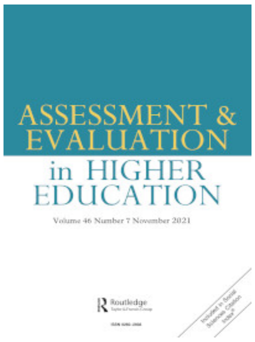

Assessment & Evaluation in Higher Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

Flipped small group classes and peer marking: incentives, student participation and performance in a quasi-experimental approach

Rabeya Khatoon & Elinor Jones

To cite this article: Rabeya Khatoon & Elinor Jones (2021): Flipped small group classes and peer marking: incentives, student participation and performance in a quasi-experimental approach, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2021.1981823

To link to this article: https://doi.org/10.1080/02602938.2021.1981823

Published online: 18 Oct 2021.

https://doi.org/10.1080/02602938.2021.1981823

Routledge
Taylor & Francis Group

Check for updates

Flipped small group classes and peer marking: incentives, student participation and performance in a quasi-experimental approach

Rabeya Khatoon[a] and Elinor Jones[b]

[a]University of Bristol, UK; [b]University College London, UK

ABSTRACT

This paper proposes a new way of flipping small group classes in quantitative courses by active reading and peer marking using the virtual learning environment. We aim to engage students in the learning material by attempting a problem followed by peer marking based on some given solution guideline before they are exposed to another similar problem to solve during the small group classes. We design a quasi-experiment to evaluate the effect of peer marking by introducing an incentive in one such problem set and not in the other. The solution to the class problem is to act as the 'incentive', to be released only to the participants of the peer assessment. Using the data of two units from two UK universities and 'incentives' as instrumental variables to participation, our quantitative findings reveal the effect of participating in one more peer marking as a 3% increase in final marks on average. The qualitative analysis based on focus group discussions shows that the process increases student engagement, satisfaction, confidence and overall learning responsibility. The challenges often lie to establish a clear understanding of the purpose and the process of peer marking to ensure student buy-in to the system.

KEYWORDS
Flipped classes; peer marking; incentives; quasi-experiment

Introduction

With technological advancement, the environment is increasingly supportive of 'active' learners rather than 'passive' listeners in higher education, where traditional lectures are often 'flipped' and the live sessions are designed as problem-solving classes. Flipped lectures often combine video-based learning outside the classroom which matches with students' media habits (Roach 2014) and is often favoured over reading text-based materials (Snyder, Paska, and Besozzi 2014). This allows students to engage with the material at their own pace and frees up classroom interaction time for active learning (Lo and Hew 2017). However, as flipping or partial flipping is becoming increasingly popular, the resulting module virtual learning environment (VLE) can become dense with learning material, thus risking students accidentally missing or intentionally ignoring these items. This issue could be addressed with a carefully planned 'incentive'.

Self and peer-assessment is also becoming a common tool in higher education, which supports modern cognitive learning theory (Ambrose et al. 2010; Gaynor 2020; Zheng, Zhang, and Cui 2020). The peer-review process in particular is recognized to enhance student learning by

CONTACT Rabeya Khatoon rabeya.khatoon@bristol.ac.uk

promoting cognitive skills, self-reflection, and intrinsic motivation with the recognition and even ‘gamification’ of learning.

In this paper, we propose a new way of flipping the classroom with online peer assessment. To our knowledge, peer assessment hasn’t been used as a tool to flip the classroom; generally, these are set within or after class time and not before (e.g. Fu, Lin, and Hwang 2019). We believe that students are more likely to engage with the learning material outside the classroom if it is in an assessment format (i.e. a peer review), even if this assessment is formative, as opposed to engaging with the videos and background reading offered in traditional flipped classrooms.

Our design flips small group classes for quantitative subjects. This is done via an additional relevant problem set for students to work on and then peer mark a student’s attempt using solutions that are provided by the lecturer. This activity takes place before a face-to-face small group class, during which students tackle similar problems to those encountered in the peer-marked problem set. The topics discussed during traditional small group classes often lag behind topics discussed in lectures. In our setting, there is a natural progression from lecture, through to immediate release of the problem set for students to work on over a short period of time, followed by the release of the solutions for peer marking so that the material is fresh in students’ minds, and then a small group class which uses the same ideas as in the peer marked problem set.

We redesign two quantitative courses in two UK higher education institutions covering both undergraduate and postgraduate students, so that both include two instances of the peer marking exercise followed by a small-group class. To assess students’ extrinsic motivation, we reward participants who take part in the peer marking process with additional solutions as ‘incentives’; specifically, solutions to the small group class problem sets, but only for one of the peer marking exercises. For the other peer marking exercise there is no incentive. In this paper, we report our findings from both quantitative and qualitative analyses of our experiment.

Literature review

Definition, goal, and variety

In recent years, flipped or inverted classrooms have gained popularity in higher education. Researchers have evaluated the use of various implementations of flipped classrooms in both large lectures (e.g. Setren et al. 2021) and small group classes (e.g. Becker and Proud 2018. See, e.g. Lundin et al. 2018 for a recent systematic review, and Strelan, Osborn, and Palmer 2020 for a meta-analysis). The effects of flipping the classroom on student performance is generally positive, but the extent of this varies widely depending on the design of flipping, student engagement, instructor motivation, group size, type of assessment and the methodology used (Strelan, Osborn, and Palmer 2020).

Peer assessment is also a widely used learning mechanism and is applied to almost all stages of education. In higher education, the efficacy of peer assessment depends on a variety of factors including whether the peer review is completed anonymously (Rotsaert, Panadero, and Schellens 2018; Kobayashi 2020), type of assessment, e.g. grading based or peer dialogue in both formative and summative assessments (Double, McGrane, and Hopfenbeck 2020), use of rubrics (Peters, Körndle, and Narciss 2018), paper-based versus online form of assessment (Wen and Tsai 2008) and rater (peer reviewer) training and frequency of peer assessment (Li et al. 2016). The effect on the peer assessor and the assesse are different too, with students generally gaining more through the process of giving feedback rather than receiving it (Li, Liu, and Steckelberg 2010).

Although both learning innovations have their benefits, flipping the classroom in terms of peer assessment has not received much attention in the literature. To the best of our knowledge, Ng and Fai (2017) is the only paper that explicitly involves both flipped classes and peer assessment, though peer assessment wasn't used as a tool to flip the classes. Ng and Fai (2017) used traditional ways of flipping based on pre-released videos while engaging students with the flipping content later by peer prepared quizzes and assessments. A clear challenge in both mechanisms is to ensure student buy-in and the indicators of student engagement with the mechanisms, especially something that could quantify the quality of engagement, still need improvement (Lundin et al. 2018).

