A comparison of self- vs. tutor assessment among Hungarian undergraduate business students

András István Kun

Department of Organization Sciences, 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-4032.

Notes on contributor

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

A comparison of self- vs. tutor assessment among Hungarian undergraduate business students

The current study analyses the self-assessment behaviour and efficiency of 163 undergraduate business students from Hungary. Using various statistical methods the results support the hypothesis that high-achieving students are more accurate in their pre- and post-examination self-assessments, and also less likely to overestimate their performance, and if they do so, the mean overestimation is lower than in the case of lower-achieving students. The study did not find a strong difference in the tendency to self-overestimation between sexes, but in their pre-examination prediction women seem to overestimate significantly more than men. An overall tendency among the students to over-rate their own examination performance is also detected, as is a tendency to increase the accuracy of self-assessment after sitting the examination.

Keywords: self-assessment, business education, higher education, students' academic performance

Introduction

The motivation to write this paper comes from the phenomenon that many tutors may experience in higher education (see among other Macdonald 2004): a large number of students seem not to be able to rationally evaluate their own knowledge and preparedness for examinations. Moreover, this is an even more serious issue for the less prepared (i.e. the lower achieving) students. A significant number of papers address the problem of differences between students' self- and tutor assessment; however, some of them use the notion of self-assessment in a broader sense, involving self-directed education in the discussion (e.g. Karnilowicz 2012, 592). Understanding how the students' self-evaluation and their achievement (e.g. their true preparedness) are connected – if they are at all – can help tutors and institutions to facilitate students to manage their own learning. Nicol and Macfarlene-Dick (2006) point out that students already assess their own work, thus higher education institutions could build on this ability. However, if this self-assessment is not accurate, then

students may set themselves inappropriate learning goals and/or mismanage their learning efforts, which will lead to lower performance both for them and for their institution. If the lowest achieving students overestimate their future performance, than they will put too little effort into learning and will not meet their expectations and goals (moreover, if they overestimate their abilities they may even set themselves unattainable goals). On the other hand, the objectives of students who underestimate themselves may be over-modest or they may waste time and resources on too much learning, and thus may be unable to accept other challenges or may miss other opportunities. Several researches have previously shown that students' self-assessment ability is learnable (e.g. Everett 1983; Pintrich 1995; Zimmerman and Schunk 2001; Ross 2006; Baartman and Ruijs 2011) – although there are other studies which do not support this finding, such as Fitzgerald, White and Gruppen (2003) –, pointing out which student groups are exposed to the phenomenon of inaccurate self-assessment and to what extent it can contribute to the efficiency of higher education institutions' actions to facilitate their students self-management.

The current study focuses its attention strictly on the measurement of higher education business students' ability to predict and evaluate their own performance in written examinations, and also the connection between this and teacher assessed achievement in the same examination.

The next section of the article briefly reviews the related empirical literature and, based on this, composes four hypotheses for the empirical research. Sections discussing the research sample and method follow, introducing the framework of the analysis described in the results section, where an explanation is given for each of the hypotheses. Based on the outcomes, the conclusion describes the implications for the hypotheses and formulates the contribution of this article to the literature, as well as pointing out the limitations of the findings.

Review of the literature

The definition of self-assessment by Boud and Falchikov (1989, 529) is "the involvement of learners in making judgements about their own learning, particularly about their achievements and the outcomes of their learning". However, a broader approach encompasses not only the act of judging the performance, but also the identification of criteria or standards, and through this process it is connected to self-directed learning (Karnilowicz 2012, 591-593). The current study analyses only a part of the phenomenon: students' ability to predict and to evaluate their examination performance relative to their externally assessed achievement; therefore the overview of the literature will also focus on this part.

There are at least two main directions in the research into students' self-assessment ability in the context of their abilities or achievement: the investigation of its accuracy (how strongly related it is to the real – tutor assessed – performance of the student) and of the tendency of students to over- or underrate themselves. The impact of other influential factors' – most frequently the students' sex – on the accuracy or self over- or underestimation is also investigated in many studies.

Based on the studies reviewed in their article, Boud and Falchikov (1989) state that there is no detectable unequivocal tendency towards over-estimation in student self-assessment: they have reviewed 17 articles where a general tendency to over-estimate was identified and 12 where it was not. In the later literature Krueger and Dunning (1999), Basnet et al. (2012), and Tejeiro et al. (2012) supported the existence of such a phenomenon, while Mehrdad, Bigdeli and Ebrahimi (2012) have found no general disposition for either under- or over-estimation.

Regarding the relationship between the students' externally measured performance and the accuracy of their self-assessment every study reviewed by the author (Boud and Falchikov 1989; Krueger and Dunning 1999; Sundström 2005; Tejeiro et al. 2012; Karnilowicz 2012) –

with the sole exception of the study by Lynn, Holzer and O'Neill (2006) – have concluded that higher-achieving students are, on average, more accurate in their self-assessment than low achievers. Tausignant and DesMarchais (2002), Edwards et al. (2003), and Eva et al. (2004) also found that the pre-assignment self-predictions are less accurate than post-assignment self-evaluations. Fitzgerald, White and Gruppen (2003) compared the self-assessment accuracy for three separate years of students and detected a relative stability over those years; however stability in time is not supported by Baartman and Ruijs (2011).

Unfortunately the term accuracy was used mistakably from the measuring point of view in several of the studies cited above, hence the accuracy of student self-assessment was conceptualised and measured as the estimated test score minus the actual test score, or, using the course grade, as the linear or non-parametric correlation of this with the actual scores (e.g. Krueger and Dunning 1999; Tausignant and DesMarchais 2002; Fitzgerald, White and Gruppen 2003; Tejeiro et al. 2012).

