Student use of PeerWise: a multi-institutional, multi-disciplinary evaluation

Student use of PeerWise: a multi-institutional, multi-disciplinary evaluation

Alison E Kay,[1] Judy Hardy,[1,2] Ross K Galloway[1]

1. University of Edinburgh, U.K.

2. Corresponding Author. Address for correspondence:
School of Physics and Astronomy
The University of Edinburgh
James Clerk Maxwell Building
King's Buildings
Peter Guthrie Tait Road
Edinburgh
EH9 3FD
UK
email: j.hardy@ed.ac.uk
tel: +44 (0)131-650 6716
fax: +44 (0)131-650 6555

Alison Kay has recently completed her PhD in the Edinburgh Physics Education Research Group at the University of Edinburgh. Her research interests lie in STEM education and the development of research-based teaching methods to engage students in active, meaningful learning. Judy Hardy is Professor of Physics Education in the School of Physics & Astronomy at the University of Edinburgh. She has a particular research interest in evaluating educational strategies that improve student learning in physics and related disciplines. Ross Galloway is a Senior Teaching Development Officer at the University of Edinburgh. His research interests include the development of student problem solving skills, diagnostic testing, and flipped classroom pedagogies.

Abstract
This study explores the relationship between engagement with an online, free to use question-generation application (PeerWise) and student achievement. Using PeerWise, students can create and answer multiple-choice questions and can provide feedback to the question authors on question quality. This provides further scope for students to engage in discussion about the question with their peers. Data on PeerWise use and examination performance was collected from over 3000 students across six large undergraduate courses (in physics, chemistry and biology) over three academic years in three research intensive UK universities. A reliable and valid measure of overall PeerWise activity was created and a multilevel model developed describing the relationship between PeerWise activity and student performance in end of course examinations. Using this approach, a significant positive association was found between students' engagement with PeerWise and their academic attainment in end of course exams, even controlling for prior ability. The implications of these findings for educators are discussed.

Structured practitioner notes
What is already known about this topic

- Activities such as writing questions, answering questions and providing and using peer feedback are associated with enhanced understanding of course materials.
- Online applications such as PeerWise enable educators to implement these types of activities in large classes.
- Students who have a higher level of engagement with PeerWise have been shown to achieve higher levels of attainment in end of course exams than students with a lower level of engagement with PeerWise.

What this paper adds

- Multilevel modelling has been shown to be an effective tool for investigating the relationship between student engagement with PeerWise and their academic achievement.
- A reliable overall measure of PeerWise activity has been created that takes account of the quality of student comments and maintains the continuous nature of the data.
- There is a significant positive association between student engagement with PeerWise and performance across a range of courses, scientific disciplines, institutions and academic years
- The benefits are related to students' overall level of engagement with the system, rather than being critically dependent on specific activities

Implications for practice

- This study suggests that PeerWise can be effective for students of all abilities and level of preparedness.
- The overall level of activity is key, so students should be encouraged to utilise the different activities available in the system in a way that suits their own needs and understanding.

Introduction
Enabling students to become creative, critical problem-solvers is essential if they are to successfully face the demands of 21st century working life (Nicol, 2010). Problem solving; thinking critically; and synthesising information demands more than just a knowledge of fact, or surface level understanding. The social context of the learning process is extremely important – knowledge is constructed through shared interactions (Vygotsky, 1978). There is a growing body of literature attesting to the educational benefit of student engagement in active and collaborative learning activities (Prince, 2004), centering on the idea that the student needs to be the key player in their own learning experience and engage meaningfully with course materials. Furthermore, universal access to the internet and the development of sophisticated computing applications with the functionality of question bank creation has hugely enhanced the potential benefit to students of question writing and peer feedback exercises in recent years.

One tool that aims to develop higher-order skills is PeerWise (Denny, n.d.), an online application which enables students to create a bank of multiple-choice questions for their classmates to answer. In addition to writing questions and answering questions posed by their peers, students may also provide feedback on question quality to the author and engage in discussion about the subject matter (Kay, Hardy, & Galloway, 2018). PeerWise highlights the value of students' input "*beyond reading… and listening*" (Collis & de Boer, 2002), and towards the creation of resources to enhance the learning of the peer group as a whole (Hamer et al., 2008). By providing opportunities for students to make a tangible contribution to the course, they gain ownership of their learning and create materials emphasising what they view as valuable (F.-Y. Yu & Liu, 2005),whilst fostering deeper engagement with, and motivation for the learning process (Fellenz, 2004).

Purpose
The aim of this study was to investigate whether student engagement with course material through PeerWise activities promotes a deeper understanding of the course content. To test this, we used a two-fold approach. Firstly, we created a reliable, valid measure of PeerWise activity. We then developed a multilevel model to investigate whether engagement with PeerWise was associated with improved examination performance in courses across a range of disciplines and universities.