An effectively designed active learning activity is essential to ensure timely engagement with the online flipped content. One of the primary goals of flipping is to free up time allocated for face-to-face teaching to allow for deeper understanding. However, this goal is not achieved as expected if the engagement rate with the flipped content is low. Flipping with online videos almost always provides a weaker incentive to engage regularly rather than last-minute cramming compared to live-only teaching deliveries (Donovan, Figlio, and Rush 2006; Lo, Hew, and Chen 2017). Flipping is not limited to providing some background videos only. The positive effects of flipped classroom on student achievement are also driven by those from active in-class engagement in problem solving (Albert and Beatty 2014; Calimeris and Sauer 2015; Setren et al. 2021). The system with flipped classes and active, in-class problem solving activities provides additional, carefully prepared tuition. If flipped videos are not drawing timely attention, the improvements on grades seen in the literature (Bergmann and Sams 2014; Becker and Proud 2018; Strelan, Osborn, and Palmer 2020) could only be due to the in-class activities. It is therefore important to separate the flipping effect from the additional tuition effect while quantifying the impact of the system on student grades.

Methods contrast

In terms of the methods used, not all the studies provide credible causal relationships between teaching innovations and student grades. Though randomized controlled experiments are the most acceptable, they are the toughest to be logistically and ethically permissible for the education sector and there are only a few in the literature.

The attempts in Deslauriers, Schelew, and Wieman (2011) and Yestrebsky (2015) on estimating the impact of the flipped classroom on student grades are considered as controlled experiments. However, they suffer from non-random assignment of students into the treatment and control groups and do not control for heterogeneity in instructor and student characteristics. Setren et al. (2021) claims their study to be the first to analyse video-based flipped classroom impact in a randomized control trial involving mathematics and economics subject groups. With individual instructor effect controlled for and random assignment of students in the traditional and flipped classrooms, they found that flipping provided short-term positive impact, though this varies with subject, student characteristics and teacher motivation. Alcalde and Nagel (2019) analysed the impact of peer instruction on student satisfaction and performance with randomly assigned treatment and control groups, and found that any positive effects are primarily short-term and driven by more interaction with the instructor rather than with peers.

Two recent meta-analyses which considered studies with a 'treatment' and 'control' group (and thereby used some form of experimental or quasi-experimental methods) are Strelan, Osborn, and Palmer (2020) who looked at the impact of the flipped classroom, and Li et al. (2020) on the effects of peer assessment. According to them, quasi-experimental studies based on credible identification strategies are more common and are often as reliable as controlled experiments. For example, Li et al. (2020), in their meta-analysis of 350 studies conducted between 1950 and 2017, found that the effect sizes from experimental (18 studies) and quasi-experimental studies (40 studies) on peer assessment are not statistically different.

Our method falls within the quasi-experimental category. We acknowledge that the data are observational rather than experimental in nature, and therefore the relationship between the intervention and outcome for a student is confounded by many factors. We use an instrumental variable technique to overcome this issue and estimate the (causal) impact of flipped classes using peer assessment on student examination performance.

Incentives

Researchers have reported various forms of incentives to engage students with out-of-class (flipped) activities. For example, Eager, Peirce, and Barlow (2014) and Setren et al. (2021) used summative out-of-class activities, Hung (2017) used quizzes with leader-boards at the start of in-class sessions, and Lai and Hwang (2016) consider self-regulated activities of goal setting and reflection to enhance engagement with flipped content. In a recent study, Huang, Hew, and Lo (2019) estimated the effect of gamification incentives for flipped classroom engagement and found that awarding participation-based and quality-based badges improved students' behavioural and cognitive engagement compared to the control group.

Likewise, the literature on peer assessment points to the struggle of engaging students with assessment. One common reason for lack of engagement with peer marking activities is a belief that they (the students) won't derive any benefit from the system, and that peer marking is an elaborate way to reduce staff workload (e.g. Hughes 2001). Studies have suggested various incentives for participation, such as marking reward (e.g. Weaver and Esposto 2012; Chevalier, Dolton, and Lührmann 2018), grade loss (e.g. Hughes 2001; Gillanders, Karazi, and O'Riordan 2020), and grading the reviewer based on the quality of their review (e.g. Gamage et al. 2017).

Our design combines flipping the classroom with using peer marking activities and setting out some incentives for participation and engagement. Lavecchia, Liu and Oreopoulos (2016) review several studies on the effects of various incentives on students, and found that participation incentives target lower-performing students and show more promise than grade-based incentives. Our context is to engage students with peer marking, and to keep students focused primarily on the content of the activity rather than the grade. We select formative rather than summative assessment with a soft participation incentive: the release of additional learning materials. It is likely to have less incentive as a result of being formative. This helps us limit issues such as reliability and validity in peer marking (Falchikov and Goldfinch 2000).

In addition, peer assessment can be more time-consuming to implement for summative assessment due to time spent in training the students, whereas formative assessment can be time-saving for the instructor while still benefitting students (Panadero 2016; Lynch and Schmid 2018). Unlike Setren et al. (2021), our study is based on small group problem-solving lessons so that the flipping is not to free up space for active learning. In the absence of flipping, the students were still supposed to solve the problems in classes with the help of the tutor, and thereby the effect of flipping on performance is not expected to come from the 'follow up in-class engagement'. Instead, we analyse the impact of giving peer feedback instead of the benefits of being assessed by a peer. Our design keeps the peer assessment anonymous, which is both more beneficial for learning (Li et al. 2020), valued by the students (Kobayashi 2020), and facilitates our large international cohort to enhance assessment and feedback experience (Chew, Snee and Price 2016).

Study design

Our study is based on two quantitative modules taught at universities in the UK, both of which running in the 2018–19 academic year. One is a second-year undergraduate (UG) statistics module with approximately 100 students, while the other is a postgraduate (PG) module in economics taught to 340 students. For both, we set up formative peer marking activities which then led to small-group flipped classes.

In each module we set up two problem sets as homework, each with specified deadlines, which the students submitted to their module virtual learning environment (VLE) in any format, e.g. MSWord document, pdf, or pictures. Submitting solutions to these problem sets is not compulsory. After the submission deadline, a peer assessment environment was created using, for example, the 'Turnitin Peermark assignment' facility in Blackboard and 'self and peer assessment' in Moodle. For both modules, the allocation of submissions to student reviewers was done randomly and anonymously. We chose anonymous peer marking to keep the focus of the assessor on the content rather than how the assesse will take their assessment (Panadero 2016). The most important factor in deciding to do this anonymously is that students are often anxious about what others think of their work: they are worried in case they make 'silly' mistakes or that they are not as clever as their peers. Anonymity makes them more likely to take part.

Each student is allocated one script to peer mark per problem set. The student reviewers are provided with solutions which are released immediately after the submission deadline. To help students to review their peer's work adequately we create rubrics, including multiple choice and free response feedback. This guides the students through the marking process and offers advice on what to emphasize when marking. We keep a record of students who completed the peer marking exercise.

