Knowing HE standards: How good are students at evaluating academic work?

Jon Guest[a]* and Robert Riegler[b]

[a]*Aston Business School, Aston University, Birmingham, United Kingdom*

Correspondence details:

Jon Guest, Department of Economics, Finance and Entrepreneurship, Aston Business School, Aston University, Aston Triangle, Birmingham B4 7ET, United Kingdom. Email: j.guest1@aston.ac.uk.

ORCID ID:

Jon Guest: 0000-0001-6139-905X.

Robert Riegler: 0000-0002-0423-5080

Word count: 6,941

Knowing HE standards: How good are students at evaluating academic work?

To become effective learners, students need to develop good evaluative judgment skills. Unfortunately, numerous studies find that self-evaluation estimates provided by undergraduates often differ significantly from the marks awarded by the tutor. This suggests that students either have a rather poor grasp of the assessment criteria, or they find it difficult to apply standards to their own work because of their emotional investment. They may demonstrate a better understanding of standards when asked to judge the work of their peers. We use data from a cohort of 2nd year undergraduates to compare the ability of students to accurately self- and peer-evaluate an assessed essay. We find that peer evaluation is more accurate, on average, than the self-evaluation but shows greater dispersion, and there is limited evidence that misconceptions about standards are consistent across self and peer-evaluation.

Keywords: Peer-evaluation, self-evaluation, evaluation design, evaluation accuracy, evaluation consistency

1 Introduction

To become effective learners, students need to (a) gain an appreciation of the required standards for the course they are studying and (b) be able to effectively apply these standards to their own work and that of others. In other words, they must develop skills of evaluative judgement. As Sadler (1989) argues, they need to know what ‘good’/high quality work looks like, be able to recognise the gaps between their own work and a ‘good’ piece of work and take appropriate action to close any gaps they have identified.

How can we measure evaluative judgement skills? One approach is to ask students to self-evaluate (SE) the quality of their assessed work and compare the results with the marks awarded by the tutors (TE), assuming the work is both valid and reliable. Numerous studies, from a wide range of different disciplines take this approach and report widely differing results (Brown & Harris, 2013; Cassidy, 2007, Lew et al., 2010).

Most studies, including those with economics undergraduates, find that most students overestimate the quality of their own work (Grimes, 2002; Guest & Riegler, 2017; Nowell, 2007). This suggests that beliefs about course standards are often below their actual levels. In other words, students think it is easier to earn higher marks than it really is. This seems to be especially true for lower attaining students (Grimes, 2002; Guest & Riegler 2017). However, issues around research design (incentives/ impression management bias) and self-evaluation may bias these results. For example, investing large amounts of effort in a piece of work can generate emotions that make it difficult for students to apply self-evaluative judgements in a consistent and dispassionate manner. One way to test this argument is to see if students find it easier to apply relevant academic standards to the work of others (i.e. peer-evaluation (PE)) where emotional attachment issues are less likely. An alternative possibility is that students will find SE easier as they are likely to invest far more time in researching, thinking and writing their own assignment than that of their peers. The impact of this additional investment may outweigh any emotional attachment bias so making SE more accurate than PE.

This paper uses data from a cohort of 2nd year undergraduates and compares their ability to SE and PE an assessed essay. The analysis focuses on three issues: (1) Is PE more accurate than SE i.e. can students apply evaluative judgements to their peers work more effectively than their own work? (2) Is any bias consistent across both SE and PE, i.e. do students who overestimate the quality of their own work also significantly overestimate the quality of their peers and vice versa? (3) What factors determine the accuracy of PE?

2 Literature review

The existing research on self and peer-evaluation can be grouped into two broad themes. One theme focusses on the role of SE and PE as a pedagogic tool and concentrates on design issues that maximise the positive impact of these activities. The accuracy of SE and PE is not of central importance. A second theme focuses more on the data generated by SE and PE and the extent to which they are consistent with TE. This provides information on some important issues such as the evaluative skills of students and the extent to which they internalise course standards. This study focuses on this second theme.

Boud and Falchikov (1989) conducted a widely cited survey of SE accuracy, that reviewed over 50 articles from a wide range of disciplines. One key finding is a tendency for lower achieving students, as judged by the marks awarded by the tutor, to overestimate the quality of their work. Some studies have focussed specifically on the SE skills of economics undergraduates, for example Grimes (2002), Nowell and Alston (2007), Ferraro (2010) and Guest and Riegler (2017). These papers find similar results. For example, Grimes (2002) reports that 64% of students in a Principles of Macroeconomics module provided SE estimates that were greater than the TEs. Guest and Riegler (2017) discuss these studies in more detail.