In general measurement accuracy is "the closeness between the measurement result and the true value of the measurand" (Rabinovich 2013:2). Hence the mean of – or similarly the correlation with – the signed error values can conceal prediction inaccuracy (for example when there are two predictions for value 0, –10 and 10, than the mean of the signed prediction errors is zero); in most cases measurement should be carried out with an appropriate method that eliminates the sign of the errors before calculating their mean (e.g. absolute value, rooted square). The practice in the current paper separates and analyses the 'accuracy' of self-assessment and the 'direction' of the self-assessment errors in relation to and from each other. Accuracy is defined as the absolute difference between the student-estimated and the actual test score, while direction is the positive or negative sign of the difference (distinguishing between under- or over-estimation). This is necessary in order not to disguise the phenomena of students' academic abilities contributing to the self-estimation ability independently of its

direction. This latter hypothesis was articulated in many of the above cited papers, however the method used to test it was occasionally inappropriate.

All the reviewed literature that has addressed the question supports without exception the idea that high-achieving students tend to overestimate their own performance less than their low-achieving fellows, and moreover, sometimes even underestimate it (Boud and Falchikov 1989; Fitzgerald et al. 1997; Krueger and Dunning 1999; Hodges, Regehr and Martin 2001; Lejk and Wyvill 2001; Edwards et al. 2003; Gramzow et al. 2003; Karnilowicz 2012). According to Edwards et al. (2003) and Macdonald (2004), there is a difference in the direction of the self-estimation errors between the two sexes: men tend to overestimate themselves more than women. However there are several studies that could not find this kind of gender-related effect (Boud és Falchikov 1989; Krueger and Dunning 1999; Lynn, Holzer and O'Neill 2006; Basnet et al. 2012).

Based on the questions and findings of the literature reviewed above, the current study forms four hypotheses:

H1: Higher-achieving students assess their examination results more accurately (measured with the absolute value of the assessment error) than their lower-achieving fellows. This hypothesis is divided into two sub-hypotheses:

H11: Higher-achieving students predict their examination results more accurately (measured with the absolute value of the pre-examination assessment error) than their lower-achieving fellows.

H12: Higher-achieving students evaluate their examination results more accurately (measured with the absolute value of the post-examination assessment error) than their lower-achieving fellows.

H2: High-achieving students tend to over-assess their examination results less than low-achieving students.

H3: Compared to female students, males tend to overestimate their own performance more.

H4: Ceteris paribus students tend to overrate their performance and this overrating is greater in pre-examination than in post-examination self-estimations.

Sample and method

The total sample consists of 163 business students from the University of Debrecen, Debrecen, Hungary, 13 of whom (2 males, 11 females) were taking part in a vocational higher education program, the others being bachelor students at the time of the examination. 70 bachelor students (24 males, 46 females) were studying on the Business Administration and Management and 80 (21 males, 59 females) on the International Business Economics major. The examination could be taken on one of two possible dates of the students' choice (in the middle or at the end of the semester). On the first date 2 test versions (identified as A and B) were used, taken by 42 and 41 students, and 4 (A, B, C, and D) on the second date, with 22, 19, 20 and 19 test-takers. To eliminate any effect deriving from the occasional differences among the test versions, the above mentioned factors are always taken into consideration as dummy variables during the following analyses. All test versions had the same structure: 20 multiple choice questions (1 correct answer from 4 choices) and 3 calculation problems. The multiple choice questions count for 20 points and the calculation problems for 50 in the total test score, thus the maximum score is 70. On both examination dates, the tests were written in two consecutive sessions, with the same versions in each session.

Before they started their exam, students were asked to predict their total multiple choice and total calculation scores (they estimated two numbers, one between 0 and 20 and the other between 0 and 50). To motivate them to predict more accurately, they were offered a percentage of their total test score if they estimated well (+10% on a perfect hit for both multiple choice and calculation questions, or +5% if the estimation was within a ±1 point

range). After the examination ended, they were asked again to make a new, final estimation of the same scores. Hence in the calculation of bonus points only the second estimation was involved, and it offered them a chance to correct their former prediction. From a research point of view, pre-examination and post-examination assessments created a possibility to examine how well students are able to re-evaluate their knowledge during the test.

In the cases of hypotheses H1 and H2 the main tools of the statistical analysis are binary logistic regression models similar to the work of Edwards et al. (2003), with one significant modification. Edwards and his team use a binary independent variable to indicate if the given self-assessment was made before or after the assignment, while in the current paper pre- and post-examination data are analysed in separate models, hence the use of the original method would duplicate each student in the sample (once performing the role of a pre- examination evaluator and once that of a post-examination evaluator). As a supplementary method for testing H2, independent samples *t*-tests are also used to compare the terciles of the highest- and lowest-achieving students. Similar analyses were also frequently used in the literature referred to above. The independent samples *t*-test is again the method selected to compare the self over-evaluation tendency of men and women (H3), and descriptive statistics, measures of association and a paired *t*-test have been chosen to analyse the overall tendency to overestimate (H4) and the differences between pre- and post-examination self-assessment within this (both in frequencies and means).

Results

Before testing the hypotheses the study provides an overview of the descriptive statistics of the sample data in Table 1, in which the reader can see how the median, mean and standard deviation values of the student pre- and post-examination self-estimations and the tutor-assigned scores are distributed by gender. Other factors (major, examination date, session, test version) have not yet been taken into consideration.

[Table 1 near here]

Table 1 suggests that both sexes overestimated their test scores, and the overestimations were higher in the pre- than in the post-examination evaluation. The self-assessment scores of female students were higher before the test and slightly lower after it than those of their male counterparts. The average male student outperformed the average female, according to the tutor assigned scores. The significance of these findings is studied through an examination of the hypotheses below. Table 2 contains the description of the variables used in further analysis.