Theoretical framework
Student-generated questions and feedback

It has long been recognised that asking students to create questions enhances their engagement with, and understanding of, course materials (Draper, 2009; Rosenshine, Meister, & Chapman, 1996), resulting in a more solid retention of concepts (Chin & Brown, 2002; Lan & Lin, 2011). The process of writing multiple-choice questions – constructing question stems, working out correct answers and distractors, and writing explanations has potential not only to further deepen understanding, but also to develop skills of problem solving and information synthesis (F. Yu & Wu, 2012). Encouraging students to produce more complex, conceptual questions further deepens engagement and understanding. (Harper, Etkina, & Lin, 2003; Lan & Lin, 2011; Wilson, 2004). Furthermore, having an authentic audience to answer the questions posed, within a platform such as PeerWise, adds purpose the process, enhancing engagement (F.-Y. Yu & Chen, 2014).

While students have limited opportunities to pose many questions throughout their educational career, their role as question answerers is well established. Although the act of testing has often been regarded as "*neutral*" (Larsen, Butler, & Roediger, 2008), repeated testing, triggering the retrieval of information, has been demonstrated to be superior to engaging in additional studying of materials (Karpicke & Roediger, 2008). Applications such as PeerWise provide opportunities for frequent practice at answering questions. This allows students to maximise the direct benefit of practising information recall, as well as the indirect benefits of enhanced revision and reflection.

Providing feedback encourages the reviewer to engage critically with the subject matter in order to identify problem areas and provide advice or guidance as to possible solutions and improvements that could be made (Nelson & Schunn, 2008). Providing feedback to their peers also encourages students to spend more time on task and promotes a greater sense of accountability for not just their own learning, but also their peerss (Li, Liu, & Zhou, 2012). When students need to go beyond their immediate initial understanding in order to critically engage with another student's work, they may have to extend their knowledge to be able to articulate their point of view (Chi, Leeuw, Chiu, & Lavancher, 1994). After having given feedback, students are encouraged to reflect on, and improve their own performance, in light of their exposure to the standards set by their peers, and perhaps having developed a deeper understanding and internalization of assessment criteria (Cho & Cho, 2010; Li, Liu, & Steckelberg, 2010; Liu & Carless, 2006; Lu & Law, 2011). Indeed, the higher the quality of feedback provided by the reviewer, the better the reviewer's subsequent performance. (Li et al., 2010).

Students can use Peerwise to write and receive feedback comments as well as to author and answer questions. Interestingly, while each of these individual activities has the potential to enhance academic achievement, studies have shown that it is through engaging *across* a range of PeerWise activities that students can best improve their understanding and academic performance (McQueen et. al , 2014). Previous studies of students overall use of PeerWise (see for example Casey et al., 2014; Denny, Hamer, Luxton-Reilly, & Purchase, 2008; Hardy et al., 2014) have used a "Multiple Measure" (MM); a combined measure of activities that can be said to represent the aggregate of the "work done" by a student using the system. However, this measure is largely based on the quantity of student contributions (number of questions written and answered, number of days active, length of comments), which does not necessarily reflect the quality.

Limitations of previous analyses

Most previous research has tended to focus on a single course, in a single year, situated within one institution (Galloway & Burns, 2014; Ryan, 2013, Denny, & Nicolson, 2011). Even those studies that have investigated PeerWise use across multiple courses (Casey et al., 2014, Hardy et al., 2014) have analysed results at the individual course level, rather than treating the data as a whole, thus limiting the statistical inferences that can be drawn when making comparisons between courses. This points towards the need for more robust and appropriate methods that allows for statistical comparison of results across courses, years and institutions by enabling data to be combined.

In addition, attempts to examine or account for the effects of prior attainment have been made by splitting students into attainment quartiles (Denny et al., 2008, Hardy et al., 2014), but this approach has a number of limitations. In particular, where the split is placed will affect the make-up of the groupings and thus the variation within and between groups. Although having a larger number of groupings allows students' differences to be modelled at a higher level of granularity than, say, a division at the median, it is sub-optimal to categorise a continuous variable as this can result in a loss of statistical power (Cohen, 1983). Dichotomization of the data also introduces problems. For example, if a class of 200 students is divided into quartiles and each quartile split at the median into high and low activity, each group will only comprise 25 students which is a relatively small sample size to carry out statistical testing. Additionally, there are inevitably many tied ranks, especially for the number of questions authored and the number of days active, where most students author the minimum number of questions, and where there is a limited range of days that students are active. This problem is compounded when using a combined measure of activity.

Multilevel analysis has the potential to provide a more robust statistical analysis than has been performed previously. Multilevel models are linear regression models but where differences between groups - in this case, courses - can be modelled at several levels (Tabachnick & Fidell, 2013). Thus multilevel modelling allows for

statistical comparison of the association between PeerWise use and student achievement across courses by enabling data to be combined. In the current work, statistical comparisons across courses have been undertaken on aggregated data by accounting for the nested nature of the data.