The problem sets and subsequent peer review takes place before students attend small group classes. The small group classes consider a set of questions that are in some sense similar to the problem sets that they had the opportunity to attempt and peer mark. This is the 'flipped' element of the experiment: the students engage with the material (problem sets) before attending a class which is dedicated to *using* rather than *teaching* new material. Thus, the students are exposed to a new problem to solve in small groups with the tutor after some preparation via individual participation in the peer-review process.

To evaluate the effect of engagement with the peer reviewing (if any) on grades, we offer an incentive to only one of the problem sets (PS1). For PS1 we release full solutions to the in-class problems to those who completed the peer marking activity. For the problem set without incentive (PS2), the solutions to the class problem (C2) are released to all students regardless of whether they completed the peer marking exercise. Figure 1 presents the study design timeline.

The students are informed from time to time with verbal notifications during the lecture and via VLE announcements about the incentives and the formative nature of the assessments. Each of the problem sets and assignments is marked by clear statements of the incentives and the procedure to follow. Note that the overall process is designed so as not to discriminate against non-participating students. For the case with an incentive, PS1, the non-participants don't get direct access to the solutions to C1, but this was not prohibitive. All students, irrespective of their participation in the peer marking process, are offered small group classes, where they are exposed to the solutions to C1 in the presence of the class tutor. Moreover, students are aware that they can get the solutions indirectly from their participant peers.

Our research objective is to evaluate the process of flipped classes via peer marking, as well as to quantify the effect of engaging with peer marking on student grades. In absence of an ideal system where students are allowed to reflect on the feedback received, our set up captures the benefits of 'giving' feedback and engaging with the system, rather than the use of feedback in helping peers with their understanding. The learning gains could come from various angles of engagement: to be able to see peer's attempts while reviewing, to note down different ways of addressing similar issues, common mistakes, and, finally, the responsibility to know more before marking peer's script. We put less weight on the quality of feedback received, which often requires the instructor to mark the peer feedback provided.

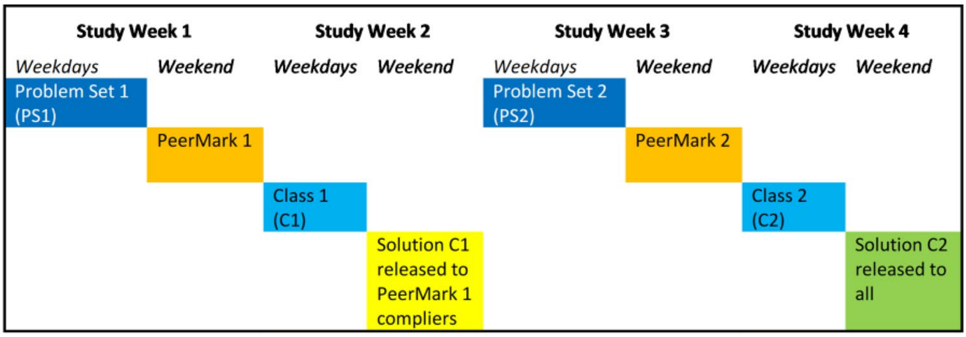

Figure 1. Study design.

Using the participation information with and without incentives, we analyse the quantitative data to estimate the effect of peer marking. In addition, we conducted two focus group discussions to undertake our process evaluation by identifying the key benefits and challenges of the process, which complements our quantitative analysis.

Data

We gather quantitative information from the respective course VLEs as well as other information on student achievement. Our key variables are:

- how many submissions each student made (between zero and two);
- whether they completed the peer marking exercises;
- final examination score for the relevant module (in %);
- previous level of achievement.

We use the final examination score of the relevant course as the outcome variable: we are interested in understanding whether the peer marking intervention impacts achievement at the end of the course.

For previous level of achievement, which we collect in order to attempt to control for level of ability, we use different methods for the undergraduate and postgraduate cohort. For the postgraduate cohort the students' undergraduate award mark was used, while for the undergraduate cohort the average Year 1 mark was used. For about 25% of the PG students, their previous score was reported as cumulative grade point average (CGPA) or the UK format of '2:1' or 'First', which is converted to % of marks obtained using the GPA pilot project conversion chart of the Higher Education Academy (2015). However, some information on the previous score is missing resulting in an effective sample of 385 students, with 291 PG students and 94 UG students.

Information on engagement with the intervention was collected as follows. We counted the number of submissions a student made and on which set of exercises, which we denote as the binary variable *submit1* if they submitted the first set and *submit2* if they submitted the second set. In addition, we recorded the number of peer marking activities completed – without considering which exercise set the peer marking belonged to – as *peermarked*, (value 0, 1 or 2), and also noted if they completed the peer marking for problem set 1 (*peermark1*) and the same for problem set 2 (*peermark2*). Note that if a student didn't submit they could not then take part in the peer marking exercise.

The summary statistics of all the variables considered are given in Table 1 and Figure 1. The submission rate was never 100% but reached as high as 76% for PG students with incentive.

Table 1. Summary statistics of the key variables.

Variable	Obs	Mean	Std. Dev.	Min	Max
final (%)	385	63.75	16.11	0	93
level (1 = PG)	385	0.76	0.43	0	1
pre_score (%)	385	73.93	11.57	3.50	92.75
peermarked	385	0.66	0.72	0	2
peermark1	385	0.49	0.50	0	1
peermark2	385	0.17	0.38	0	1
submit	385	1.17	0.82	0	2
submit1	385	0.68	0.47	0	1
submit2	385	0.49	0.50	0	1

The rate for UG students is less than 50%, which is reduced to a quarter without the incentive. There is a clear pattern of increased submission with the incentive irrespective of the level of study. Whether a UG or PG student, the peer mark rate is very low, within the 16–18% range without incentive.

Laying this submission and peer mark pattern over student preparation/motivation and outcome, Figure 3 presents the average final marks compared to the previous average scores. The graph in panel (a) shows the relationship over the number of problem sets peer marked, and panel (b) shows the relationship over the number of problem sets submitted. In both cases, there is an increasing pattern over the number of problem sets peer marked/submitted. The effect of both submission and peer marking seems to be more for UG students than the PG students, which is evident from the average differentials of the final marks and their standard errors. The increasing pattern of the previous score over submission and peer mark seems reasonable, too, reflecting that the more prepared/motivated students are more engaging with the process.

We organized two focus group discussions, one for each institution, to collect qualitative information from students about the intervention. To ensure ethical compliance and gather the best possible information, each focus group discussion was conducted by the author affiliated with the other institution. Each focus group discussion consisted of five students, who discussed the peer marking intervention at the end of the teaching term so that the experience was fresh in students' memories. The participants self-enrolled themselves for the focus group discussions which were advertised on the unit VLEs. The focus group discussions were conducted in-person on the university premises. The anonymous recordings were then transcribed using a university transcription service. We also used end-of-course anonymous student evaluations for the two courses as an additional source of qualitative data for the study.