[Table 2 near here]

Testing the H1 hypothesis

H11 and H12 are tested with linear regression models, where the dependent is the accuracy of the students pre-test and post-test estimations (measured with the absolute difference between the student estimated scores and the tutor assigned scores), while a function of the tutor assigned test score is an independent variable (among others). The functions of MCSCORE, CPSCORE and TTSCORE are selected in order to maximize the ratio that the models can explain from the variance of the dependent variable (R^2). Self-assessment accuracy models are estimated for multiple choice questions, calculation problems and the total test score independently, each in two versions. The first contains all the available independent variables (Model 1); the other is restricted to those that are significant at least on the 10% level (Model 2). Statistics of the regression models are shown in Table 3 for the pre- and in Table 4 for the post-test estimations.

[Table 3 near here]

According to both Model 1 and Model 2, self-predictions of multiple choice scores are more accurate if the student is male, the test is written in the second session and the student is more

prepared (that is the test earns a higher score when assessed by the tutor). Although the linear relation would also be significant, the logarithmic function of MCSCORE has a slightly stronger explanatory power, and so it is used in Table 3. There were no significant differences in accuracy among majors and test versions. In the case of calculation problems, both Model 1 and Model 2 show a significant (at the 1% level) linear, positive connection between the accuracy of students' prediction and the tutor's assessment (better students again seem to be more accurate). In the calculation problem case the test version also plays a role in accuracy as a situational factor, in that vocational higher education students tend to be more accurate than bachelor students. Actually, in the models of the total test score self-estimations these latter outcomes are echoed: tutor scoring relates negatively to self-assessment mistakes, those on the vocational higher education course are more accurate, and three of the test versions were proved to facilitate a more correct estimation when compared to the others. Based on all the 6 regression models above, the H11 hypothesis should be considered as supported, as it argues that students better at learning are better in pre-examination self-assessment, too.

In the case of post-test self-assessment, we can find somewhat different outcomes (see Table 4). Hence the main findings are the same for both Model 1 and Model 2, so they are not explained individually. Student-estimated multiple choice scores are again in a positive logarithmic relationship with accuracy (thus correlate negatively with the average number of mistakes), although the explanatory power is slightly lower than in the case of pre-examination assessment. Unlike the pre-examination self-assessment models, sex has no significant effect on accuracy, but session has. Estimation accuracy on calculation scores is significantly connected to tutor assessment via a cubic function, that decreases accuracy up to $CPSCORE \leq 11.8846$, increases it when $11.8847 \leq CPSCORE \leq 46.6488$, and decreases it again if $46.6489 \leq CPSCORE$. Among the other independent variables, only the dummy variable of vocational higher education has a significant positive relationship with accuracy.

The explanatory power of both Models 1 and 2 is much lower than in the case of pre-examination predictions. For the total test score the accuracy of self-evaluation is greater for vocational higher education students and for those who are assigned a higher score by the tutor (TTSCORE2 is significant). Both models' explanatory power is much lower again than in the pre-examination case. Although H12 is supported for MCSCORE and TTSCORE, the connection between self-assessment accuracy and students' achievement is found to be weaker after the test is taken than before. Thus, self-evaluation seems to change following the experience with the examination. For CPSCORE, H12 is not supported. According to the findings, for very low and extremely high CPSCOREs the relation to accuracy is negative, while from low-medium to high scores it is positive.

[Table 4 near here]

Hence H11 is supported and H12 is mostly supported (except for calculation scores); H1 as the hypothesis which frames them could also be mostly accepted, with the exception of calculation scores after the test writing.

Testing the H2 hypothesis

Contrary to H1, which emphasized the extent of the estimation errors independently of their positive or negative sign, the second hypothesis focuses directly on the sign, asking if a higher achieving student tends to underestimate his/her own preparedness (measured with the tutor-assigned score) less or more than those who are not as ready to take the exam.

Dividing the sample into terciles by tutor's assigned multiple choice, calculation problem or total test scores, – using the independent *t*-test – significant differences are identified in the mean assessment errors between the first and third terciles in both the pre- and post examination cases. For multiple choice scores, the mean difference of the self-assessment error between the highest- and the lowest-achieving terciles is 4.0417 (t = 7.1121,

$p < 0.01$) in the pre-examination case, and 3.2024 ($t = 5.4203$, $p < 0.01$) in the post-examination case. For calculation problem scores the mean difference between the higher and lowest terciles is 10.6389 ($t = 5.6589$, $p < 0.01$) in the pre-examination case, and 4.8679 ($t = 2.7975$, $p < 0.01$) in the post-examination case. The same differences for total test scores are 11.9107 ($t = 5.8298$, $p < 0.01$) and 5.8023 ($t = 3.2799$, $p < 0.01$). This is equivalent to saying that the lowest-achieving one third of the students tend to overestimate their own performance before and after the examination. Thus H2 is supported.

However, if one wishes to identify a purer connection between students' achievement (i.e. their score assigned by the tutor) and the likeliness to under- or over-estimate their own performance, without the influence of other factors, binary logistic regression could be a more appropriate tool for analysis. In all constructed models (they can be found in Table 5) the dependent variable is the likeliness that a student over-assesses him/herself. Those cases where the tutor-assigned score was 0 or maximal, leaving no chance for under- or over-assessment error, are left out of the sample. Cases where the student estimated his/her own performance without error were also neglected, being unimportant for this question.