Courses and course context

This study included six large undergraduate courses in three research intensive UK universities: Edinburgh and Glasgow (Scottish), and Nottingham (English), spanning three scientific disciplines; physics, chemistry and biology. Class sizes ranged from 90 to 279 students. The study covered academic years 2011-12, 2012-13 and 2013-14.

Instructors were motivated to use PeerWise by a desire to promote deeper engagement with course materials and the development of higher-order skills. In acknowledgement of the challenging nature of the task, students in all courses were given considerable guidance in writing effective multiple-choice questions through scaffolding workshops, and had opportunities to practise writing, answering and commenting on questions before the PeerWise system went 'live'. This scaffolding did not focus on the user interface or mechanics of the PeerWise system, but rather addressed the pedagogic considerations needed for good quality multiple-choice questions. For example, the scaffolding workshops featured student activities highlighting effective and ineffective aspects of questions (e.g. poor-quality distractors, inadvertent cues, etc.) plus group activities on question design to familiarise students with the process. The incorporation of PeerWise was part of wider curriculum changes to promote active learning, for example, the adoption of electronic voting using 'clickers'; Peer Instruction episodes; and the flipped classroom approach.

The PeerWise assessment was worth between 1% and 6% of the course mark. Each course stipulated a minimum level of engagement. (Table S1). The PeerWise assignment replaced a 'traditional' hand-in exercise, thus ensuring that students did not have an increased assessment load. Moreover, teaching staff did not monitor the system to ensure accuracy – the platform was regulated by students themselves, both in relation to academic quality and non-academic disputes. A lack of staff regulation may raise concerns about the overall quality of the repository, however research indicates that most students (over 90%) write accurate, clear questions, and in most instances where errors occur, they are identified and corrected by other students on the course (Bates, Galloway, Riise, & Homer, 2014; Bottomley & Denny, 2011; Denny, Luxton-Reilly, & Simon, 2009). Moreover, erroneous questions tend to be given a much lower quality rating by students (Denny, Luxton-Reilly, & Simon, 2009) and a significant positive correlation between quality rating and the number of answers to questions has also been reported (Denny, Hanks, & Simon, 2010). Quality ratings therefore have the effect of promoting accuracy and provide a rough filter for students to determine the usefulness of a particular question (Denny, Luxton-Reilly, & Simon, 2009). Aside from the small differences in the marking of the PeerWise assignment, course organisers confirmed that course structures remained the same across the study period – suggesting that within each course, performance would be broadly comparable across the three academic years.

The number of students included in the analysis, and the percentage of the total number of enrolled students this represents is shown in Table S2. Students were only included in the analysis if they had a measure of exam score, prior ability and the Multiple Measure. Therefore, not all students enrolled in the class have been included in the analysis.

Method

Variable construction

Three variables were constructed: exam score, prior ability, and the multiple measure of engagement – a derived variable comprising levels of engagement with each of the activities undertaken on PeerWise.

Exam score: The dependent variable of interest throughout this work is student attainment, as measured by performance in end of course exams. To assist with meaningful interpretation of the results, the final exam scores for each course remain unstandardized, thus ranging from 0-98. The overall mean exam score is 61.331 with a standard deviation of 16.207.

Prior Ability: Past academic performance has regularly been demonstrated to be a strong predictor of future performance. (Hattie, 2008). To account for the fact that stronger students will generally perform better in exams than weaker students and that they may also engage with PeerWise to a greater level, prior ability was statistically controlled for, to isolate its effect on exam score.

Course organisers were asked to identify an assessment, undertaken prior to the introduction of PeerWise as a measure of a student's ability. In Physics 1A, the Force Concept Inventory (FCI) (Hestenes, Wells, & Swackhamer, 1992) – measuring student understanding of the Newtonian concept of force – was chosen. The FCI was undertaken prior to any teaching. Students enrolled on Physics 1B have (mostly) completed Physics 1A, therefore Physics 1A exam score was used as a proxy for ability. Similarly, Chemistry 1A is a prerequisite for Chemistry 1B). In both Physics 2 (Glasgow) and Foundations of Chemistry (Nottingham), compulsory class tests were chosen. Genes and Gene Action (GGA), a second-year course used exam performance on Molecules, Genes and Cells (MGC), a prerequisite first year course. These assessments were chosen on the basis that they

were designed to assess not only knowledge and understanding, but also aspects of problem solving and critical thinking. For example, the Physics courses under consideration conform to the expectations of Institute of Physics accreditation, which mandates some element of problem solving in potentially unrehearsed contexts. Similarly, the Chemistry courses are part of a Royal Society of Chemistry accredited programme, where assessment is required to rigorously test breadth of knowledge and its application to problems at a threshold level. As such, these assessments were considered reliable indicators of ability.

