in Geography', 0 otherwise.
CHEMISTRY	1 if the student's major is 'BSc in Chemistry', 0 otherwise.
ENVIRONMENT	1 if the student's major is 'BSc in Environmental Studies', 0 otherwise.
MATHS	1 if the student's major is 'BSc in Mathematics', 0 otherwise.
EARTH	1 if the student's major is 'BSc in Earth Science', 0 otherwise.

Table 3. Descriptive statistics of the standardised test scores by examinations and by turns

Sample	Sex	Turns (according to TURN2)									
		1st					2nd				
		N	STDSCORE				*N*	STDSCORE			
			median	mean	std. dev.	s.e.m.		median	mean	std. dev.	s.e.m.
1	M	24	-0.237	-0.366	1.158	0.236	27	-0.308	-0.382	0.933	0.180
	F	55	0.110	0.133	0.805	0.109	49	0.337	0.237	1.057	0.151
	Total	79	0.068	-0.019	0.947	0.107	76	0.068	0.017	1.052	0.121
2	M	23	-0.130	-0.097	1.333	0.278	20	0.106	0.189	0.722	0.161
	F	51	0.246	-0.022	1.041	0.146	49	0.116	0.039	0.885	0.127
	Total	74	0.246	-0.046	1.131	0.132	69	0.116	0.083	0.839	0.101
3	M	8	-0.814	-0.655	0.495	0.175	13	0.126	-0.358	0.839	0.233
	F	24	0.743	0.383	1.174	0.240	24	0.185	0.021	0.882	0.180
	Total	32	-0.086	0.124	1.134	0.201	37	0.185	-0.112	0.875	0.144
4	M	10	-0.333	-0.424	0.723	0.229	7	0.550	0.322	1.076	0.407
	F	21	0.841	0.508	0.986	0.215	22	-0.392	-0.394	0.901	0.192
	Total	31	0.315	0.208	1.000	0.180	29	-0.242	-0.221	0.977	0.181
5	M	31	-0.295	-0.439	1.087	0.195	34	-0.295	-0.202	0.936	0.161
	F	49	0.340	0.009	0.825	0.118	46	0.340	-0.026	1.085	0.160
	Total	80	0.022	-0.164	0.954	0.107	80	0.022	-0.101	1.022	0.114
6	M	7	0.290	0.082	1.176	0.444	19	0.124	0.193	0.914	0.210
	F	27	0.008	-0.091	1.108	0.213	40	-0.081	-0.045	0.930	0.147
	Total	34	0.047	-0.055	1.106	0.190	59	-0.005	0.032	0.923	0.120
7	M	6	0.354	0.193	0.778	0.318	6	0.565	0.217	1.143	0.467
	F	10	-0.042	0.015	0.815	0.258	15	0.014	-0.174	1.150	0.297
	Total	16	0.059	0.082	0.780	0.195	21	0.014	-0.062	1.134	0.247
8	M	26	0.054	-0.160	1.041	0.204	21	-0.124	-0.144	0.958	0.209
	F	29	0.018	0.006	1.041	0.193	36	0.231	0.196	0.952	0.159
	Total	55	0.018	-0.073	1.035	0.140	57	0.018	0.070	0.960	0.127
9	M	21	-0.202	-0.125	0.855	0.187	18	-0.163	-0.242	1.208	0.285
	F	23	0.384	-0.032	0.969	0.202	24	0.328	0.321	0.933	0.190
	Total	44	0.010	-0.076	0.907	0.137	42	0.308	0.080	1.083	0.167
10	M	16	0.045	-0.201	1.113	0.278	21	0.072	-0.077	0.859	0.188
	F	17	0.222	-0.068	1.091	0.265	16	0.667	0.373	0.889	0.222
	Total	33	0.204	-0.133	1.086	0.189	37	0.339	0.117	0.889	0.146
11	M	19	-0.624	-0.511	0.809	0.186	14	0.124	-0.001	0.945	0.253
	F	23	0.676	0.652	0.784	0.164	24	0.134	0.152	0.960	0.196
	Total	42	0.134	0.126	0.980	0.151	38	0.134	0.096	0.944	0.153
12	M	7	-0.133	-0.235	0.602	0.227	10	-0.328	0.203	1.216	0.385
	F	36	-0.005	0.084	1.029	0.172	25	-0.547	-0.154	0.936	0.187
	Total	43	-0.026	0.032	0.974	0.149	35	-0.547	-0.052	1.018	0.172
13	M	19	0.347	0.382	0.924	0.212	12	-0.028	0.271	1.081	0.312
	F	24	-0.495	-0.147	0.840	0.171	23	-0.457	-0.208	1.046	0.218
	Total	43	-0.254	0.086	0.907	0.138	35	-0.336	-0.044	1.067	0.180
Total	M	217	-0.202	-0.214	1.026	0.070	222	-0.010	-0.066	0.964	0.065
	F	389	0.142	0.095	0.968	0.049	393	0.116	0.039	0.979	0.049
	Total	606	0.022	-0.016	1.001	0.041	615	0.068	0.002	0.974	0.039