[Table 5 near here]

The binary logistic regression models presented in Table 5 show a relatively good explanatory power (measured by the Nagelkerke R^2), especially when forecasting over- and underestimations of the total test scores (last column). The squared TTSCORE is able to increase this explanatory power (the change in the Nagelkerke R^2) by more than 0.1 in all of the six tested models. However, while in the cases of multiple choice or calculation problem assessments there is a lower possibility the better students will overestimate themselves, in the total score case the connection between performance and the tendency to overestimate is not linear. At lower scores (up to 44.2210) the better the tutor's assessment, the greater the chance of self-overestimations, and above a score of 44.2211 the relationship becomes negative. The

explanatory power – measured with the Nagelkerke R^2 – of the post-examination self-assessment models (see table 6) is from 0.0293 to 0.5933 less than in the pre-examination cases for multiple choice, calculation problem and total test scores. However, the main findings remain the same (both multiple choice and calculation problem scores affect the chance that students over-assess their own performance linearly and negatively), with the only exception that this time the likeliness of self-overestimation is a linear – and not a quadratic – function of the total test score. The H2 hypothesis is thus mostly supported by the binary logistic regression analysis, with the only exception being that in Model 2 for the pre-examination total scores, the negative relationship between self-overassessment of total scores and TTSCORE is not supported for lower-achieving students (below app. 44 points form 70) but only for high-achievers.

[Table 6 near here]

Testing the H3 hypothesis

To compare the two sexes in their under- or over-estimating habits in self-assessment, independent samples *t*-tests are used on pre- and post-examination multiple choice, calculation problem and total test scores. Mean, median and standard deviation values for the self-evaluations and also for the tutor assessments are presented in Table 1, showing that both sexes overestimate their performance. The remaining question is, which gender has a stronger tendency to overestimate? The relevant independent *t*-test statistics are in given Table 7 below.

[Table 7 near here]

Both genders overestimated their test results and, contradicting the hypothesis, the results show a significant difference between the two genders in pre-examination multiple choice and total test scores, where female students are proven to make a greater overestimation of their

performance. In all other cases it cannot be refuted that the two sexes act indifferently. Consequently, H3 should be rejected.

Testing the H4 hypothesis

Table 8 convincingly supports the first part of hypothesis H4 for both genders (and thus also for the total sample), since there are far more students who over-assess than under-assess their multiple choice, calculation problem and total test scores. It should also be noted, however, that for every gender and test-type combination there are fewer students who underestimated their test results (by from 4.08 to 4.48 percentage points) and accurately estimated them (by from 0.88 to 6.87 percentage points), and more who over-estimated them (by from -16.36 to -6.12 percentage points) before than after the examination. These differences are statistically significant (for the association measures see Table 8) in the case of calculation and total test scores but not significant (with the only exception of the Goodman – Kruskal tau test) for the multiple choice scores. The overall results support the H4 hypothesis.

[Table 8 near here]

Besides the frequencies, the scale of under- and overestimations is also important to evaluate students' sense of reality in self-assessment. This paper joins paired *t*-tests to compare means of pre- and post-examination self-assessment errors (see Table 9). The overestimation is significantly lower after than before the examination for every score on the total sample and also on the subsamples by gender, except with the multiple choice estimation of the male students, where the mean difference is not significant, but fittingly directed (there is a lower average overestimation after than before sitting the test). Thus H4 is supported once again.

[Table 9 near here]

Conclusions

The main contribution of the analysis given above is to provide new empirical results to the

literature from a sample of Hungarian business higher education students. The findings support the hypotheses that students with higher educational achievement are also better at both predicting (prior to an examination) and evaluating (right after an examination) their own academic performance in terms of accuracy as measured by absolute estimation errors (hypotheses H1, H11, H12), which is in accordance with the conclusions of Boud and Falchikov (1989), Krueger and Dunning (1999), Sundström (2005), Tejeiro et al. (2012), and Karnilowicz (2012), but not with the findings of Lynn, Holzer and O'Neill (2006). On the other hand, the results also support the idea (as do the studies in the literature reviewed above) that they are less likely to overestimate their own results, which means they have a more realistic attitude (H2). Edwards et al. (2003) and Macdonald (2004) identified a higher tendency to self-overestimation in the case of male students compared to their female fellows (H3). These findings are not supported by the current analysis; moreover, in some cases (pre-examination self-estimation of multiple choice and total test scores) women seem to overestimate significantly more than men. An overall tendency of the students to overestimate their own performance is identified (supporting the results of Krueger and Dunning 1999; Basnet et al. 2012; and Tejeiro et al. 2012), as is the more marked tendency to do so before the examination rather than right after it (H4). The latter finding echoes the results of Tausignant and DesMarchais (2002), Edwards et al. (2003), and Eva et al. (2004). As an additional contribution, the article pointed out that the self-assessment accuracy and the tendency to over- or underestimate should be addressed and measured separately, otherwise the true patterns of both phenomena might remain disguised.

Since the analysed sample of students is from one program and one year of a given university, sitting an examination in a given undergraduate subject, any generalisation of the results should be approached with caution. However, most of the findings are supported by at least a part of the previous empirical studies, thus in their context the results of this paper can

contribute to a better understanding of the big picture. From another point of view, the differences which presumably exist between countries, nationalities, cultures, or academic areas in such a culture-dependent question as self-evaluation represent a good opportunity for further research.

Acknowledgements

The author would like to thank the help given by Bence Kovács who contributed to the data collection.

Notes on contributor

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

References

Baartman, L., and L. Ruijs. 2011. "Comparing students' perceived and actual competence in higher vocational education." *Assessment & Evaluation in Higher Education* 36 (4): 385–398. doi: 10.1080/02602938.2011.553274.

Basnet, B., M. Basson, C. Hobohm, and S. Cochrane. 2012. *Students' self-assessment of assignments – is it worth it?* Proceedings of the 2012 AAEE Conference, Melbourne, Victoria. url: http://www.aaee.com.au/conferences/2012/documents/abstracts/aaee2012-submission-194.pdf.

Boud, D., and N., Falchikov. 1989. "Quantitative studies of student self-assessment in higher education: a critical analysis of findings." *Higher Education* 18 (5): 529–549. doi: 10.1007/BF00138746.

Edwards, R. K., K. R. Kellner, C. L. Sistrom, and E. J. Magyari. 2003. “Medical student self-assessment of performance on an obstetrics and gynecology clerkship.” *American Journal of Obstetrics and Gynecology* 188 (4): 1078–1082. doi: 10.1067/mob.2003.249.