Pre-scores were standardized within each course and centred at the mean before their inclusion in the analysis since it cannot be assumed that a student that scores 60% in one course has the same ability as a student scoring 60% on a different course. As the mean for each of the courses is zero, they are therefore also centred at the grand mean – the overall mean of all the courses when aggregated.

Multiple Measure: The aggregate, or multiple measure (MM), of PeerWise activity used in the current study comprises a summed score based upon the standardized values of four activities undertaken using PeerWise - the number of questions written and answered and the number of *quality* comments given and received (where the feedback should be used to enhance understanding, thus improving future performance). The measures based quality comments distinguish between students who contributed posts filled with emojis or irrelevant text from those who wrote longer more detailed comments, and considers the benefit to students of *receiving* feedback from their peers, and differ from those used in previous studies (Casey et al., 2014; Denny, Hamer, Luxton-Reilly, & Purchase, 2008; Hardy et al., 2014), which used the number of days active on the system and comment length (along with the number of questions written and answered). Submissions from the start of the course until the date of the first sitting of the final exam were included, however duplicated or deleted questions were omitted, as were submissions from after the final exam until the resit.

All comments submitted to the system were coded, (n = 80262) according to a three-point scheme. Table S3 outlines the types of comment falling under each code within each course. The broad categories of the three-point scale increase reliability, resulting in more robust, error-free analysis of each student's contribution. Details of reliability testing carried out on the coding is included in the supplementary information. It should be noted that quality of submission was not assessed in the PeerWise assignment. Comments at level 1 were irrelevant or nonsensical – submitted to increase a student's participation score (n = 17106 – 22%). Comments at level 2 were relevant to the question but were more simplistic in nature, for example a general "*good question*" or "*thank you*" type comment. (n = 32438 – 40%) 78% of all comments written therefore provided relevant feedback to the question author, or contributed to wider discussion about the question or concepts it addressed. Although level 2 comments demonstrate engagement with the exercise and contribute to the development of the online community, only comments coded at level 3 were included in the analysis (n = 30398 – 38%), as a focus of this work was to move towards creating a measure of quality, These comments were more sophisticated, perhaps commenting on a specific aspect of the question, or identifying an error, or an alternative solution. Given the high number of comments, and that the platform was an unmoderated space for student interaction, instructors were very encouraged by the quality of the engagement of students.

Within each course, the number of submissions for each of the four measures – authoring questions; answering questions; giving quality comments; and receiving quality comments – were standardized. These scores were then summed and an average taken, to create the MM. The reliability and unidimensionality of the measure was then confirmed ($\alpha > 0.7$), indicating measurement of the same underlying construct – engagement with the PeerWise system – and hence supporting the construct validity of the MM. It should be noted that for the purposes of this analysis, dimension reduction was undertaken purely to describe the structure of the data, thus the score on the measure is simply the mean score of the standardized activity levels.

Multilevel Model

In this study, a random intercept model was used. Random intercept models allow the mean value of the dependent variable of each group (course) to vary from the overall mean value, whilst assuming that the relationships or slopes between the independent and dependent variables remain constant across courses. Tis means that random intercept models are suitable for investigating whether PeerWise attainment and prior ability can explain any differences across courses in average student exam attainment, and whether any effects of PeerWise engagement remain significant when accounting for prior ability.

Initial testing was undertaken to confirm that course level differences in exam score did indeed exist. First, a null model was constructed with no predictor variables to determine: a) the amount of variation in exam score that can be attributed to course differences and b) whether the variation between courses is statistically significant and therefore worth modelling further. The overall mean exam score across all courses was 61.33. The intraclass correlation coefficient was calculated to be 0.0986 – indicating that around 10% of the total variation in exam score is attributable to course differences. This is the first indication that there are course-level differences worth further modelling (Robson & Pevalin, 2016). As a second method of determining whether there are differences between courses in exam score, the model fit of the null model (allowing mean exam score to vary across courses) was compared to the single-level model where course means were not allowed to vary.

The deviance statistic (–2LL) was calculated to assess the fit of the model. The higher the –2LL, the less optimal the model fit. The difference of 229.13 with 1 *d.f.* is highly significant, with a p value of < .001 thus demonstrating that there are significant course-level effects on exam score that are worth further examination
Changes in the –2LL can be tested to ascertain whether the addition of fixed or random effects significantly improves the fit of the model. If the model fit is significantly improved, then the newly added parameter explains a significant proportion of the variance and should therefore be left in the model. The –2LL is often taken as be the most accurate determinant of model fit, and thus it is suggested that, just as with effect sizes, the magnitude of the variance coefficients are considered, rather than their significance (Robson & Pevalin, 2016). This approach is adopted in the current work where the primary purpose of using multilevel modelling is to account for the nested structure of the data and to focus on the fixed effects in the model – the relationship between PeerWise activity and exam score (Kreft & De Leeuw, 1998).