Falchikov and Goldfinch (2000) carried out a widely cited meta-analysis of PE studies and found a mean correlation of 0.69 between PE and TE. Somewhat surprisingly, the accuracy of PE did not vary by subject area or by the level of the course. In addition, having a single student peer-evaluate the quality of the work was just as accurate as having multiple students carry out the task. Li et al. (2016) conducted a similar meta-analysis of more recent studies and found significant variation in the results with reported Pearson correlations varying from -0.19 to 0.98. The average correlation was 0.63, similar to that reported by Falchikov and Goldfinch. The authors found the correlation between PE and TE was stronger when (a) participation was voluntary as opposed to compulsory, (b) the course was at graduate rather than undergraduate level, (c) the process was paper-based rather than computer-assisted, (d) feedback comments were provided instead of just a mark, and (e) the process was not anonymised. Mostert and Snowball (2013) appears to be the only published paper that uses PE with economics students. This research focuses on the impact on learning as measured by student perceptions of the activity rather than trying to measure accuracy.

Very few studies have attempted to compare how accurately the same group of students SE and PE the same piece of written work. Some that have use very small sample sizes. For example, Lindblom-Ylänne et al. (2006) compared SE, PE and TE of just fifteen students on a law degree while Kiliç (2016) conducted a similar study with fifteen students on a teacher-training course.

Stefani (1994) is the most widely cited study of both SE and PE that uses a larger sample. Eighty students completed a SE exercise while fifty-seven students completed a PE exercise on the same assessment – a laboratory practical report. Both SE and PE were remarkably accurate which may reflect the nature of the work. The average SE score was 72.7 percent compared with an average TE of 75.3 percent while the average PE score was 74.4 percent compared with a TE score of 74 percent. The correlation coefficient between SE and TE was 0.93 whilst that between PE and TE was 0.89. However, the group of students that completed the SE was not the same as the one that completed the PE. Stefani commented that ‘An ideal situation would be self, peer and tutor assessment occurring within the same class.’

One of the only studies the authors could find with a larger sample size, the same students and a written assessment was that by Hassan et al. (2014). As part of this research, they analysed the SE and PE accuracy of 80 accounting students enrolled on a module in business law. The assessment was an essay-writing task that was part of an in-class test. On average, both SE and PE were greater than TE although PE was slightly more accurate (66.3 percent, 63.1 percent and 56.8 percent respectively). The variance in PE was also greater than TE and SE with standard deviations of 17.16, 16.47 and 14.55 respectively. A summary of the research is available in Table A1 of the appendix.

The literature review finds a very limited number of papers that compare the ability of the same group of students to accurately self- and peer-evaluate their work. Those that are published, typically use small sample sizes and very few are based on an extended piece of written work. They also provide little incentive for students to take the evaluation activities seriously. For example, Hassan et al. (2014) found that many students were unhappy with the failure of many of their peers to engage with the exercise. They conclude that ‘Going forward, it would be useful if some system could also be devised that motivates more students’.

This study attempts to add to the literature by comparing the ability of a relatively large sample of students to self- and peer-evaluate an extended piece of writing. It also provides incentives for students to (a) engage in the process and (b) overcome impression management issues i.e. concerns that their evaluations will influence the marker.

3 Methodology and data

3.1 Research Design

The data for this study comes from a 1,500 word assessed essay that students had to write as part of the coursework on a module in microeconomics.[1] This is a mandatory second year module on an economics undergraduate degree programme. The student population on this module was 131, of which 110 students agreed to participate in this study.[2] The sample consists of 33 female students (30%), 76 students with an UK education background (69%), 8 repeating students (7%) and 11 visiting students (10%).

To complete the assignment, students had to submit electronic versions of three pieces of work on the Virtual Learning Environment (VLE): (i) an essay, (ii) a self-evaluation grading sheet and (iii) a peer-evaluation grading sheet. The deadline to submit the SE form was the same as the essay. PE forms had to be submitted a week later.[3]

The tutors provided detailed written and verbal guidance for all three pieces of work. Much of this focussed on the assessment criteria, as some previous studies found that explicit and well-understood criteria improve the accuracy of evaluations (Falchikov & Boud, 1989). The day after the submission deadline, each student received an anonymised version of a classmate’s essay. The random distribution function of the ‘The Workshop’ tool in Moodle was used for this purpose. The students had one week to submit a completed peer-evaluation form. Both students and tutors used the same assessment criteria for all of the marking activities.

Researchers need to consider a number of factors when carrying out this type of research. These include (a) the validity and reliability of the assessment, (b) the

[1] The authors completed the University Ethical Approval process and the project gained approval.

[2] The characteristics of students who did not agree to participate in the exercise appeared to be random.

[3] The order of the self- and peer-evaluation could have an impact on the overall results and will be discussed on pages 11-12.

incentive for students to provide SE and PE that truly reflect their evaluation beliefs, (c) sample selection, (d) the ordering of the activities, (e) anonymity, and (f) the number of assessors.

To measure accuracy, the coursework must be both valid and reliable. An assessment is valid if it accurately measures what it purports to measure. This is typically the students' understanding of some specific part of the module content and/or particular leaning outcomes. The nature, wording, meaning and interpretation of the essay question was discussed with a number of experienced colleagues before its release to students in an attempt to address this issue. The assessment is reliable if all the tutors marking the work have an objective and consistent conception of standards. In other words, they all agree on what good work looks like and can apply this standard/scale to different pieces of work in a consistent manner. Marking an open response style of assessment, such as an essay, introduces a greater level of subjectivity into the marking process and potentially reduces reliability. Many studies have found large variations in the marks awarded by different tutors for the same essay (Bloxham et al., 2016). The potential for inconsistency is lower in this study as there were only two tutors marking the assessments. Great care was also taken to moderate all essays in an attempt to maintain consistency between the two markers.