Eva, K. W., J. P. W. Cunnington, H. I. Reiter, D. R. Keane, and G. R. Norman. 2004. “How Can I Know What I Don’t Know? Poor Self Assessment in a Well-Defined Domain.” *Advances in Health Sciences Education* 9 (3): 211–224. doi: 10.1023/B:AHSE.0000038209.65714.d4.

Everett, M .S. 1983. “Influence of trait anxiety on self-grading.” *Educational Directions* 8 (1): 4–9.

Fitzgerald, J. T., L. D. Gruppen, C. B. White, and W. K. Davis. 1997. *Medical student self-assessment abilities: Accuracy and calibration.* Presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, April. url: http://files.eric.ed.gov/fulltext/ED410296.pdf

Fitzgerald, J. T, C. B White, and L. D. Gruppen. 2003. “A longitudinal study of self-assessment accuracy.” *Medical Education* 37 (7): 645–649. doi: 10.1046/j.1365-2923.2003.01567.x.

Gramzow, R. H., A. J. Elliot, E. Asher, and H. A. McGregor. 2003. “Self-evaluation bias and academic performance: Some ways and some reasons why.” *Journal of Research in Personality* 37 (2): 41–61. doi: 10.1016/S0092-6566(02)00535-4

Hodges, B., G. Regehr, and D. Martin. 2001. “Difficulties in Recognizing One’s Own Incompetence: Novice Physicians Who Are Unskilled and Unaware of It.” *Academic Medicine: Journal Of The Association Of American Medical Colleges* 76 (10 Suppl): 87–89.

Karnilowicz, W. 2012. “A Comparison of Self-Assessment and Tutor Assessment of Undergraduate Psychology Students.” *Social Behavior and Personality* 40 (4): 591–604. doi: 10.2224/sbp.2012.40.4.591.

Kruger, J., and D. Dunning. 1999. “Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated self-assessments.” Journal *of Personality and Social Psychology* 77 (6): 1121–1134. doi: 10.1037/0022-3514.77.6.1121.

Lejk, M., and M. Wyvill. 2001. “The effect of the inclusion of self-assessment with peer assessment of contributions to a group project: A quantitative study of secret and agreed assessments.” *Assessment & Evaluation in Higher Education* 26 (6): 551–561. http://doi.org/brg2sw

Lynn, D. J., C. Holzer, and P. O’Neill. 2006. “Relationships between self-assessment skills, test performance, and demographic variables in psychiatry residents.” *Advances in Health Sciences Education* 11 (1): 51–60. doi: 10.1007/s10459-005-5473-4.

Macdonald, A. 2004. “Student self-evaluation of coursework assignments: a route to better perception of quality.” *Learning and Teaching in Higher Education* 1 (1): 102–107. url: http://insight.glos.ac.uk/tli/resources/lathe/Documents/issue%201/case%20studies/macdonald.pdf

Mehrdad, N., S. Bigdeli, and H. Ebrahimi. 2012. “A Comparative Study on Self, Peer and Teacher Evaluation to Evaluate Clinical Skills of Nursing Students.” *Procedia – Social and Behavioral Sciences* 47 (1): 847–1852. doi: 10.1016/j.sbspro.2012.06.911.

Nicol, D., and D. Macfarlane-Dick. 2006. “Formative assessment and self-regulated learning: a model and seven principles of good feedback practice.” *Studies in Higher Education* 31 (2): 199–218. doi: 10.1080/03075070600572090

Pintrich, P. R. 1995. *Understanding self-regulated learning*. San Francisco, CA: Jossey-Bass.

Rabinovich, S. G. 2013. *Evaluating Measurement Accuracy: A Practical Approach*. New York, NY: Springer.

Ross, J. A. 2006. "The Reliability, Validity, and Utility of Self-Assessment." *Practical Assessment, Research & Evaluation* 11 (10) url: http://pareonline.net/getvn.asp?v=11&n=10

Sundström, A. 2005. *Self-assessment of knowledge and abilities: A literature study*. EM No 54. Umeå, Sweden: Umeå University. url: http://www.edusci.umu.se/digitalAssets/60/60577_em541.pdf.

Tousignant, M., and J. E. DesMarchais. 2002. "Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study." *Advances In Health Sciences Education* 7 (1): 19–27. doi: 10.1023/A:1014516206120

Tejeiro, R. A., J. L. Gomez-Vallecillo, A. F. Romero, M. Pelegrina, A. Wallace, and E. Emberley. 2012. "Summative Self-Assessment in Higher Education: Implications of Its Counting towards the Final Mark." *Electronic Journal of Research in Educational Psychology* 10 (2): 789–812. url: http://www.investigacion-psicopedagogica.org/revista/new/english/ContadorArticulo.php?707.

Zimmerman, B. J. and D. H. Schunk. 2001. *Self-regulated learning and academic achievement: theoretical perspective*s. Mahwah, NJ: Lawrence Erlbaum Associates

Table 1. Descriptive statistics of test score estimations and real test scores by gender

	Pre-examination estimations (N = 163)						Post-examination estimations (N = 162)					
	Multiple choice			Calculation problem			Multiple choice			Calculation problem		
	Median	Mean	S.D.	Median	Mean	S.D.	Median	Mean	S.D.	Median	Mean	S.D.
M	10.0000	10.4894	2.9479	27.0000	26.2979	10.8766	10.0000	10.2128	3.3940	20.0000	22.1277	12.8767
F	12.0000	11.7586	2.6560	28.0000	26.5603	11.7149	10.0000	10.0609	2.7025	20.0000	19.6696	13.4175
T	11.0000	11.3926	2.7943	28.0000	26.4847	11.4468	10.0000	10.1049	2.9098	20.0000	20.3827	13.2703