Note: Tests were taken in their assigned turn. *N* = number of students, s.e.m. = standard error of the mean, M = male, F = female. Median and mean values are underlined in the turn where they are greater.

Table 4. Statistics of the independent samples *t*-tests

Sample	Levene's test		*t*-test						
	F	sig.	*t*	df.	sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Male	0.463	0.497	-1.563	437.000	0.119	-0.149	0.095	-0.335	0.038
Female	0.033	0.855	0.797	780.000	0.426	0.056	0.070	-0.081	0.192
Total	0.324	0.569	-0.305	1219.000	0.760	-0.017	0.057	-0.128	0.094

Note: Dependent is STDSCORE, grouping variable is TURN2.

Table 5. Descriptive statistics of the variables in the linear regression models

Variable	Mean	Std. Dev.	Variable	Mean	Std. Dev.
STDSCORE	-0.007	0.987	ML	0.088	0.284
SEX	0.641	0.480	BIOLOGY	0.063	0.243
YEAR	2.310	0.783	PHYSICS	0.003	0.050
TRIAL	1.016	0.142	GEOGRAPHY	0.030	0.172
EBA	0.120	0.326	CHEMISTRY	0.045	0.208
BAM	0.419	0.494	ENVIRONMENT	0.021	0.142
IBE	0.145	0.352	MATHS	0.017	0.130
VHE_BAM	0.011	0.103	EARTH	0.018	0.133
BIBSC	0.003	0.057	TURN2	0.504	0.500
BIMSC	0.017	0.130			

N = 1221

Table 6. Statistics of two linear regression models

Variables	Dependent: STDSCORE			
	Model 1		Model 2	
	Coefficients	*t*-statistics	Coefficients	*t*-statistics
Constant	0.367	1.521	–	–
SEX	0.143	2.418**	0.168	2.976***
YEAR	0.101	1.973*	0.148	3.616***
TRIAL	-0.735	-3.725***	-0.546	-5.055***
EBA	-0.020	-0.211	–	–
IBE	0.222	2.627***	0.268	3.374***
VHE_BAM	-0.674	-2.397**	-0.567	-2.058**
BIBSC	-1.018	-2.109**	-0.956	-1.983**
BIMSC	-0.186	-0.812	–	–
ML	0.223	1.727*	0.332	2.889***
BIOLOGY	0.455	3.770***	0.517	4.458***
PHYSICS	-0.365	-0.657	–	–
GEOGRAPHY	-0.310	-1.865*	–	–
CHEMISTRY	0.105	0.760	–	–
ENVIRONMENT	-0.392	-1.967**	–	–
MATHS	-0.117	-0.538	–	–
EARTH	-0.520	-2.443**	-0.443	-2.122**
TURN2	0.018	0.325	0.023	0.423
R^2	0.067		0.059	
adjusted R^2	0.054		0.052	
F	5.093***		7.628***	
N	1221		1221	
R^2 change[a]	0.000		0.000	
F change[a]	0.105		0.179	

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

[a] after entering TURN2 into the models

Note: Since it has the greatest frequency, the omitted major became the BAM in Model 1.