Another major issue is the extent to which the SE and PE data truly reflect the evaluation beliefs of the students. This may not be the case because of (a) low cognitive effort levels, and (b) impression management concerns.

Applying assessment criteria is a cognitively demanding and time-consuming task. In a study by Hanrahan and Isaacs (2001), the students report finding the process difficult because of (a) limited prior experience of marking and (b) uncertainty over standards. The participants also claimed that many of their peers had not taken the evaluation activities seriously because they did not count towards the final grade and had simply chosen what they believed to be an average mark. Hassan et al. (2014) finds similar results. This suggests that without appropriate incentives, the students will not exert enough effort to carry out the activity effectively. The previous experience of the authors is also consistent with these findings i.e. grading sheets completed in the last few minutes prior to the deadline.

Even if students do exert appropriate effort levels, their evaluations may still not reflect their true beliefs because of marker-influence concerns. Students who believe they have submitted a low quality piece of work may avoid conveying this information because they fear it will negatively influence the assessment judgments of the tutor. They may instead submit deliberately inflated SEs. There is also a possibility that students who believe they have submitted high quality assignments may deliberately underrate the quality of their work for fear of appearing ‘big headed’.

In this research, the tutors use a marks incentive scheme to both reduce the potential for impression management bias and encourage students to exert effort on the evaluation activities. The design was influenced by the work of Race (2020) who suggests the use of a five percent incentive to encourage students to engage with feedback. The authors’ original intention was to duplicate this approach but a panel of internal assessors at the university judged this was too generous. Therefore, a three percent bonus was adopted and works as follows. If either the students’ SE or PE are within three percentage points of TE then three bonus marks are added to the TE mark.[4]

[4] For example, if a student estimated that an appropriate mark for their essay was 56 percent and the tutor awards it 55 percent then they would receive a final mark of 58 percent, i.e. 3 percentage points are added to the mark provided by the tutor. The same mark incentive was provided for peer-evaluation accuracy.

Therefore, if both SE and PE were accurate it was possible for the students to gain six extra marks. In a further attempt to minimise the chances of impression management bias, the students had to submit essay and self-evaluation forms via different links on the VLE. Both written and verbal guidance highlighted that tutors would not access either of the evaluation forms until the marking was completed.

Two other potential problems with the research design are the impact of (a) sample selection and (b) the ordering of the evaluation activities. If SE and PE are voluntary then only the most motivated students may submit completed forms and the data will not be representative of the whole cohort. This could significantly bias any results. To address this issue both SE and PE were 'gatekeeper activities' for the coursework. If the students failed to submit either of the evaluation sheets, they received a mark of zero for their essay.

The ordering of the activities could also influence the results, as there is evidence that evaluation accuracy improves with experience (Lopez & Kossack, 2007). Hence, the results may change depending on whether SE occurs before PE or vice versa. Interestingly, in Hassan et al. (2014) both evaluation activities took place in class, simultaneously. Therefore, students would have access to a peer's essay before making their final SE. In theory, a research design could test for this by having half the cohort undertake SE before PE while the other half undertake PE before SE. Unfortunately, this was impossible given the constraints in this research but could be an interesting avenue for future work. The decision to implement SE before PE was taken in this study as many students carry out informal SE as they are researching and writing their assessments. Therefore, to some extent, SE will always occur before PE, even if the formal SE set by the tutor occurs afterwards.

One issue specific to the design of PE is anonymity. As previously discussed, Li et al. (2016) found studies that use non-anonymous PE had estimates more closely aligned to TE. One potential explanation for this finding is that when the process is anonymous the students take it less seriously. However, PE was conducted anonymously in this study as (a) the marks bonus provides an incentive for students to take it seriously, and (b) a number of students reported being uncomfortable with a non-anonymised process. Li et al. (2016) also reported greater correlation between PE and TE in studies where participants had to provide both comments and marks. This study took the same approach.

3.2 Empirical Analysis Method

The analysis section consists of three parts. Firstly, an initial comparison is made between SE, PE and TE. This section focuses on the mean values and the variance in the data. Secondly, the study investigates whether any biases or misconceptions about course standards are consistent across both SE and PE. To answer this question this section presents the results of correlation and simple univariate regression analysis. Finally, a more comprehensive econometric approach is used to identify the determinants of peer-evaluation inaccuracy. The following model is estimated:

$$PE\ Inaccuracy_i = \beta_0 + \beta_1 X_i + \beta_2 Own\ Essay\ Mark_i + \beta_3 Peer\ Essay\ Mark_i + \beta_4 Over\ Marker_i + \beta_5 Under\ Marker_i + \varepsilon_i \tag{1}$$

The measure of *PE Inaccuracy* is the absolute difference between the marks awarded by the student and the tutor. The larger the absolute value of this variable, the greater the level of inaccuracy. However, this measure does not differentiate between those who either grade more or less generously than the tutor. To see if this has any effect both *Over Marker* and *Under Marker* dummy variables[5] are included. The *Over Marker* variable captures the impact of PE estimates that are at least three marks above the tutor mark whereas the *Under Marker* variable captures PE estimates that are at least 3 marks below. The X_i explanatory variable captures the following student characteristics: gender, repeating the module and UK schooling. One might expect those students who have previously studied in the UK to have a greater appreciation of the standard required at UK universities. From the experience of repeating the module, some students might develop a better understanding of the material. However, repeating students may be academically weaker and that might increase their inaccuracy. Some studies have also found a gender effect on self-confidence and accuracy when evaluating work (Beyer, 1999).