The assumptions for multilevel modelling are broadly similar to those of linear regression. As with single level linear regression models, these assumptions were tested in the current work by producing normality plots and scatter plots at each level of analysis. This was undertaken for the "best model" in the analysis. The plots show some deviation from normality, however deletion of the most outlying cases at the student-level did not materially change the results of the analysis. All students were therefore included in the final analysis.

Although using multilevel models is preferable to simply aggregating course data without regard for the hierarchical, grouped data-structure, the small sample size at level 2 must be considered as a potential limitation. Whilst it has been suggested that at a minimum, there must be at least 10 level 2 units to conduct the analysis, a value of 20 is thought to be appropriate, and 30 has been cited as the smallest number to ensure that standard error and estimate display as little bias as possible. Caution should therefore be exercised in interpreting the random coefficients, however the fixed part of the model will be unaffected (Maas & Hox, 2005). Given the relatively small sample size of courses in this study, it is worth firstly interpreting the random intercept models, then using any significant effects in the random slope models as indicative of relationships worthy of future study. Preliminary analysis of the random effects may be a starting point for future analyses which may incorporate course-level variables to explain any emerging differences between courses, in the relationship between PeerWise activity and attainment.

Results

As described above, there exist course-level differences worth further modelling. Table 1 outlines the development of a model to determine the relationship between the multiple measure and exam score. Four models (Models 0-3) were developed; these are defined in Table 1. Model 0 shows that there is evidence of variation in exam score amongst courses. On adding the multiple measure and allowing the intercept to vary to create Model 1, a positive relationship between the multiple measure and exam score emerges. The –2LL is significantly reduced, indicating that model fit has been improved (χ^2 = 226.91, 1 *d.f.*, p < .001). This relationship persists with the addition of prior ability in model 2 (χ^2 = 190.67, 1 *d.f.*, p < .001). Although the multiple measure and prior ability both have a positive fixed relationship with exam score, only by allowing the regression slopes to vary according to course can it be determined whether the strength of relationship is in fact consistent across all courses. Given the limited number of units at level 2, extreme caution must be exercised when interpreting level 2 random effects. Model 3 demonstrates that whilst there is a significant fixed relationship between the multiple measure and exam score when controlling for prior ability, the relationship between the multiple measure and exam score also varies across courses (χ^2 = 828.32, 2 *d.f.*, p < .001), however this should be interpreted tentatively.

Across the 18 courses in question, there exists a significant positive relationship between overall engagement with PeerWise activity and performance in end of course exams – the greater the level of activity on the system, the higher a student's exam score. Ability has been included in the model to remove its effects from the relationship between exam score and PeerWise activity. This research demonstrates that the relationship between exam score and PeerWise activity persists, even when controlling for a students' prior ability.

Discussion

This is one of the most comprehensive studies of PeerWise undertaken to date, comprising over 3000 students in six courses, across three disciplines, in three universities, across three academic years. It is also the first study to report the use of multilevel modelling to investigate the benefits of engaging with PeerWise. The findings provide robust evidence to support the use of PeerWise as an effective tool for enhancing student learning and achievement. They also point to the flexible nature of the system; the positive association between PeerWise activity and student performance is not strongly dependent on the course, the material under study, the instructor or the implementation details. Furthermore, the benefits are related to the overall level of engagement, rather than being critically dependent on the specific activities that students undertake. This flexibility can perhaps be explained by the fact that different PeerWise activities target different aspects of the learning process. There is built-in differentiation within tasks – students can participate at a level that is most appropriate for their level of performance and understanding, for example, by writing questions that are complex, or choosing to answer more or less challenging questions. Some students may lack confidence to critique their peers' contributions –

choosing to focus on the question authoring or answering aspects of the course. This flexibility of the system makes engagement with PeerWise beneficial to students of all abilities. It also ensures that opportunities to engage with tasks requiring higher-order skills such as evaluation are not limited to students of higher abilities – all students can engage with tasks at different levels at different times to best develop their weaknesses and stretch their understanding.

Students with a strong understanding of the material may find themselves able to write challenging questions, synthesising a range of topics; or may further their knowledge by providing an explanation for a fellow student who does not understand a concept; or develop a more elegant solution to a question. Students who are struggling to understand a topic or who need a stronger foundation of knowledge may find it most useful to practice answering questions to aid retention of concepts, to test their knowledge and to provide feedback on their performance. Moreover, students with a strong understanding in one curricular area might find they need more support in another. One of the clear strengths of the PeerWise system is that it allows students to self-sort their mode of engagement in this way without any additional input from the course instructor.