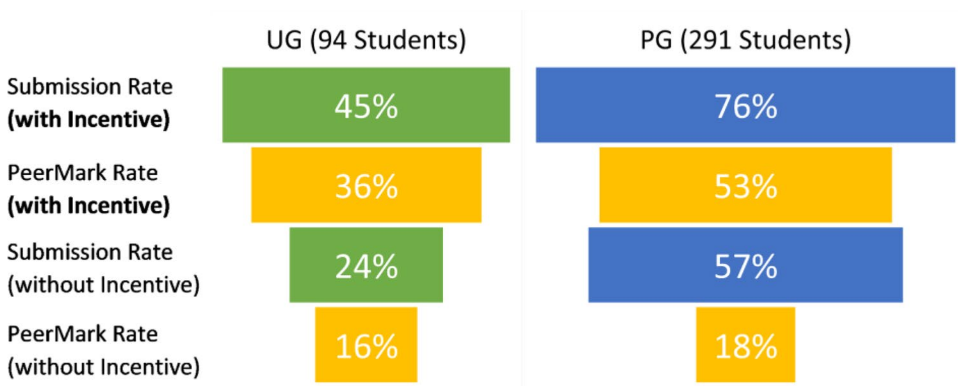

Figure 2. Summary statistics: rate of participation and peermark with and without incentive.

Methodology

It is never straightforward to estimate the effect of a pedagogical innovation on student learning due to our inability to design pure scientific experiments, for example using a randomized control trial, which would not be ethically compliant. An alternative is to fit a regression model, though it is impossible to collect all possible confounders of the relationship between final mark and engagement with the peer marking activity: whether a student engages in peer marking or not is most likely to be related to many other unobserved conditions, including but not limited to their preparation, ability, motivation, and various factors affecting their availability of sufficient time and space to study. If we simply built a regression model with final score as outcome and engagement with the peermarking intervention as a covariate, the resulting regression coefficient(s) would give estimates of association between final mark and peer marking engagement which combines the effect of all of these factors (and more).

We design our quasi-experiment to use instrumental variable regression to overcome the unobserved heterogeneity and lack of additional information on circumstances affecting student exam performances. This allows us to estimate the causal effect of the intervention on the final score. We use the incentive variable $peermark1_i$ as an instrumental variable for $peermarked_i$.

To be a valid instrumental variable, $peermark1_i$ needs to satisfy two conditions, namely, 'relevance' and 'exogeneity'. First of all, it needs to be related to our variable of interest, $peermarked_i$. This condition is satisfied as the number of PS peer marked depends on peermark1, and the dependence is higher than $peermark1_2$, as PS1 has the incentive and PS2 didn't. This is also evident from the student comments from the focus group discussions:

> the reason I wouldn't is because for problems like two the lecturer says the answer will be given to everyone anyway. (PG student)

> had to do in order to get something… the first time because it was the first one, we had this incentive of having the answers if we did it. (UG student)

To fulfil the exogeneity condition, the instrumental variable shouldn't be directly related to the outcome variable other than via the endogenous variable. In our case, this condition can be translated as the 'incentive', that is, the solution to C1 is not directly related to the final marks. This may sound contradictory at first, as having access to an additional set of solutions does positively affect the students' final marks. However, the quasi-experiment has been designed carefully to control for this, and the incentive was only the online release of C1 solution to those who did peermark1, but not to prohibit the access to C1 solution from other sources, e.g. solving C1 with the tutor in the class and collecting it from peers. This was acknowledged by the students in one of the focus group discussions:

> I know that even though I did not do it, I can get the answers like from my friends who have done that. (PG student)

Therefore, it is not very unrealistic to assume that incentive as the instrumental variable is not directly related to the final marks, thereby preserving exogeneity. It is not the availability of the additional solution that yields higher marks, but the practice by engaging in peer marking to enhance the understanding of the subject matter and the knowledge gathered on the overall marking scheme. The incentive here influenced the participation in the practice and thereby had an indirect positive effect on the final marks.

Our primary regression model with incentive (peermark1) as the instrumental variable is:

$$final_i = \beta_0 + \beta_1 \ peermarked_i + \beta_2 \ submit_i + \beta_3 \ prescore_i + \beta_4 \ level_i + u_i \quad (1a)$$

$$peermarked_i = \gamma_0 + \gamma_1 peermark1_i + \gamma_2 submit_i + \gamma_3 prescore_i + \gamma_4 level_i + v_i \quad (1b)$$

where, $final_i$ = % of final marks in the unit obtained by student i; $peermarked_i \in [0,1,2]$, the number of problem sets peermarked by student i; $submit_i \in [0,1,2]$, the number of problem sets submitted by student i; $prescore_i$ = % of marks obtained on average in the previous year by student i; $level_i = 1$ if PG, 0 if UG; $peermark1_i = 1$ if student i peer marked PS1, 0 otherwise. We also consider the two PS submission variables: $submit1_i$ and $submit2_i$ instead of $submit_i$ for comparison.

Equation (1b) is the first stage regression, regressing the possible endogenous variable on the instrumental variable and the same other control variables as in equation (1a). Equation (1a) is the structural equation with $peermarked_i$ being our variable of interest. We believe final marks could be influenced by 'submission' itself even if the student is not engaged in peer marking, including the variable. The previous achievements, measured by 'prescore', are important but insufficient determinants of students' innate ability. Finally, it is important to distinguish between the 'level' of study, PG or UG, which also works as the control for the institution and course instructor.

Findings

Overall analysis

The regression results obtained from simple ordinary least squares estimation of equation (1a) and the instrumental variable estimation are given in Table 2. Column (1) comes from an ordinary least squares analysis of the final score on peermarked, adjusted for submit, prescore and level. The results here cannot be treated as causal as we have not accounted for all (unobserved) confounders of the score-peermarked relationship. Meanwhile, column (2) comes from a two-stage least squares analysis where we use the instrumental variable in order to assess whether peermarked is causal for score. The estimated coefficients of $peermarked$ from both ordinary least squares and instrumental variable estimation methods are almost equivalent, yielding a 3.48% increase in final marks per problem set peer marked and are statistically significant at 1% and 5% level respectively. There is a further 2.85% increase in final marks per problem set submitted

Table 2. Estimation results with $peermark1$ as instrumental variable.

	(1)	(2)	(3)	(4)
VARIABLES	OLS	IV	OLS	IV
peermarked	3.481***	3.487**	3.387***	3.136*
	(1.203)	(1.610)	(1.219)	(1.602)
submit1			3.690*	3.875*
			(2.183)	(2.269)
submit2			2.227	2.336
			(1.714)	(1.761)
pre_score	0.405***	0.405***	0.402***	0.402***
	(0.0702)	(0.0697)	(0.0708)	(0.0699)
level	3.969**	3.970**	3.952**	3.901**
	(1.723)	(1.730)	(1.728)	(1.736)
submit	2.856**	2.852*		
	(1.402)	(1.486)		
Constant	25.16***	25.16***	25.18***	25.19***
	(4.897)	(4.867)	(4.891)	(4.843)
Observations	385	385	385	385
R-squared	0.239	0.239	0.239	0.239

Robust standard errors in parentheses.
*** $p < 0{:}01$,
** $p < 0{:}05$,
*$p < 0{:}1$. The sample consists of 94 UG students from University 1 and 291 PG students from University 2 in one academic year.

using both methods. As expected, a higher previous score increases final marks, representing higher student ability. The marks are higher for PG students compared to UG students.