Own Essay Mark is the tutors' agreed mark for the students' own essay. This variable captures both the students' academic ability and the time/effort exerted in writing this particular essay. We expect this variable to have a negative impact on *PE Inaccuracy*. Spending more time on the essay, will develop expertise on the topic area and so assist accurate grading. We also anticipate that academically stronger students will have a better appreciation of the standards.

Peer Essay Mark is the tutors' agreed mark for their classmate's essay that the student evaluates. This variable might influence levels of inaccuracy in a number of conflicting ways. One possibility is the higher the quality of the work, the easier it is to evaluate. This may be the case for a couple of reasons. Firstly, if the essay is well structured and written, it takes less cognitive effort to read and mark. Secondly, higher

[5] The dummy variables enable us to compare the level of inaccuracy of both over markers and under markers compared to accurate students i.e. those with evaluation estimates within the range of +/- 3 percentage points of the tutor marks.

quality work may be easier to recognise than assignments of a lower standard. If this effect is present, then the *Peer Essay Mark* should have a negative effect on *PE Inaccuracy*. Another possibility is that risk-averse undergraduates will avoid awarding high marks. Even if they believe the work is excellent, students may not have the confidence to award a mark above 75 percent.

Another possibility is that students find it easier to peer-evaluate work that is of similar quality to their own. For example, a student who submitted a high quality essay may find it easier to grade another high quality as opposed to a low quality piece of work. To capture this effect, we employ another variable (*Diff in Essay Mark)*. This is calculated by subtracting *Own Essay Mark* from the *Peer Essay Mark*. A positive value for this variable indicates a situation where students evaluate a higher quality essay than their own. A negative value indicates a situation where students evaluate a lower quality essay than their own.

Three different models are estimated: Model 1 is the baseline model that is based on Equation 1. In contrast to Model 1, Model 2 controls for the difference in essay marks instead of controlling for Peer Essay Mark and Own Essay Quality separately. Finally, Model 3 adds a specific ability measure to Model 2 in an attempt to get a clearer picture of the relationship between the difference in essay marks and PE accuracy.

4 Analysis and results

4.1 Self vs peer-evaluation

As discussed in previous sections, SE may be more or less accurate than PE. Table 1 illustrates the mean values for TE, SE and PE. Both average SE and PE are greater than TE. However, whilst the average over-confidence of SE (3.13 marks) is statistically significant, the average over-confidence of PE (0.81 marks) is not (see Table 1).

	Variable	Obs	Mean	Std. Dev.	Min	Max	t test	sd test	*ρ*
Self-	Tutor Mark (TE_{SE})	110	56.51	9.83	3	78			0.65
evaluation	Student Mark (SE)	110	59.64	8.23	0	72	$p = .011$	$p = .065$	$p < .001$
Peer-	Tutor Mark (TE_{PE})	110	55.54	8.97	25	80			0.51
evaluation	Student Mark (PE)	110	56.35	12.31	0	87	$p = .574$	$p = .001$	$p < .001$
SE & PE	SE - TE_{SE}	110	3.13	7.65	-18	20			0.15
inaccuracy	PE - TE_{PE}	110	0.81	10.88	-32	26	$p = .026$	$p < .001$	$p = .128$

Notes: 1. SE and PE rows: Unpaired two-way t-tests were undertaken with unequal variances assumed.
2. Due to the random allocation of 131 essays, the mean tutor marks differ as the sample of essays that were self-evaluated was not identical to the sample that was peer-evaluated. For example, a student's essay that was self-evaluated may not be in the peer-evaluation sample.
3. SE & PE inaccuracy rows: Inaccuracy is measured be taking the difference between student's and tutor's evaluations. Paired one-way t-test was undertaken to test whether SE inaccuracy is larger than PE inaccuracy.

Table 1 Difference between the mean SE and PE estimates and mean tutor marks

These simple mean values suggest that PE is more accurate than SE. However, there is far greater dispersion in PE inaccuracy as shown by the standard deviation data in Table 1[6]. Figure 1 is another useful way to illustrate this dispersion. The 45-degree lines show all the observations where the student evaluations are identical to those of the tutor. Any points above the 45-degree line in both diagrams illustrate students who over-estimate the quality of the work whilst any points below illustrate those who underestimate the quality of the work. Around half of the students provide overconfident self-evaluation estimates and we find a bunching effect around 60 percent. The middle fifty percent of the sample (i.e. those between the 25 to 75 percentiles) provide self-evaluation estimates of between 56 and 64 percent. While many students overestimate the quality of the work of their peers, this bunching effect is not as strong.[7] The middle fifty percent provide PEs of between 50 to 65 percent. These

[6] While we do not find evidence that SE estimates are more dispersed than the TE, PE estimates are significantly more dispersed than the TE (see Table 1, Std. Dev. column). The PE estimates are also significantly more dispersed than the SE estimates (not shown in the Table).