Occasionally there may be issues around suitability of posts, the accuracy or quality of questions, or the quality of the comments. It is inevitable that some students will participate at minimal levels, however it is important not to assume that all students who engage at times at a more superficial level do not also make insightful contributions. The community aspect of PeerWise is crucial to foster a safe space for learning, therefore it is arguably desirable that students behave in a similar manner online as they do offline, interacting on a social level as well as discussing their work. It could be argued that this indicates that students are using the system authentically, taking ownership of their space, and thus producing deeper engagement with the task. There is currently functionality for students to report offensive or problematic questions, however, that the space is student-owned is critical to the mindset in which students should be operating when using PeerWise. Given the small proportion of the overall course mark assigned to PeerWise, unchecked errors have minimal consequences on student marks. If the proportion of marks assigned to PeerWise were to increase, there may be a need to moderate submissions more systematically, however, it is the view of the authors that any more heavy-handed staff contributions may risk damaging the co-operative, student-led dynamic of the system.

It is worth noting that in all of the courses in this study, PeerWise only contributed a small (typically <5%) percentage of the overall course mark. It is also low cost in terms of staff effort, requiring little time to set up and maintain. Thus PeerWise potentially offers educators a ‘low-cost low-risk’ intervention that supports student learning across the class. More generally, it points to the educational benefits of providing students with a space where they can participate in the creation and discussion of course material in a structured but student-led way.

Limitations

The use of exam score as a measure of either attainment, or prior ability may have some limitations. Practising answering multiple-choice questions may not produce directly transferable benefits where there is no multiple-choice component in the final exam. Moreover, PeerWise encourages the development of higher-order skills such as reasoning and problem solving. These skills take time to develop and require a change in a student’s approach to learning, which may perhaps be better assessed after a longer period. Some may also argue whether exams can be considered the best assessment of higher cognitive skills such as problem solving. Although exams may risk merely testing recall, in the courses included in this study, higher order skills were explicitly targeted, for example, Physics 1A and 1B comprised open book exams, with the aim of encouraging students to apply their knowledge. In all other courses students were taught in an interactive manner by staff committed to incorporating an evidence based pedagogy into their practice. There are also potential limitations of the MM. The activities comprising the MM have been given equal weighting. This assumes that each activity is of the same value in terms of both its learning benefits to, and the level of effort required from students, in its completion. The different balance of the activities in each course makes it difficult to place a meaningful weighting on each activity. Moreover, it is not clear how to judge, for example, the number of questions answered that would be equivalent to writing one question. Although giving each activity the same weighting may not accurately reflect the real balance in terms of the benefits of PeerWise, this compromise nevertheless ensures comparability across courses.

Table 1: Course effects on the relationship between MM and exam score

	Model 0		Model 1		Model 2		Model 3	
	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
Fixed Effects								
Intercept	60.84*	1.24	60.84*	1.24	60.84*	1.24	60.84*	1.24
z MM[g] estimate			1.04	0.08	0.91*	0.07	1.09	0.13
z Pre-score[h] estimate					1.39*	0.09	1.42*	0.09
Random Effects								
Course Level Variance								
Intercept variance	26.22	9.25	26.36	9.25	26.48	9.25	26.50	9.27
Covariance: Course and MM							0.04	0.70
Slope variance: MM							0.21	0.11
Student Level Variance								
Student variance	239.61	6.13	225.73	5.78	208.00	5.32	204.688	2.25
Deviance (–2*log likelihood)	25593.30		25410.07		25159.41		25129.90	
N: course	18		18		18		18	
N: students	3071		3071		3071		3071	

* Coefficient is approximately twice its standard error.

Acknowledgements
AK thanks the Higher Education Academy for a PhD studentship through the Mike Baker Doctoral Programme. The authors would like to thank Dr Heather McQueen, Dr Peter Kirsop, Dr Morag Casey and Dr Kyle Galloway for giving access to their students' data and for numerous discussions about how PeerWise has been implemented within their courses. Thanks must also be extended to Dr Paul Denny for providing PeerWise analytics for each of the courses analysed. Finally, thank you to all the students whose data has been analysed during this work.

Open Data
The data used in this research contains summative assessment marks. We are therefore unable to make this publicly available.

Ethics
This research followed the UKRIO's Code of Practice for Research. It took place under the auspices of each participating institutions' Ethics Committee.

Conflict of Interest
There is no conflict of interest in the work reported above.