We present two further columns in Table 2, columns (3) and (4), with estimations from a slightly changed model. This time, we include the two problem sets' submission data separately, including $submit1$ and $submit2$, instead of the combined variable $submit$. The instrumental variable estimate in column (4) is now slightly different from ordinary least squares in column (3), yielding a 3.13% increase in final marks per problem set peer marked. These estimates also reveal that the impact of submitting problem set one was higher than problem set 2. While the coefficient of $submit1$ is statistically significant, that of $submit2$ isn't.

Overall, our quantitative analysis reveals a positive and statistically significant effect of peer marking on final marks. The combined effect is not very high per problem set that is peer marked but still sufficient to be considered effective.

Heterogeneous effects analysis

It is also of interest whether the intervention benefits particular groups of students more than others, specifically whether the process is more beneficial for higher-achieving students or not. To do so, we subdivide our sample with the students gaining a 'first', that is 70% and over in their previous examination, and those who received a score under 70%. Table 3 reports the instrumental variable estimation results for these subgroups.

The results reveal that the peer marking process benefits UG students with a pre-score of less than 70% the most, with an estimated 14.79% increase in the final score per problem set peer marked for these students. Interestingly, when considering PG students, those with a pre-score of at least 70% gain the most from the intervention. The groups standing to gain the least are the PG students with a pre-score below 70% and UG students with a pre-score of at least 70%. However, there is no significant effect of submitting a problem set for any of the groups under this heterogeneous effect analysis.

Robust analysis

Though our quasi-experiment is carefully designed to capture the exogenous variation in peer-marking via the incentive, one may argue that there is endogeneity in choosing to peer mark having submitted. To address this issue, we modify our model in equations (1a) and (1b) to that in equations (2a) and (2b) where we include $submit1$ as another instrumental variable along with $peermark1$. We argue that the real source of exogenous variation comes from the

Table 3. Instrumental variable estimation results: heterogeneous effects.

	prescore > 69%		prescore < 70%	
VARIABLES	PG	UG	PG	UG
peermarked	4.562**	−3.890	2.510	14.79**
	(2.198)	(4.834)	(2.716)	(7.141)
submit	2.055	7.851	2.324	−8.049
	(2.099)	(4.944)	(2.340)	(5.623)
pre score	0.383	0.769**	0.159	0.590**
	(0.300)	(0.340)	(0.134)	(0.245)
Constant	30.46	1.986	47.42***	10.54
	(25.43)	(24.923)	(8.612)	(14.78)
Observations	173	43	118	51
R-squared	0.085	0.384	0.059	0.244

Robust standard errors in parentheses.
*** $p < 0.01$,
** $p < 0.05$,
* $p < 0.1$.

variation in submission across PS1 and PS2 and the variation in incentives. The incentive effect then only affects those students who have (exogenously) submitted the first problem set, so there is an incentive effect for that piece of work for those students, but no incentive effect for the students who did not (exogenously) submit PS1 nor an incentive effect for any student who submitted PS2.

$$final_i = \beta_0 + \beta_1 peermarked_i + \beta_2 submit2_i + \beta_3 prescore_i + \beta_4 level_i + u_i \quad (2a)$$

$$peermarked_i = \gamma_0 + \gamma_1 \text{peermark1}_i + \gamma_2 \text{submit1}_i + \gamma_3 \text{submit2}_i + \gamma_4 \text{prescore}_i + \gamma_5 \text{level}_i + \text{v}_i. \quad (2b)$$

The comparative results are presented in Table 4, for full sample as well as PG and UG sub-samples, as their submission rates are quite different (Figure 3). We note that our sample size is relatively small, particularly for the UG students. We explore iterated generalized method of moments (IGMM) estimator (given our overidentified setting with two instrumental variables in equation (2b) which is thought to have some gains in finite sample efficiency over simple instrumental variable estimates; Hall 2005; Hansen, Heaton and Yaron 1996) and present the results in Table 4.

Columns (1) to (4) in Table 4 presents the full sample estimates for instrumental variable and iterated generalized method of moments in two similar specifications, (1) and (2) including the variable $submit$ and (3) and (4) with $submit2$. The estimated effects of peer marking one additional script are quite similar (sometimes higher) to our estimates in Table 2; there is, on average, at least a 3% increase in final marks. The estimates for PG (columns 5 to 8) turn out to be slightly higher than the estimates for the UG level. All of these models satisfy Hansen's J-test of over-identifying restrictions.

Discussion

The above quantitative analysis, though limited in its scope given the small sample and the quasi-experimental structure, provides some evidence of the benefits of peer marking. For educators, the benefits are far more than that, though the challenges are no less.

Benefits of flipping classes with peer marking

It is a common complaint that it is often hard to engage students with flipped content on time (Akçayır and Akçayır 2018). In our setting, since flipping is in an assessment format, it is possible to motivate students with minimal incentives releasing additional solutions and marking scheme. To quote from one of our focus group discussions:

> the biggest incentive of doing peer review would be solutions after it.

The peer assessment process helps increase the amount of feedback students receive, without significantly increasing the marking load for the educator (Lynch and Schmid 2018). The process helps fulfil the common student demand for 'more practice questions' in quantitative units.

> I think it was really good, for example, tutorial, when we had the question to do it on our own it was like a mock small exam.

Flipping quantitative small group classes can increase student engagement during face-to-face classes, making them active and effective. Our analysis of the overall end of term student evaluation reveals an increase in both UG and PG levels, the average score from 3.98 to 4.50

Table 4. Robustness checks.