Table 7. Descriptive statistics of the variables of all 3 turns of the sample

Variable	Turn	Mean	Std. Dev.	Variable	Turn	Mean	Std. Dev.
SCORE	1st	40.825	6.010	PHYSICS	1st	–	–
	2nd	41.225	6.440		2nd	0.013	0.112
	3rd	43.653	6.276		3rd	–	–
	Total	41.864	6.339		Total	0.004	0.065
STDSCORE	1st	-0.164	0.954	GEOGRAPHY	1st	0.150	0.359
	2nd	-0.101	1.022		2nd	0.163	0.371
	3rd	0.285	0.996		3rd	0.120	0.327
	Total	0.001	1.006		Total	0.145	0.353
SEX	1st	0.613	0.490	CHEMISTRY	1st	0.313	0.466
	2nd	0.575	0.498		2nd	0.200	0.403
	3rd	0.640	0.483		3rd	0.240	0.430
	Total	0.609	0.489		Total	0.251	0.435
YEAR	1st	2.013	0.251	ENVIRONMENT	1st	0.088	0.284
	2nd	2.100	0.377		2nd	0.175	0.382
	3rd	2.013	0.116		3rd	0.040	0.197
	Total	2.043	0.274		Total	0.102	0.304
TRIAL	1st	1.000	0.000	MATHS	1st	0.063	0.244
	2nd	1.000	0.000		2nd	0.063	0.244
	3rd	1.013	0.116		3rd	0.107	0.311
	Total	1.004	0.065		Total	0.077	0.267
BIOLOGY	1st	0.263	0.443	EARTH	1st	0.125	0.333
	2nd	0.288	0.456		2nd	0.100	0.302
	3rd	0.333	0.475		3rd	0.160	0.369
	Total	0.294	0.456		Total	0.128	0.334

$N = 235$ (80 in the first turn, 80 in the second turn, and 75 in the third turn)

Table 8. Results of the independent samples *t*-tests comparing the first turn to the third

Sample	Levene's test		*t*-test						
	F	sig.	*t*	df.	sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Male	0.002	0.965	-1.332	56.000	0.188	-0.390	0.293	-0.977	0.197
Female	1.291	0.259	-2.703^{***}	95.000	0.008	-0.463	0.171	-0.802	-0.123
Total	0.853	0.357	-2.866^{***}	153.000	0.005	-0.449	0.157	-0.758	-0.139

Dependent: STDSCORE, grouping variable: TURN13, $N = 155$

$^{***}p < 0.01$

Table 9. Results of the independent samples *t*-tests comparing the second turn to the third

Sample	Levene's test		*t*-test						
	F	sig.	*t*	df.	sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Male	0.186	0.668	-0.577	59.000	0.566	-0.154	0.266	-0.686	0.379
Female	7.848	0.006	-2.458***	85.774	0.016	-0.498	0.203	-0.901	-0.095
Total	3.311	0.071	-2.375***	153.000	0.019	-0.385	0.162	-0.706	-0.065

Dependent: STDSCORE, grouping variable: TURN23, $N = 155$

** $p < 0.05$

Table 10. Results of the independent samples *t*-tests comparing the merged first and second turns to the third turn

Sample	Levene's test		*t*-test						
	F	sig.	*t*	df.	sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Male	0.094	0.760	-1.108	90.000	0.271	-0.266	0.240	-0.744	0.211
Female	5.617	0.019	-3.032***	103.560	0.003	-0.480	0.158	-0.794	-0.166
Total	2.762	0.098	3.014***	233.000	0.003	0.417	0.138	0.144	0.690

Dependent: STDSCORE, grouping variable: TURN3, $N = 235$

*** $p < 0.01$

Table 11. Statistics of the linear regression models for H2

Variables	Dependent: STDSCORE			
	Model 1		Model 2	
	Coefficients	*t*-statistics	Coefficients	*t*-statistics
Constant	-0.648	-0.598	–	–
SEX	0.292	2.171**	0.271	2.796***
YEAR	0.303	1.313	–	–
TRIAL	0.058	0.061	–	–
PHYSICS	-1.334	-1.412	–	–
GEOGRAPHY	-0.787	-3.826***	-0.695	-4.245***
CHEMISTRY	-0.201	-1.190	–	–
ENVIRONMENT	-0.703	-3.117***	-0.622	-3.057***
MATHS	-0.626	-2.533**	-0.547	-2.355**
EARTH	-0.640	-3.062***	-0.571	-3.210***
TURN2	0.097	0.642	–	–
TURN3	0.418	2.756***	0.369	2.935***
R^2	0.182		0.161	
adjusted R^2	0.141		0.139	
F	4.495***		7.335***	
N	235		235	
R^2 change[a]	0.028		0.032	
F change[a]	7.595***		8.615***	

** $p < 0.05$; *** $p < 0.01$. [a] after entering TURN3 into the models.

Note: Since it has the greatest frequency, in Model 1 the omitted major became the BIOLOGY.