References

Barak, M., & Rafaeli, S. (2004). On-line question-posing and peer-assessment as means for web-based knowledge sharing in learning. *International Journal of Human-Computer Studies*, *61*(1), 84–103. http://doi.org/10.1016/j.ijhcs.2003.12.005

Bates, S. P., Galloway, R. K., Riise, J., & Homer, D. (2014). Assessing the quality of a student-generated question repository. *Physical Review Special Topics - Physics Education Research*, *10*(2), 020105. http://doi.org/10.1103/PhysRevSTPER.10.020105

Berry, J. W., & Chew, S. L. (2008). Improving Learning through Interventions of Student-Generated Questions and Concept Maps. *Teaching of Psychology*, *35*(4), 305–312. http://doi.org/10.1080/00986280802373841

Bottomley, S., & Denny, P. (2011). A participatory learning approach to biochemistry using student authored and evaluated multiple-choice questions. *Biochemistry and Molecular Biology Education*, *39*(5), 352–361. http://doi.org/10.1002/bmb.20526

Casey, M. M., Bates, S. P., Galloway, K. W., Galloway, R. K., Hardy, J. A., Kay, A. E., … McQueen, H. A. (2014). Scaffolding student engagement via online peer learning. *European Journal of Physics*, *35*(4). http://doi.org/10.1088/0143-0807/35/4/045002

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. In A. Kozulin (Ed.), *Vygotsky's educational theory and practice in cultural context* (pp. 39–64). Cambridge University Press.

Chi, M. T. H., Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting Self-Explanations Improves Understanding. *Cognitive Science*, *18*(3), 439–477. http://doi.org/10.1207/s15516709cog1803_3

Chin, C., & Brown, D. E. (2002). Student-generated questions: A meaningful aspect of learning in science. *International Journal of Science Education*, *24*(5), 521–549. http://doi.org/10.1080/09500690110095249

Cho, Y. H., & Cho, K. (2010). Peer reviewers learn from giving comments. *Instructional Science*, *39*(5), 629–643. http://doi.org/10.1007/s11251-010-9146-1

Cohen, J. (1983). The Cost of Dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253.

Collis, B., & de Boer, W. (2002). A changing pedagogy in e-learning: From acquisition to contribution. *Journal of Computing in Higher Education*, *13*(2), 87–101.

Collis, B., & Moonen, J. (2006). The contributing student: Learners as co-developers of learning resources for reuse in web. In M. Khine & D. Hung (Eds.), *Engaged Learning with Emerging Technologies* (pp. 49–67). New York, NY: Springer.

Denny, P. (n.d.). PeerWise. Online learning tool.

Denny, P., Hamer, J., Luxton-Reilly, A., & Purchase, H. (2008). PeerWise: students sharing their multiple choice questions. In *Proceedings of the Fourth international Workshop on Computing Education Research* (pp. 51–58). New York, NY, USA: ACM. http://doi.org/10.1145/1404520.1404526

Denny, P., Hanks, B., & Simon, B. (2010). PeerWise: Replication study of a student-collaborative self-testing web service in a U.S. setting. In *SIGCSE* (pp. 421–453).

Denny, P., Luxton-Reilly, A., Hamer, J., & Purchase, H. (2009). Coverage of course topics in a student generated MCQ repository. In *ItiCSE* (Vol. 41, pp. 11–15). Paris, France: ACM. http://doi.org/10.1145/1595496.1562888

Denny, P., Luxton-Reilly, A., & Simon, B. (2009). Quality of Student Contributed Questions Using PeerWise. In *Proceedings of the Eleventh Australasian Conference on Computing Education - Volume 95* (pp. 55–63). Darlinghurst, Australia, Australia: Australian Computer Society, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1862712.1862724

Denny, P., Simon, B., & Micou, M. (2010). Evaluation of PeerWise as an educational tool for bioengineers. In *ASEE Annual Conference and Exposition*. Louisville, Kentucky.

Draper, S. W. (2009). Catalytic assessment: understanding how MCQs and EVS can foster deep learning. *British Journal of Educational Technology*, *40*(2), 285–293. http://doi.org/10.1111/j.1467-8535.2008.00920.x

Falkner, K., & Falkner, N. J. G. (2012). Supporting and structuring "contributing student pedagogy ” in computer science curricula. *Computer Science Education*, *4*(December 2012), 413–443.

Fellenz, M. R. (2004). Using assessment to support higher level learning: the multiple choice item development assignment. *Assessment & Evaluation in Higher Education*, *29*(6), 703–719. http://doi.org/10.1080/0260293042000227245

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8410–8415. http://doi.org/10.1073/pnas.1319030111

Galloway, K. W., & Burns, S. (2014). Doing it for themselves: students creating a high quality peer-learning environment. *Chem. Educ. Res. Pract.* http://doi.org/10.1039/C4RP00209A

Goldstein, H. (2011). *Multilevel Statistical Models* (4th ed.). Chichester: John Wiley and Sons.

Hamer, J. (2006). Some experiences with the “contributing student approach.” In *ItiCSE* (pp. 68–72). Bologna, Italy: ACM.

Hamer, J., Cutts, Q., Jackova, J., Luxton-Reilly, A., Mccartney, R., Purchase, H., … Sanders, K. (2008). Contributing student pedagogy. *Inroads - SIGCSE Bulletin*, *40*(4), 194–212.