	Full Sample				PG				UG			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
VARIABLES	IV	IGMM	IV	IGMM	IV	IGMM	IV	IGMM	IV	IGMM	IV	IGMM
peermarked	3.284**	3.302**	4.752***	4.754***	3.425**	3.374**	4.461***	4.238***	2.438	2.399	4.018**	4.095**
	(1.579)	(1.578)	(1.299)	(1.299)	(1.680)	(1.677)	(1.533)	(1.523)	(4.208)	(4.208)	(2.033)	(2.021)
submit	2.970**	2.882*			2.311	2.194			2.319	2.358		
	(1.493)	(1.486)			(1.581)	(1.573)			(4.022)	(4.022)		
submit2			2.505	1.879			1.860	1.527			1.693	1.499
			(1.783)	(1.746)			(2.040)	(2.024)			(2.775)	(2.730)
pre_score	0.405***	0.409***	0.419***	0.436***	0.294***	0.299***	0.301***	0.308***	0.888***	0.886***	0.909***	0.902***
	(0.0695)	(0.0694)	(0.0712)	(0.0716)	(0.0723)	(0.0722)	(0.0739)	(0.0739)	(0.123)	(0.123)	(0.118)	(0.117)
level	3.929**	3.962**	4.640***	4.882***								
	(1.731)	(1.729)	(1.649)	(1.644)								
Constant	25.16***	24.97***	24.90***	23.89***	38.28***	38.20***	39.04***	39.03***	−6.840	−6.780	−7.896	−7.433
	(4.857)	(4.863)	(5.020)	(5.058)	(5.949)	(5.961)	(6.024)	(6.041)	(8.105)	(8.106)	(7.975)	(7.881)
Observations	385	385	385	385	291	291	291	291	94	94	94	94
R-squared	0.239	0.239	0.232	0.232	0.111	0.111	0.106	0.105	0.574	0.574	0.575	0.575
J-Test (p value)	0.563	0.563	0.091	0.091	0.492	0.492	0.177	0.177	0.752	0.752	0.697	0.697

Robust standard errors in parentheses.
*** $p < 0.01$,
** $p < 0.05$,
* $p < 0.1$. IGMM refers to iterated Generalized Method of Moments estimator. J-Test is the Hansen's J-test of over-identifying restrictions with the null hypothesis that the IVs are exogenous.

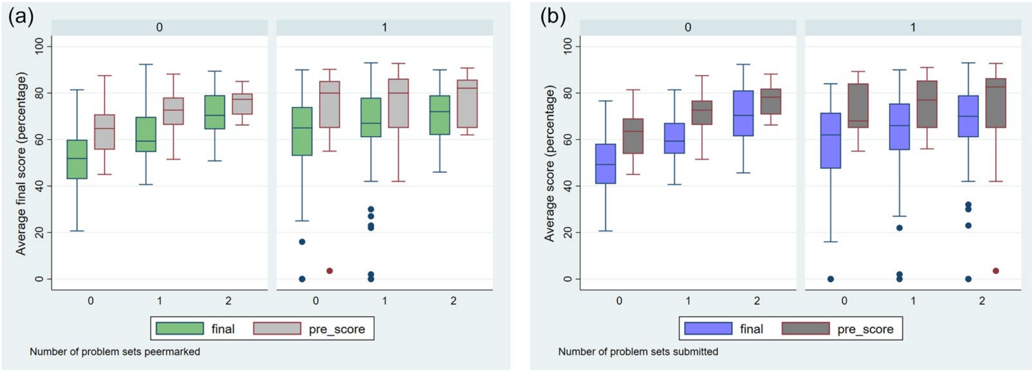

Figure 3. Summary statistics: average pre and final scores by number of problem sets submitted and peermarked in the UG and PG level. (a) Average score by number of problem sets peermarked (b) Average score by number of problem sets submitted

in PG, 4.56 to 4.68 in UG, both out of 5. The process enhances timely thinking and reflection, especially for those who generally don't see what feedback they received till late. This is also recognized by a student in the focus group discussion who commented as such:

> doing the marking and I'm looking at the solutions thinking about why an answer is right or wrong, I think it was and additional perk of doing the peer review.

Moreover, the process helps instant revision, improving student confidence as expected:

> Because after I submitted the homework, I can do the peer review so just a good time to help me do revision.

> during… doing the peer review and I find the student's answer is the same as me, I have more confidence.

Possible challenges and suggestions to achieving student buy-in to the system

Despite the benefits discussed, it is often a challenge to achieve student buy-in to the system. This is especially true if the assessment is 'for learning', i.e. a piece of formative assessment, rather than for credit (Cohen and Williams 2019). In our investigation, around half of the PG students participated in the peer marking exercise when an incentive was offered. For the UG students, the challenge is even higher: fewer than half of the UG students submitted the problem sets when an incentive was offered, with far fewer completing the subsequent peer marking. This may be linked with level of study: those registered on postgraduate courses have already proven their motivation to further their learning in their chosen subject.

Getting access to additional solutions seems to be an effective incentive, which is evident in our analysis and is supported by student comments during the focus group discussions. This is specifically so if the problem set felt hard to the students:

> if we did have the answers anyway then maybe I wouldn't have done it.

> for some hard questions… you just actually want to have a look at the right answers.

There is also the other side of the story; some students may struggle with peer marking for relatively harder questions:

> people have no idea what the question's about, how they can be even more helpful on marking someone else's script when they can't even understand the questions themselves.

> if the students… who I do the peer review with, approach in a different way to solve the problem, I may not figure out why he did it in that way, I would take it wrong.

While designing the problem sets, the educator needs to keep this in mind. A possible solution to this dilemma could be to provide additional resources, e.g. referring to specific textbook examples in the given solution guidelines to help those who might struggle to the point that they don't participate.

We use unit-specific rubrics or marking criteria while marking summative assessments, especially in large units with multiple markers. These detailed marking criteria are not released to the students. We have implemented similar marking criteria in peer marking tasks to help students mark their peer's script. These were implemented by adding a few 'tick-box' and 'free response' type questions in the peer mark system. Gaining an insight into the marking criteria through a peer marking exercise is also an incentive for students:

> I think it was a good opportunity for me to see how another student wrote the answers and also to see the marking scheme.

An important challenge is to entice students to continue to engage with peer-marking activities beyond the first exercise. Students quickly become demotivated if their allocated peer marker doesn't complete the marking or the quality of feedback is poor. This is evident from a student comment, 'Sometimes like you didn't receive the feedback and… it will discourage you'. Educators need to clearly convey the message that it is not a matter of 'being as capable as the lecturer' in order to assess others' work; it is about applying and practicing their capability and confidence through peer marking (Chew, Snee and Price 2016). The key is to make the students aware of the benefits of engagement, including that the process is likely to be more beneficial to the assessor than the assesses (Kim 2009). In addition, some technology, e.g. *peergrade.io,* seems to support conditioning receipt of peer review on own submission of peer review, which may mitigate this problem. Another possibility is to allocate more than one submission per student marker. This way each student is more likely to receive at least some quality feedback on their work. Indeed, there is also anecdotal evidence that asking the recipient of the feedback to grade or assess the quality of the feedback they received may also motivate students to provide thoughtful feedback.

In short, an incentive as small as releasing additional solutions helps increase student participation in formative peer assessment. Moreover, the educator needs to make sure that they convey a clear message to students that they aren't expected to provide complete and absolutely correct feedback and shouldn't expect to receive such feedback either. That is, the process needs to be more of engagement, practice and reflection rather than solely about the content of the feedback itself.

Setting up the online system of submission and peer marking, the release of the solutions (the incentive), and timely announcements from the course leader are often time-consuming tasks. However, we believe this is no more work than selecting or recording videos for flipped lectures or classes. Though under our suggested process the educator sets twice the number of problem sets, these can come from suggested textbook or past examination papers. Indeed, students may well be more motivated to engage with the process if the problem sets – and hence the solutions and marking scheme – is based on material that is connected with summative assessment.