	Tutor's assessment (N = 163)					
	Multiple choice			Calculation problem		
	Median	Mean	S.D.	Median	Mean	S.D.
M	8.0000	8.1064	2.2959	13.0000	15.3404	13.4415
F	7.0000	7.5948	2.4847	9.2500	12.8922	12.9019
T	8.0000	7.7423	2.4358	11.0000	13.5982	13.0655

Notes: M = males, F = females, T = total, S.D. = standard deviation

Table 2. Explanation of variables

Variable name	Explanation
SEX	1 if the student is female, 0 if male.
VHE	1 if the student is on a vocational higher education program, 0 if not.
BAM	1 if the student is on the business administration and management major, 0 if not.
IBE	1 if the student is on the international business economics major, 0 if not.
V1	1 if the test was taken on the 1st examination date and it was version A, 0 if not.
V2	1 if the test was taken on the 1st examination date and it was version B, 0 if not.
V3	1 if the test was taken on the 2nd examination date and it was version A, 0 if not.
V4	1 if the test was taken on the 2nd examination date and it was version B, 0 if not.
V5	1 if the test was taken on the 2nd examination date and it was version C, 0 if not.
V6	1 if the test was taken on the 2nd examination date and it was version D, 0 if not.
SESSION	1 if the test was taken in the second session and 0 if in the first.
MCSCORE	The raw multiple choice score.
CPSCORE	The raw calculation problem score.
TTSCORE	MCSCORE + CPSCORE
DIFMC1	(Student's pre-examination estimation of MCSCORE) – MCSCORE
DIFCP1	(Student's pre-examination estimation of CPSCORE) – CPSCORE
DIFTT1	(Student's pre-examination estimation of TTSCORE) – TTSCORE
DIFMC2	(Student's post-examination estimation of MCSCORE) – MCSCORE
DIFCP2	(Student's post-examination estimation of CPSCORE) – CPSCORE
DIFTT2	(Student's post-examination estimation of TTSCORE) – TTSCORE
ADIFMC1	Absolute value of DIFMC1
ADIFCP1	Absolute value of DIFCP1
ADIFTT1	Absolute value of DIFTT1
ADIFMC2	Absolute value of DIFMC2
ADIFCP2	Absolute value of DIFCP2
ADIFTT2	Absolute value of DIFTT2
OEMC1	1 if the student overestimated its multiple choice scores before exam, 0 if not.
OECP1	1 if the student overestimated its calculation problem scores before exam, 0 if not.
OETT1	1 if the student overestimated its total test scores before exam, 0 if not.
OEMC2	1 if the student overestimated its multiple choice scores after exam, 0 if not.
OECP2	1 if the student overestimated its calculation problem scores after exam, 0 if not.
OETT2	1 if the student overestimated its total test scores after exam, 0 if not.

Table 3. Linear regression models for the pre-examination student vs. tutor assessment differences

Dependent	ADIFMC1		ADIFCP1		ADIFTT1	
Variables	Model1	Model2	Model1	Model2	Model1	Model2
Constant	11.5425***	10.6720***	19.4892***	21.2814***	27.2323***	28.9837***
	(8.9564)	(9.2761)	(7.3878)	(19.2781)	(8.4967)	(18.6117)
SEX	0.8450**	1.0128**	-0.1163	–	1.0368	–
	(2.0539)	(2.5940)	(-0.0837)		(0.6473)	
VHE	-0.3946	–	-6.1835**	-6.3650***	-6.3459**	-6.7244**
	(-0.5731)		(-2.5966)	(-2.7889)	(-2.3108)	(-2.5535)
BAM	-0.0425	–	0.8892	–	1.4615	–
	(-0.1114)		(0.6879)		(0.9834)	
V2	-0.3837	–	-0.3549	–	-0.9059	–
	(-0.7612)		(-0.2078)		(-0.4606)	
V3	-0.4503	–	-7.9343***	-7.3527***	-9.0094***	-8.2004***
	(-0.7379)		(-3.7630)	(-4.0013)	(-3.7140)	(-3.8756)
V4	-0.7026	–	-4.0187*	-3.4020*	-5.1042**	-4.1861*
	(-1.0830)		(-1.8490)	(-1.7812)	(-2.0362)	(-1.8981)
V5	-0.9604	–	-5.2025**	-4.8138**	-7.2529***	-6.5061***
	(-1.5308)		(-2.4406)	(-2.5728)	(-2.9556)	(-3.0163)
V6	-0.8876	–	-1.0023	–	-1.8705	–
	(-1.3750)		(-0.4582)		(-0.7430)	
SESSION	-0.6707*	-0.6734*	1.3315	–	0.7380	–
	(-1.8329)	(-1.8864)	(1.0721)		(0.5164)	
ln(MCSCORE)	-3.3539***	-3.2307***	–	–	–	–
	(-6.8140)	(-6.7859)				
CPSCORE	–	–	-0.3540***	-0.3537***	–	–
			(-7.1993)	(-7.3216)		
TTSCORE	–	–	–	–	-0.4051***	-0.4080***
					(-7.5819)	(-7.7624)
F	6.2719***	20.0810***	6.4919***	12.8182***	7.2222***	14.158***
R^2	0.2921	0.2748	0.2993	0.2899	0.3221	0.3108
adjusted R^2	0.2455	0.2611	0.2532	0.2673	0.2775	0.2888
R^2 change	0.2162***	0.2100***	0.2389***	0.2425***	0.2564***	0.2645***

Note: $N = 163$. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$; t-statistics in parentheses. As it contains the greatest number of cases, the IBE is omitted from the majors and the V1 is omitted from the test-versions

Table 4. Linear regression models for differences in the post-examination student vs. tutor assessment