[7] The number of overconfident self-evaluators (53) is slightly higher than the number of generous peer-evaluators (48).

results clearly show students using a wider range of marks when judging the quality of their peers' work and is consistent with the findings of Hassan et al. (2014) and Kearney et al. (2016).

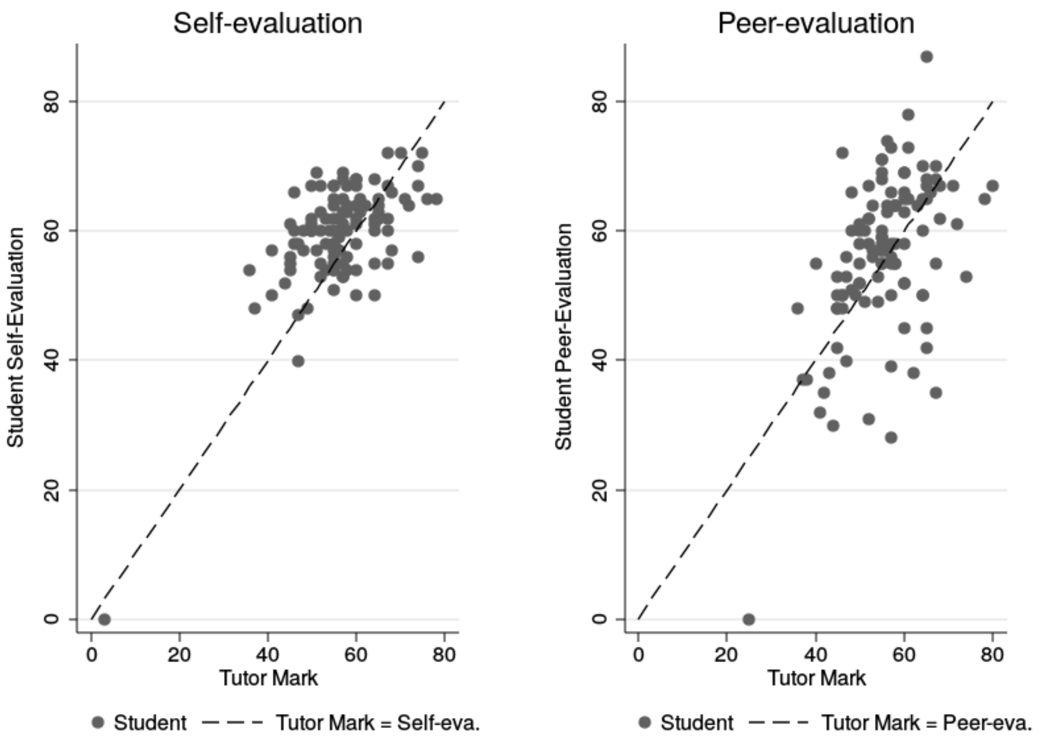

Figure 1 Comparison of SE and PE estimates of students (the two outliers at the bottom of the figures are dropped from the statistical analysis)

This analysis of the data shows that simple comparisons of the mean scores is misleading. The greater dispersion of PE inaccuracy in comparison to SE inaccuracy suggests that many students find it also difficult to judge the quality of their classmates' work.

4.2 Is there consistency between peer- and self-evaluation?

Another interesting question is whether any misconceptions about course standards remain consistent across both SE and PE estimates. For example, are students who significantly overestimate the quality of their own work also more likely to overestimate the quality of their peers? If this is the case, then SE estimates may be a good indicator of misconceptions. A more worrying alternative is that many students have an ill-defined conception of the standards and effectively make random guesses. In this case, there will be no relationship between SE and PE and the misaligned beliefs between tutors and student will be much more difficult to close. To investigate this issue, we split the data into four different quadrants as illustrated in Figure 2. *Quadrant B* contains data on the students who both overestimated the quality of their own work and that of their peers while *quadrant C* illustrates those who underestimated both the quality of their own work and that of their peers. In both of these quadrants, there is some level of consistency to the bias. *Quadrant A* contains data on those students who underestimated the quality of their own work while overestimating the quality of their peers while *quadrant D* contains the data on those who overestimated the quality of their own work while underestimating the quality of their peers. Both of these quadrants, illustrate cases of inconsistency.