Conclusion

In this paper, we propose a new way of flipping small group classes by active reading and peer marking. Our aim is to engage students in the learning material by attempting a problem and engaging in peer marking, for which they are given full solutions, before they are exposed to a similar problem to solve during small group classes. The solution to the in-class problem acts as an 'incentive', to be released only to those who took part in the peer marking exercise.

We designed a quasi-experiment with the view of evaluating the effect of peer marking by introducing this incentive in one such problem set and not in the other. Using 'incentives' as instrumental variables to participation, the effect of participating in one additional peer marking exercise yields a 3% increase in final marks on average. Heterogeneous effects analysis shows that the effect can be as high as a 14% increase in marks, on average, for relatively less able undergraduate participants.

While the quantitative impacts are impressive, the qualitative insights gained are no less. The process increases student engagement, satisfaction, confidence and overall responsibility for their learning. The challenge often lies in establishing a clear understanding of the purpose and the process of the peer marking to ensure student buy-in to the system.

Acknowledgments

This study benefited from a small teaching excellence fund from the University of Bristol, UK. We are grateful to the two anonymous referees of the journal, and Edmund Cannon for his useful comments on an earlier draft. We would like to thank the students who participated in our study for their time and feedback and the participants of the Development of Economics Education Conference 2019 at the University of Warwick.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributor

Rabeya Khatoon is a Senior Lecturer at the School of Economics, University of Bristol, UK. She is a senior fellow of Advance HE. Rabeya obtained her PhD in Econometrics from the University of Manchester.

Elinor Jones is an Associate Professor (Teaching) at the Department of Statistical Science, University College London, UK. Elinor graduated in Mathematics and Statistics from the University of Warwick before completing a PhD in Probability Theory at the University of Manchester.

ORCID

Rabeya Khatoon http://orcid.org/0000-0001-7966-9206

References

Akçayır, G., and M. Akçayır. 2018. "The Flipped Classroom: A Review of Its Advantages and Challenges." *Computers & Education* 126: 334–345. doi:10.1016/j.compedu.2018.07.021.

Albert, M., and B. J. Beatty. 2014. "Flipping the Classroom Applications to Curriculum Redesign for an Introduction to Management Course: Impact on Grades." *Journal of Education for Business* 89 (8): 419–424. doi:10.1080/08832323.2014.929559.

Alcalde, P., and J. Nagel. 2019. "Why Does Peer Instruction Improve Student Satisfaction More than Student Performance? A Randomized Experiment." *International Review of Economics Education* 30: 100149. doi:10.1016/j.iree.2018.10.001.

Ambrose, S. A., M. W. Bridges, M. DiPietro, M. C. Lovett, and M. K. Norman. 2010. *How Learning Works: Seven Research-Based Principles for Smart Teaching*. John Wiley & Sons, USA.

Becker, R., and S. Proud. 2018. "Flipping Quantitative Tutorials." *International Review of Economics Education* 29: 59–73. doi:10.1016/j.iree.2018.01.004.

Bergmann, J., and A. Sams. 2014. *Flipped Learning: Gateway to Student Engagement*. Arlington, VA: International Society for Technology in Education.
Calimeris, L., and K. M. Sauer. 2015. "Flipping out about the Flip: All Hype or is There Hope?" *International Review of Economics Education* 20: 13–28. doi:10.1016/j.iree.2015.08.001.
Chevalier, A., P. Dolton, and M. Lührmann. 2018. "Making It Count': Incentives, Student Effort and Performance." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181 (2): 323–349. doi:10.1111/rssa.12278.
Chew, E., H. Snee, and T. Price. 2016. "Enhancing International Postgraduates' Learning Experience with Online Peer Assessment and Feedback Innovation." *Innovations in Education and Teaching International* 53 (3): 247–259. doi:10.1080/14703297.2014.937729.
Cohen, A. J., and A. L. Williams. 2019. "Scalable, Scaffolded Writing Assignments with Online Peer Review in a Large Introductory Economics Course." *The Journal of Economic Education* 50 (4): 371–387. doi:10.1080/00220485.2019.1654951.
Deslauriers, L., E. Schelew, and C. Wieman. 2011. "Improved Learning in a Large-Enrollment Physics Class." *Science* 332 (6031): 862–864. doi:10.1126/science.1201783.
Donovan, C., D. N. Figlio, and M. Rush. 2006. "Cramming: The Effects of School Accountability on College-Bound Students." Technical Report. National Bureau of Economic Research.
Double, K. S., J. A. McGrane, and T. N. Hopfenbeck. 2020. "The Impact of Peer Assessment on Academic Performance: A Meta-Analysis of Control Group Studies."
Eager, E. A., J. Peirce, and P. Barlow. 2014. "Math Bio or Biomath? Flipping the Mathematical Biology Classroom." *Letters in Biomathematics* 1 (2): 139–155. doi:10.30707/LiB1.2Eager.
Falchikov, N., and J. Goldfinch. 2000. "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks." *Review of Educational Research* 70 (3): 287–322. doi:10.3102/00346543070003287.
Fu, Q. K., C. J. Lin, and G. J. Hwang. 2019. "Research Trends and Applications of Technology-Supported Peer Assessment: A Review of Selected Journal Publications from 2007 to 2016." *Journal of Computers in Education* 6 (2): 191–213. doi:10.1007/s40692-019-00131-x.
Gamage, D., M. E. Whiting, T. Rajapakshe, H. Thilakarathne, I. Perera, and S. Fernando. 2017. "Improving Assessment on MOOCs through Peer Identification and Aligned Incentives." In Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale, April, 315–318. doi:10.1145/3051457.3054013.
Gaynor, J. W. 2020. "Peer Review in the Classroom: Student Perceptions, Peer Feedback Quality and the Role of Assessment." *Assessment & Evaluation in Higher Education* 45 (5): 758–775. doi:10.1080/02602938.2019.1697424.
Gillanders, R., S. Karazi, and F. O'Riordan. 2020. "Loss Aversion as a Motivator for Engagement with Peer Assessment." *Innovations in Education and Teaching International* 57 (4): 424–433. doi:10.1080/14703297.2020.1726203.
Hall, A. R. 2005. *Generalized Method of Moments*. Oxford: Oxford University Press.
Hansen, L. P., J. Heaton, and A. Yaron. 1996. "Finite-Sample Properties of Some Alternative GMM Estimators." *Journal of Business & Economic Statistics* 14 (3): 262–280. doi:10.2307/1392442.
Higher Education Academy. 2015. "Grade Point Average: Report of the GPA Pilot Project 2013-14."
Huang, B., K. F. Hew, and C. K. Lo. 2019. "Investigating the Effects of Gamification-Enhanced Flipped Learning on Undergraduate Students' Behavioral and Cognitive Engagement." *Interactive Learning Environments* 27 (8): 1106–1126. doi:10.1080/10494820.2018.1495653.
Hughes, I. 2001. "But Isn't This What You're Paid for? The Pros and Cons of Peer and Self Assessment." *Planet* 3 (1): 20–23. doi:10.11120/plan.2001.00030020.
Hung, H. T. 2017. "Clickers in the Flipped Classroom: Bring Your Own Device (BYOD) to Promote Student Learning." *Interactive Learning Environments* 25 (8): 983–995. doi:10.1080/10494820.2016.1240090.
Kim, M. 2009. "The Impact of an Elaborated Assessee's Role in Peer Assessment." *Assessment & Evaluation in Higher Education* 34 (1): 105–114. doi:10.1080/02602930801955960.
Kobayashi, M. 2020. "Does Anonymity Matter? Examining Quality of Online Peer Assessment and Students' Attitudes." *Australasian Journal of Educational Technology* 36 (1): 98–110. doi:10.14742/ajet.4694.