Dependent Variables	ADIFMC2		ADIFCP2		ADIFTT2	
	Model1	Model2	Model1	Model2	Model1	Model2
Constant	9.9080***	9.4009***	6.7543***	7.3034***	11.9512***	13.1036***
	(8.0233)	(9.1329)	(2.6316)	(6.6690)	(3.4454)	(15.7527)
SEX	-0.3257	–	-0.3937	–	-0.3535	–
	(-0.8243)		(-0.3162)		(-0.2532)	
VHE	-0.0449	–	-4.7129**	-4.9558**	-5.0253**	-6.0571***
	(-0.0680)		(-2.1949)	(-2.4483)	(-2.0904)	(-2.6913)
BAM	0.0363	–	0.9821	–	1.4755	–
	(0.0990)		(0.8468)		(1.1394)	
V2	-0.0088	–	-0.3176	–	-0.5613	–
	(-0.0181)		(-0.2044)		(-0.3273)	
V3	-0.1835	–	-0.9046	–	-1.8619	–
	(-0.3141)		(-0.4752)		(0.8778)	
V4	-0.4580	–	-2.0337	–	-3.5163	–
	(-0.7374)		(-1.0459)		(-1.6183)	
V5	-0.0103	–	-1.5021	–	-2.2938	–
	(-0.0172)		(-0.7842)		(-1.0700)	
V6	-0.2986	–	-0.2911	–	-2.0342	–
	(-0.4724)		(-0.1453)		(-0.9101)	
SESSION	-0.6710*	-0.6748**	0.8150	–	-0.3015	–
	(-1.9024)	(-1.9882)	(0.7260)		(-0.2407)	
ln(MCSCORE)	-2.8613***	-2.7718***	–	–	–	–
	(-6.0722)	(-6.1794)				
CPSCORE	–	–	0.8316***	0.8456***	–	–
			(2.8675)	(3.0661)		
$CPSCORE^2$	–	–	-0.0439**	-0.0443***	–	–
			(-2.5992)	(-2.7521)		
$CPSCORE^3$	–	–	0.0005**	0.0005***	–	–
			(2.1302)	(2.2633)		
TTSCORE	–	–	–	–	0.2442	–
					(1.3857)	
$TTSCORE^2$	–	–	–	–	-0.0069**	-0.0026***
					(-2.2421)	(-3.2664)
F	4.1500***	20.9737***	1.8789**	5.1221***	1.9445**	7.4907
R^2	0.2156	0.2087	0.1314	0.1154	0.1248	0.0861
adjusted R^2	0.1636	0.1988	0.0615	0.0929	0.0606	0.0746
R^2 change[a]	0.1915***	0.1900***	0.0770***	0.0780***	0.0799***	0.0613***

Note: $N = 162$. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$; t-statistics in parentheses; [a] R^2 change after the test scores were included in the models. As it contains the greatest number of cases, the IBE is omitted from the majors and the V1 is omitted from the test-versions.

Table 5. Binary logistic regression models (pre-exam)

Dependent	OEMC1		OECP1		OETT1	
Variables	Model1	Model2	Model1	Model2	Model1	Model2
Constant	19.7634	5.6200***	2.6478	3.1127***	-36.2221	–
	(0.0000)	(14.6788)	(0.7652)	(8.4916)	(0.0000)	
SEX	-1.2005	1.5566**	-0.9629	–	-1.8298*	2.4673***
	(2.6511)	(5.7949)	(1.5448)		(3.1334)	(8.4544)
VHE	-18.0922	–	3.4659**	2.4481*	-15.8611	–
	(0.0000)		(4.7362)	(3.3445)	(0.0000)	
BAM	-0.4763	–	-0.0562	–	0.3923	–
	(0.4178)		(0.0060)		(0.1802)	
V2	0.1338	–	-0.4255	–	14.7657	3.3527*
	(0.0076)		(0.1566)		(0.0000)	(3.3637)
V3	1.9893	–	0.2145	–	-20.7458	–
	(2.2725)		(0.0254)		(0.0000)	
V4	1.8580	–	-0.5346	–	19.7266	–
	(1.5199)		(0.1796)		(0.0000)	
V5	2.3440	–	1.4215	–	20.2739	–
	(3.1208)		(1.5885)		(0.0000)	
V6	0.6549	–	0.8986	–	19.1881	–
	(0.1634)		(0.7952)		(0.0000)	
SESSION	-0.3796	–	-0.8857	–	0.3796	–
	(0.2410)		(1.3589)		(0.1620)	
MCSCORE	-0.4705***	-0.4804***	–	–	–	–
	(8.5259)	(9.9887)				
CPSCORE	–	–	-0.1827***	-0.1548***	–	–
			(20.0162)	(23.1207)		
TTSCORE	–	–	–	–	0.2403	0.2418***
					(2.6633)	(21.1435)
TTSCORE^2	–	–	–	–	-0.0061**	-0.0055***
					(4.7499)	(21.8284)
N	148	148	157	157	159	159
OX^2	27.9902***	19.0349***	51.2533***	44.1851***	38.5295***	174.0293***
HLX^2	3.9617	1.4377	6.8505	5.0196	3.8750	7.1090
NR^2	0.3843	0.2691	0.5467	0.4815	0.5741	0.8871
NR^2 change[a]	0.1451	0.1681	0.4852	0.4787	0.2745	0.1096

Note: $N = 162$; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$; Wald statistics in parentheses; OX^2 = Omnibus test X^2; HLX^2 = Hosmer – Lemeshow X^2; NR^2 = Nagelkerke R^2; [a] NR^2 change after the test scores were included in the models. As it contains the greatest number of cases, the IBE is omitted from the majors and the V1 is omitted from the test-versions.