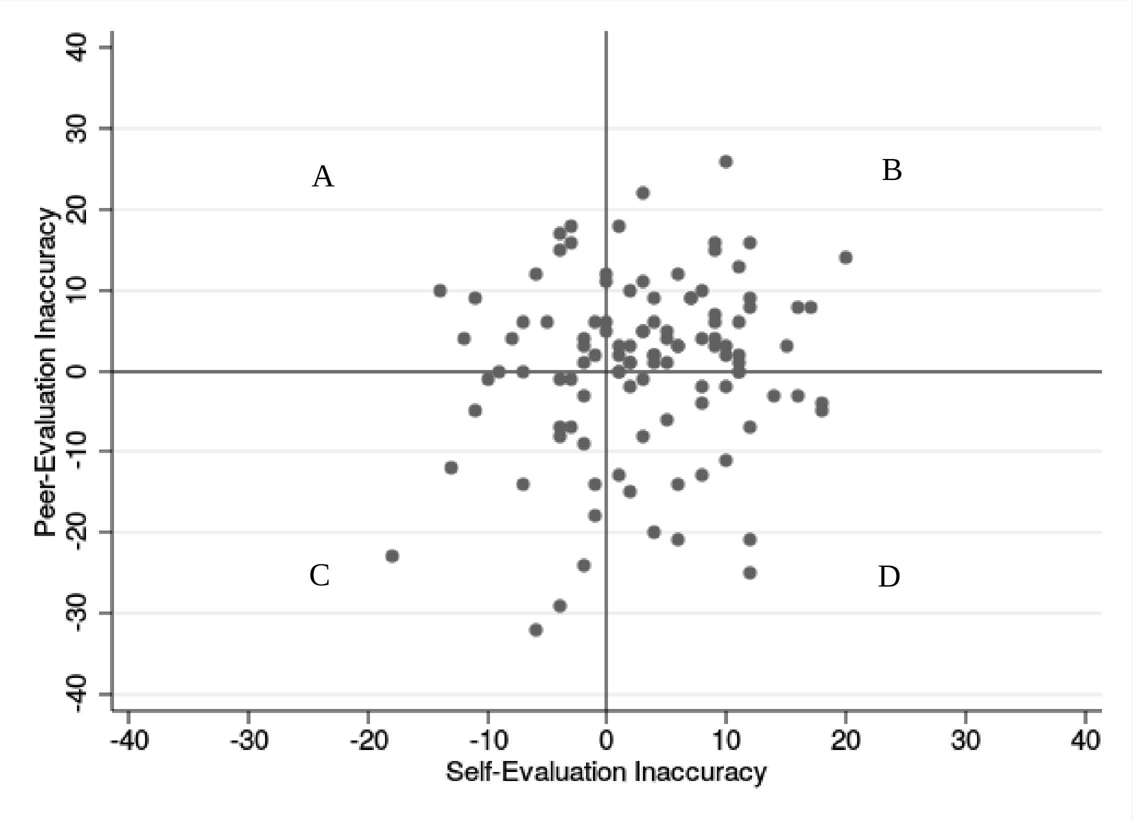

Figure 2 Relationship between generosity and confidence

The results in Figure 2 show some consistency. For example, 42% of students both over-estimated the quality of their own work and that of their peers (quadrant B) while 15% of students under estimated both the quality of their own and that of their peers (Quadrant C). However, 15% of students under-estimated the quality of their own work and over-estimated the quality of their peers (quadrant A) while 19% over-estimated the quality of their own work while under-estimating the quality of their peers work (quadrant D). Even though we found some level of consistency, the correlation coefficient between SE and PE inaccuracy[8] is only 0.18 and this is statistically non-

[8] This inaccuracy measure takes on both positive and negative values. Evaluations below the tutor mark have a negative value while evaluations above the tutor mark have a positive value.

significant at the 5 percent significance level.[9] A simple regression of PE inaccuracy on SE inaccuracy provides a quantitative measure on the strength of the relationship between the two variables – see Table A2 in the appendix for more details. The results show that a single percentage point increase in SE inaccuracy is associated with a 0.25 percentage point increase in PE inaccuracy, but it is again statistically non-significant at the 5 percent level. However, as the respective *p*-value is 0.065, the relationship needs to be investigated in more detail because of the potential impact of omitted variable bias. This is the focus of the next subsection.

4.3 What factors determine peer-evaluation accuracy?

The regression results based on Equation 1 are presented in Table 2.[10] The dependent variable is *PE Inaccuracy* - the absolute difference between the marks awarded by the student and the tutor. Values closer to zero for this variable indicate PEs that are closer to the judgements of the tutor. Model 1, in Table 2, is the base-line regression that controls specifically for X_i (a number of student characteristics including gender), *Own Essay Mark* (the tutor's agreed mark for the student's own essay), and *Peer Essay Mark* (the tutor's agreed mark for the essay the student peer-evaluates). As anticipated, there is a significant negative relationship between *Own Essay Mark* and *PE Inaccuracy*. If a student develops greater expertise in the topic area and/or is academically more able then it improves their ability to evaluate the work of their peers.

[9] One outlier was dropped from the sample to ensure reliable results. Therefore, this calculation used 109 observations.

[10] One outlier was dropped from all regressions.