Table 12. Means and standard deviations

Variable	Mean	Std. Dev.	Variable	Mean	Std. Dev.
STDSCORE	-0.063	0.990	BIOLOGY	0.353	0.479
SEX	0.632	0.484	PHYSICS	0.030	0.171
YEAR	1.995	0.474	GEOGRAPHY	0.139	0.347
TRIAL	1.005	0.071	CHEMISTRY	0.179	0.384
EBA	0.005	0.071	ENVIRONMENT	0.030	0.171
BAM	0.045	0.207	MATHS	0.075	0.263
IBE	0.040	0.196	EARTH	0.075	0.263
BIMSC	0.005	0.071	TURN2	0.448	0.499
ML	0.020	0.140			

$N = 201$

Table 13. Differences between turns: independent samples *t*-tests (Dependent: STDSCORE)

Sample	Levene's test		*t*-test						
	F	sig.	*t*	df.	sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Male	0.011	0.917	-2.784***	72.000	0.007	-0.618	0.222	-1.061	-0.176
Female	0.130	0.719	-1.375	125.000	0.172	-0.243	0.177	-0.592	0.107
Total	0.163	0.687	-2.761***	199.000	0.006	-0.382	0.138	-0.654	-0.109

$N = 201$ (First turn: males = 43, females = 68, total = 111; second turn: males = 31, females = 59, total = 90).
*** $p<0.01$.

Table 14. Results of turn-changing vs. non-changing students: independent samples *t*-tests

Sample	Levene's test		*t*-test						
	F	sig.	*t*	df.	sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
ASSIGNED TO THE FIRST TURN BUT SAT THE EXAMINATION IN THE SECOND									
Male	0.003	0.960	1.712*	243.000	0.088	0.317	0.185	-0.048	0.681
Female	0.326	0.569	0.389	445.000	0.697	0.053	0.136	-0.215	0.321
Total	0.234	0.629	1.322	690.000	0.187	0.145	0.110	-0.070	0.361
ASSIGNED TO THE SECOND TURN BUT SAT THE EXAMINATION IN THE FIRST									
Male	0.042	0.837	-0.912	258.000	0.363	-0.155	0.170	-0.489	0.179
Female	1.248	0.265	-1.905*	455.000	0.058	-0.244	0.128	-0.496	0.008
Total	0.427	0.514	-2.119**	715.000	0.034	-0.219	0.103	-0.421	-0.016

Dependent: STDSCORE, $N = 1399$ (First to second turn: male changers = 31, male non-changers = 214, female changers = 59, female non-changers = 388, total changers = 90, total non-changers = 602; second to first turn: male changers = 43, males non-changers = 217, females changers = 68, females non-changers = 389, total changers = 111, total non-changers = 606).

* $p < 0.1$; ** $p<0.05$.

Table 15. Linear regression models for the same test effect among turn-changers

Variables	Dependent: STDSCORE			
	Model 1		Model 2	
	Coefficients	*t*-statistics	Coefficients	*t*-statistics
Constant	-0.542	-0.490	–	–
SEX	0.071	0.456	–	–
YEAR	0.038	0.230	–	–
TRIAL	0.403	0.403	–	–
EBA	-0.876	-0.880	–	–
BAM	-0.185	-0.504	–	–
IBE	-0.190	-0.499	–	–
BIMSC	-0.781	-0.774	–	–
ML	0.272	0.497	–	–
PHYSICS	-0.248	-0.585	–	–
GEOGRAPHY	-0.561	-2.426**	-0.569	-3.098***
CHEMISTRY	-0.249	-1.230	–	–
ENVIRONMENT	-0.588	-1.403	–	–
MATHS	-0.194	-0.695	–	–
EARTH	-0.024	-0.083	–	–
TURN2	0.303	1.997**	0.179	1.747*
R^2	0.086		0.056	
adj R^2	0.012		0.046	
F	1.167		5.845***	
N	201		201	
R^2 change[a]	0.020		0.015	
F change[a]	3.986**		3.051*	

* $p < 0.1$; ** $p<0.05$; *** $p<0.01$. [a] after entering TURN2 into the models.

Note: since it has the greatest frequency, in Model 1 the omitted major became the BIOLOGY.