Table 6. Binary logistic regression models (post-exam)

Dependent Variables	OEMC2		OECP2		OETT2	
	Model1	Model2	Model1	Model2	Model1	Model2
Constant	0.0404	3.8964***	0.5666	1.9591**	4.2436***	5.4344***
	(0.0002)	(16.8510)	(0.0618)	(5.7480)	(15.6454)	(24.1548)
SEX	-0.4763	–	0.2030	–	0.0480	–
	(0.7425)		(0.1066)		(0.0068)	
VHE	0.1716	–	2.1183**	2.0681**	17.7834	–
	(0.0201)		(4.2505)	(4.5677)	(0.0000)	
BAM	0.7783	0.8896*	-0.6157	–	-0.0088	–
	(1.9277)	(3.0665)	(1.1172)		(0.0003)	
V2	1.0032	–	-0.2520	–	1.0394	-1.7499**
	(1.1880)		(0.1176)		(1.1143)	(4.0541)
V3	-0.8880	–	1.1138	–	-0.4695	–
	(0.7328)		(0.8315)		(0.2163)	
V4	1.4996	–	0.5596	–	-1.4860*	–
	(2.0176)		(0.3466)		(2.9830)	
V5	1.1813	–	0.9776	–	-1.0244	–
	(1.4090)		(1.2239)		(1.3479)	
V6	1.5723	–	0.7372	–	-1.0045	–
	(2.3006)		(0.7963)		(1.5179)	
SESSION	-0.3244	–	-0.0494	–	0.3160	–
	(0.3416)		(0.0083)		(0.3254)	
MCSCORE	-0.3369***	-0.3204***	–	–	–	–
	(8.3230)	(8.8608)				
CPSCORE	–	–	-0.1436***	-0.1350***	–	–
			(27.8407)	(29.9074)		
TTSCORE	–	–	–	–	-0.0834***	-0.0825***
					(17.6237)	(19.8093)
N	140	140	148	148	156	156
OX^2	17.5147*	11.6846***	57.1656***	51.3195***	33.9976***	28.4151***
HLX^2	10.4497	4.2507	10.0661	8.1923	4.5668	6.9018
NR^2	0.2061	0.1403	0.4944	0.4522	0.3455	0.2938
NR^2 change	0.1087	0.1175	0.4273	0.4520	0.2099	0.2458

Note: $N = 162$; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$; Wald statistics in parentheses; OX^2 = Omnibus test X^2; HLX^2 = Hosmer – Lemeshow X^2; NR^2= Nagelkerke R^2; [a] NR^2 change after the test scores were included in the models. As it contains the greatest number of cases, the IBE is omitted from the majors and the V1 is omitted from the test-versions.

Table 7. Comparison of female and male self-estimations with independent *t*-test

Variables	Levene's *F*	*t*	Mean difference	Std. Error Mean	95% Conf. Int. of the Diff. Lower	Upper
DIFMC1	0.4194	-3.4189***	-1.7808	0.5209	-2.8094	-0.7522
DIFCP1	0.0012	-1.5065	-2.7107	1.7993	-6.2640	0.8427
DIFTT1	0.3611	-2.2520**	-4.4915	1.9944	-8.4301	-0.5528
DIFMC2	2.4947	-0.6915	-0.3545	0.5127	-1.3670	0.6580
DIFCP2	0.0097	0.0708	0.1133	1.6003	-3.0470	3.2737
DIFTT2	0.1340	-0.1385	-0.2412	1.7415	-3.6805	3.1981

Note: $N = 163$ (47 males and 116 females); * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 8. Pre- and post examination frequencies of under-, accurate and over-estimations

	Males									Females									Total								
	Multiple choice			Calculation problem			Total score			Multiple choice			Calculation problem			Total score			Multiple choice			Calculation problem			Total score		
	U	A	O	U	A	O	U	A	O	U	A	O	U	A	O	U	A	O	U	A	O	U	A	O	U	A	O
Pre	8	5	34	8	1	40	7	1	41	5	10	101	11	5	102	4	3	111	13	15	135	19	6	142	11	4	152
Post	10	6	31	10	3	36	9	2	38	11	16	88	22	13	82	15	4	98	21	22	119	32	16	118	24	6	136
X^2	0.4516			0.6788[a]			0.7033a			4.5246			9.5760***			8.0728**			4.2115			9.1915**			6.4260**		
LR	0.4522			0.6858			0.7105			4.5937			9.7797***			8.6790**			4.2379			9.3385***			6.5706**		
Φ	0.0693			0.0850			0.0865			0.1400			0.2036***			0.1869**			0.1138			0.1682**			0.1406**		
T	0.0034			0.0044			0.0048			0.0133**			0.0299***			0.0266***			0.0090*			0.0201***			0.0162***		

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$; U = underestimation, A = accurate estimation, O = overestimation; X^2 = Pearson chi-square; LR = Likelihood ratio; LL = Linear by linear association; T = Goodman – Kruskal tau (the dependent is the overestimation category variable); a = the minimum expected count for the accurate self-assessment was 1.5 for both pre- and post-examination cases.

Table 9. Pre- and post-examination size of under-, accurate and over-estimations

Variable	Sample	*N*	Mean	Std. Dev.	S. E. M.	95% Conf. Int. Diff.		*t*
						Lower	Upper	
DIFMC1 – DIFMC2	Male	47	0.2766	2.5427	0.3709	-0.4700	1.0232	0.7458
	Female	115	1.6957	2.4501	0.2285	1.2430	2.1483	7.4216***
	Total	162	1.2840	2.5525	0.2005	0.8879	1.6800	6.4024***
DIFCP1 – DIFCP2	Male	47	4.1702	7.2721	1.0607	2.0350	6.3054	3.9314***
	Female	115	6.9043	8.3193	0.7758	5.3675	8.4412	8.8999***
	Total	162	6.1111	8.1034	0.6367	4.8538	7.3684	9.5987***
DIFTT1 – DIFTT2	Male	47	4.4468	8.2720	1.2066	2.0180	6.8756	3.6854***
	Female	115	8.6000	8.8722	0.8273	6.9611	10.2389	10.3948***
	Total	162	7.3951	8.8804	0.6977	6.0172	8.7729	10.5990***

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.