Variable	**Model 1**	**Model 2**	**Model 3**
Female	0.83	0.77	0.86
	(0.98)	(0.98)	(0.89)
UK Education	-0.22	-0.18	-0.33
	(0.96)	(0.97)	(0.91)
Repeater	-2.89	-2.69	-4.14***
	(1.77)	(1.75)	(1.27)
Own Essay Quality	-0.10**		
	(0.05)		
Peer Essay Quality	0.03		
	(0.06)		
Diff. in Essay Mark		0.07**	0.09**
		(0.03)	(0.04)
Prev. Econ Module Mark			-0.07
			(0.06)
Over-Marker (>3%)	8.40***	8.49***	8.64***
	(0.85)	(0.86)	(0.88)
Under-Marker (<-3%)	11.93***	11.65***	11.77***
	(1.58)	(1.53)	(1.45)
Constant	5.44	1.76**	5.14
	(4.77)	(0.70)	(3.10)
No. of Observations	109	109	95
R^2	0.49	0.49	0.53

*Legend: * p<0.1; ** p<0.05; *** p<0.01*
Due to the presence of heteroscedasticity in all models, robust standard errors are used
Robust Std. err. in brackets
Model 1: Baseline regression
Model 2: Employs difference in Peer and Own Essay mark as explanatory variable
Model 3: Adds a specific ability measure to model 2.

Table 2 OLS regression results for the determinants of peer-evaluation inaccuracy

The student characteristics of gender and progressing through the UK education system were non-significant in all three models. There is a consistently negative sign for the coefficient of the repeater dummy, but it is only statistically significant for Model 3.

The coefficient for the *Peer Essay Mark* is positive but statistically non-significant. Therefore, this study finds no evidence that students are better able to evaluate higher quality essays than lower quality essays. The coefficient on the *Under-Marker* dummy variable is also greater than on the *Over-Marker* dummy variable.[11]

[11] In all estimated models, the Under-Marker coefficients are larger than the Over-Marker coefficients at the 5% significance level.

This means that the inaccuracy of under-marking students is on average greater than those who over-mark.

Model 2 is a second regression that includes the *Diff in Essay mark* variable to test for the possibility that students find it more difficult to peer-evaluate an essay that differs in quality to their own. The results indicate a positive but relatively small effect, which is statistically significant.

Students appear to be more inaccurate at peer evaluating an essay that has a higher TE than their own, e.g. a student who peer-evaluate an essay that is ten marks higher than her own essay mark, will be on average 0.7 marks more inaccurate. Although relatively small, there are two potential explanations for this finding: Firstly, students find it easier to evaluate work that is either of similar or lower quality than their own. Secondly, the result is a statistical artefact of the data. Students who have submitted higher quality essays are more likely to evaluate a peer's essay that is of lower quality than their own. Therefore, it could be their ability rather than the lower quality peer essay that causes the more accurate evaluation estimates. In an attempt to reduce the potential explanatory power of this latter point, a specific ability measure is added to Model 3. We chose the student's module mark from a previous intermediate economics module as this was judged to be the best indicator of academic ability for the essay in this study. As shown in Table 2, the inclusion of this variable does not change the overall results. The new ability measure has the expected negative sign, even though it is not statistically significant, and the tutor mark differences remain positive and statistically significant.[12]

[12] We also tried other ability measures, e.g. the exam marks of this module (Intermediate Economics 2), Intermediate Economics 1 and Econometrics, as well as the overall module

In summary, the mean PE estimates are more accurate than the SE estimates. However, there is much greater variation in PE inaccuracy. There is limited evidence that misconceptions about course standards are consistent across both SE and PE. Finally, a more comprehensive statistical analysis reveals that students who submit higher quality essays are more accurate at evaluating the work of their peers. We also find evidence that students find it more difficult to accurately PE work that has a higher TE than their own.

5 Conclusion

To become effective learners, students need to acquire good evaluative judgment skills. In other words, they need to develop a good understanding of the relevant academic standards and be able to apply these standards to their own work and the work of others. A whole stream of research has tried to measure capabilities in this area by asking students to self-evaluate their work and compare the results with the marks awarded by tutors. The results vary significantly with many studies finding a tendency for students to overestimate the quality of their own work. This suggests that beliefs about course standards are often below their actual levels. However, these results may be sensitive to research design issues and particular issues students face when trying to apply standards to their own work in a dispassionate manner.

To test this hypothesis we compare the ability of a relatively large number of students to both SE and PE an assessed piece of extending writing. Care was taken with the research design to address a number of important issues such as validity, reliability, incentives, impression management, sample selection, anonymity and the number of

marks of Intermediate Economics 1 and Econometrics. All coefficient were negative but only the exam mark for this module was statistically significant.

peer assessors. There are two competing arguments concerning the relative accuracy of SE and PE. On the one hand, PE may be more accurate than SE as it lacks the same level of personal involvement so students may find it easier to apply their understanding of the standards in a more objective manner. An alternative argument is that, PE may be less accurate as students are likely to invest more time in researching and writing their own essay so improving their ability to evaluate the work.

The simple mean values suggest that PE is more accurate than SE. However, this statistic on its own can be misleading. PE inaccuracies are significantly more dispersed than SE inaccuracies indicating that many students do not find PE any easier than SE.

We were also interested to see if any biases or misconceptions about course standards are consistent across both SE and PE. If this is the case, then SE may be a good indicator of the extent to which students comprehend course standards or any misapprehensions that they have. We find only limited evidence that any biases are consistent across both types of evaluation. This further supports the idea that the students in our sample are uncertain about course standards.