Lai, C. L., and G. J. Hwang. 2016. "A Self-Regulated Flipped Classroom Approach to Improving Students' Learning Performance in a Mathematics Course." *Computers & Education* 100: 126–140. doi:10.1016/j.compedu.2016.05.006.

Lavecchia, A. M., H. Liu, and P. Oreopoulos. 2016. Behavioral economics of education: Progress and possibilities. In *Handbook of the Economics of Education*, edited by E. A. Hanushek, S. Machin, and L. Woessmann. Vol. 5, pp. 1-74. Elsevier. doi:10.1016/B978-0-444-63459-7.00001-4.

Li, L., X. Liu, and A. L. Steckelberg. 2010. "Assessor or Assessee: How Student Learning Improves by Giving and Receiving Peer Feedback." *British Journal of Educational Technology* 41 (3): 525–536. doi:10.1111/j.1467-8535.2009.00968.x.

Li, H., Y. Xiong, C. V. Hunter, X. Guo, and R. Tywoniw. 2020. "Does Peer Assessment Promote Student Learning? A Meta-Analysis." *Assessment & Evaluation in Higher Education* 45 (2): 193–211. doi:10.1080/02602938.2019.1620679.

Li, H., Y. Xiong, X. Zang, M. L. Kornhaber, Y. Lyu, K. S. Chung, and H. K. Suen. 2016. "Peer Assessment in the Digital Age: A Meta-Analysis Comparing Peer and Teacher Ratings." *Assessment & Evaluation in Higher Education* 41 (2): 245–264. doi:10.1080/02602938.2014.999746.

Lo, C. K., and K. F. Hew. 2017. "A Critical Review of Flipped Classroom Challenges in K-12 Education: Possible Solutions and Recommendations for Future Research." *Research and Practice in Technology Enhanced Learning* 12 (1): 4– 22. doi:10.1186/s41039-016-0044-2.

Lo, C. K., K. F. Hew, and G. Chen. 2017. "Toward a Set of Design Principles for Mathematics Flipped Classrooms: A Synthesis of Research in Mathematics Education." *Educational Research Review* 22: 50–73. doi:10.1016/j.edurev.2017.08.002.

Lundin, M., A. B. Rensfeldt, T. Hillman, A. Lantz-Andersson, and L. Peterson. 2018. "Higher Education Dominance and Siloed Knowledge: A Systematic Review of Flipped Classroom Research." *International Journal of Educational Technology in Higher Education* 15 (1): 1–30. doi:10.1186/s41239-018-0101-6.

Lynch, D., and B. Schmid. 2017. "Peer Evaluation: Enhancing Learning Opportunities and Reducing Marking Effort." In Proceedings of the Canadian Engineering Education Association (CEEA), University of Toronto. doi:10.24908/pceea.v0i0.10368.

Ng, V., and C. M. Fai. 2017. "Engaging Student Learning through Peer Assessments." In Proceedings of the 2017 International Conference on E-Education, E-Business and E-Technology, September, pp. 30–35. doi:10.1145/3141151.3141165.

Panadero, E. 2016. "Is It Safe? Social, Interpersonal, and Human Effects of Peer Assessment: A Review and Future Directions." In *Handbook of Human and Social Conditions in Assessment*, edited by G. T. L. Brown & L. R. Harris, 247–266. New York: Routledge.

Peters, O., H. Körndle, and S. Narciss. 2018. "Effects of a Formative Assessment Script on How Vocational Students Generate Formative Feedback to a Peer's or Their Own Performance." *European Journal of Psychology of Education* 33 (1): 117–143. doi:10.1007/s10212-017-0344-y.

Roach, T. 2014. "Student Perceptions toward Flipped Learning: New Methods to Increase Interaction and Active Learning in Economics." *International Review of Economics Education* 17: 74–84. doi:10.1016/j.iree.2014.08.003.

Rotsaert, T., E. Panadero, and T. Schellens. 2018. "Anonymity as an Instructional Scaffold in Peer Assessment: Its Effects on Peer Feedback Quality and Evolution in Students' Perceptions about Peer Assessment Skills." *European Journal of Psychology of Education* 33 (1): 75–99. doi:10.1007/s10212-017-0339-8.

Setren, E., K. Greenberg, O. Moore, and M. Yankovich. 2021. "Effects of Flipped Classroom Instruction: Evidence from a Randomized Trial." *Education Finance and Policy* 16 (3): 363–387. doi:10.1162/edfp_a_00314.

Snyder, C., L. M. Paska, and D. Besozzi. 2014. "Cast from the past: Using Screencasting in the Social Studies Classroom." *The Social Studies* 105 (6): 310–314. doi:10.1080/00377996.2014.951472.

Strelan, P., A. Osborn, and E. Palmer. 2020. "The Flipped Classroom: A Meta-Analysis of Effects on Student Performance across Disciplines and Education Levels." *Educational Research Review* 30: 100314. doi:10.1016/j.edurev.2020.100314.

Weaver, D., and A. Esposto. 2012. "Peer Assessment as a Method of Improving Student Engagement." *Assessment & Evaluation in Higher Education* 37 (7): 805–816. doi:10.1080/02602938.2011.576309.

Wen, M. L., and C. C. Tsai. 2008. "Online Peer Assessment in an in Service Science and Mathematics Teacher Education Course." *Teaching in Higher Education* 13 (1): 55–67. doi:10.1080/13562510701794050.

Yestrebsky, C. 2015. “Flipping a Large First-Year Chemistry Class: Same-Semester Comparison with a Traditionally Taught Large-Lecture Class.” In *Best Practices for Flipping the College Classroom* edited by J.B. Waldrop, M.A. Bowdon, 27–38. Routledge.

Zheng, L., X. Zhang, and P. Cui. 2020. “The Role of Technology-Facilitated Peer Assessment and Supporting Strategies: A Meta-Analysis.” *Assessment & Evaluation in Higher Education* 45 (3): 372–386. doi:10.1080/02602938.2019.1644603.