Finally, a more comprehensive statistical analysis reveals a number of factors that influence the accuracy of PE. Students who submit higher quality essays (based on TE) appear to be more accurate at evaluating the work of their colleagues whereas the quality of the essay they peer evaluate appears to have no impact. Differences between the quality of their own work and the quality of the work they evaluate do seem to have a small positive effect. Students appear to be more inaccurate at peer evaluating an essay that has a higher TE than their own and more accurate at peer evaluating an essay that has a lower TE than their own.

One major limitation with the study is that we were unable to test whether the ordering of the evaluation activities influence the results. This is an interesting avenue for future work.

Anecdotal evidence[13] also suggests that the three-percentage point incentive scheme might not provide strong enough incentives for the majority of students to fully engage with the activities and/or overcome impression management concerns. Another area for further research would be to carry out similar evaluation activities with stronger mark incentives to see if it has an impact on engagement and accuracy. Also, due to the cross-sectional nature of the data, we cannot control for unobserved time-invariant student specific characteristics. Therefore, a larger research project could follow students over time and employ a panel-data estimator.

References

Beyer, S. (1999). Gender differences in the accuracy of grade expectancies and evaluations. *Sex Roles, 41*(3-4), 279-296.

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let’s stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466-481.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*(5), 529-549.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322.

Ferraro, P. J. (2010). Know thyself: Competence and self-awareness. *Atlantic Economic Journal, 38*(2), 183-196.

[13] Comments made by students in two informal focus groups.

Grimes, P. W. (2002). The overconfident principles of economics student: An examination of a metacognitive skill. *The Journal of Economic Education, 33*(1), 15-30.

Guest, J., & Riegler, R. (2017). Learning by doing: Do economics students self-evaluation skills improve?*, International Review of Economics Education, 24*(2017), 50-64.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self-and peer-assessment: The students' views. *Higher Education Research & Development, 20*(1), 53-70.

Kearney, S., Perkins, T., & Kennedy-Clark, S. (2016). Using self-and peer-assessments for summative purposes: analysing the relative validity of the AASL (Authentic Assessment for Sustainable Learning) model. *Assessment & Evaluation in Higher Education, 41*(6), 840-853.

Kiliç, D. (2016). An Examination of Using Self-, Peer-, and Teacher-Assessment in Higher Education: A Case Study in Teacher Education. *Higher Education Studies, 6*(1), 136-144.

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., & Suen, H. K. (2016). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245-264.

Lindblom-ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer- and teacher-assessment of student essays. *Active Learning in Higher Education, 7*(1), 51-62.

Mostert, M., & Snowball, J. D. (2013). Where angels fear to tread: Online peer-assessment in a large first-year class. *Assessment & Evaluation in Higher Education, 38*(6), 674-686.

Nowell, C., & Alston, R. M. (2007). I thought I got an A! Overconfidence across the economics curriculum. *The Journal of Economic Education, 38*(2), 131-142.

Race, P. (2020). *The lecturer's toolkit: a practical guide to assessment, learning and teaching* (5th ed.). Routledge.

Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69-75.

Appendix

Authors	Type of Assessment	Evaluation Task	Explicit Incentive	Sample Size	Findings
Ferraro (2010)	3 multiple choice exams	SE, voluntary	Yes	105	Overconfidence, but positive correlation between competence and SE accuracy.
Grimes (2002)	Exams	SE, voluntary	No	253	Overconfident students were less accurate than underconfident students.
Guest and Riegler (2017)	Individual written Essay	SE, compulsory	Yes	78	Overconfidence, but SE accuracy slightly improves over time
Hassan et al. (2014)	Presentations and in-class test essay	SE, PE compulsory	No	18 (pres.) 80 (essays)	SE and PE are inaccurate, both averages are above TE average
Kiliç (2016)	Presentation	SE, PE compulsory	Yes	15	PE is found to be significantly higher than SE and TE. TE and SE are similar
Lindblom-Ylänne et al. (2006)	Essays	SE, PE compulsory	No	15	SE and PE was similar to TE
Mostert and Snowball (2013)	Individual Essay	PE, compulsory	No	400	Evaluation accuracy was not assessed
Nowell and Alston (2007)	Module mark, consisting of Exam, Tests, Homework.	SE, voluntary	No	715	Overconfidence. No statistically significant relationship between GPA and overconfidence
Stefani (1994)	Laboratory practical report	SE, PE compulsory	Yes	80 (SE) 57 (PE)	Both PE and SE were accurate

Table A1 Summary of the current literature on SE and PE

Dep. Var.: Peer-evaluation inaccuracy	Coefficients
Self-evaluation inaccuracy	0.25*
	(0.13)
Constant	0.3
	(1.09)
No. of Observations	109
R^2	0.03

*Legend: * p<0.1; ** p<0.05; *** p<0.01*
Std. err. in brackets
Both inaccuracy measures can take positive and negative values

Table A2 Simple OLS regression of PE inaccuracy on SE inaccuracy