Hardy, J., Bates, S. P., Casey, M. M., Galloway, K. W., Galloway, R. K., Kay, A. E., … McQueen, H. A. (2014). Student-Generated Content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education*, (May), 1–15. http://doi.org/10.1080/09500693.2014.916831

Harper, K. a., Etkina, E., & Lin, Y. (2003). Encouraging and analyzing student questions in a large physics course: Meaningful patterns for instructors. *Journal of Research in Science Teaching*, *40*(8), 776–791. http://doi.org/10.1002/tea.10111

Hattie, J. (2008). *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. Routledge.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158. http://doi.org/10.1119/1.2343497

Hooper, A. S. C., Park, S. J., & Gerondis, G. (2011). Promoting student participation and collaborative learning in a large Info 101 class: Student perceptions of PeerWise web 2.0 technology. In *Proceedings of Higher Education Research and Development Society of Australasia International Conference*.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science (New York, N.Y.)*, *319*(5865), 966–8. http://doi.org/10.1126/science.1152408

Kay, A. E., Hardy, J., & Galloway, R. K. (2018). Learning from peer feedback on student-generated multiple choice questions: Views of introductory physics students. *Physical Review Physics Education Research*, *14*(1), 10119. http://doi.org/10.1103/PhysRevPhysEducRes.14.010119

Kreft, I., & De Leeuw, J. (1998). *Introducing Multilevel Modeling*. London: Sage Publications Ltd.

Lan, Y., & Lin, P. (2011). Evaluation and improvement of students question-posing ability in a web-based learning environment. *Australasian Journal of Educational Technology*, *27*(4), 581–599.

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2008). Test-enhanced learning in medical education. *Medical Education*, *42*(10), 959–966. http://doi.org/10.1111/j.1365-2923.2008.03124.x

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology*, *41*(3), 525–536. http://doi.org/10.1111/j.1467-8535.2009.00968.x

Li, L., Liu, X., & Zhou, Y. (2012). Give and take: A re-analysis of assessor and assessee's roles in technology-facilitated peer assessment. *British Journal of Educational Technology*, *43*(3), 376–384. http://doi.org/10.1111/j.1467-8535.2011.01180.x

Liu, N., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279–290.

Lu, J., & Law, N. (2011). Online peer assessment: effects of cognitive and affective feedback. *Instructional Science*, *40*(2), 257–275. http://doi.org/10.1007/s11251-011-9177-2

Luke, D. A. (2004). *Multilevel Modeling*. Thousand Oaks, California: Sage Publications.

Luxton-Reilly, A., Hamer, J., Denny, P., Luxton-Reilly, A., & Hamer, J. (2008). The PeerWise system of student contributed assessment questions. In *Proc. 10th Australasian Computing Education Conference* (Vol. 78, pp. 69–74).

Maas, C. J., & Hox, J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Journal of Research*

Methods for the Behavioral and Social Sciences, *1*(3), 86–92. http://doi.org/10.1027/1614-1881.1.3.86

McQueen, H. A., Shields, C., Finnegan, D. J., Higham, J., & Simmen, M. W. (2014). Peerwise provides significant academic benefits to biological science students across diverse learning tasks, but with minimal instructor intervention. *Biochemistry and Molecular Biology Education: A Bimonthly Publication of the International Union of Biochemistry and Molecular Biology*, 1–11. http://doi.org/10.1002/bmb.20806

Moore, J. W. (1989). Tooling Up for the 21st Century. *Journal of Chemical Education*, *66*(I), 15–19.

Nelson, M. M., & Schunn, C. D. (2008). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science*, *37*(4), 375–401. http://doi.org/10.1007/s11251-008-9053-x

Nicol, D. (2010). *The foundation for graduate attributes: Developing self-regulation through self and peer assessment*.

Palmer, E., & Devitt, P. (2006). Constructing multiple choice questions as a method for learning. *Annals of the Academy of Medicine, Singapore*, *35*(9), 604–8.

Prince, M. (2004). Does Active Learning Work? A Review of the Research. *Journal of Engineering Education*, *93*(July), 223–231. http://doi.org/10.1002/j.2168-9830.2004.tb00809.x

Purchase, H., Hamer, J., Denny, P., & Luxton-Reilly, A. (2010). The quality of a PeerWise MCQ repository. In *Proceedings of the Twelfth Australasian Conference on Computing Education - Volume 103* (pp. 137–146). Darlinghurst, Australia, Australia: Australian Computer Society, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1862219.1862238

Robson, K., & Pevalin, D. (2016). *Multilevel Modeling in Plain Language*. Lomdon: Sage Publications Ltd.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching Students to Generate Questions: A Review of the Intervention Studies. *Review of Educational Research*, *66*(2), 181–221. http://doi.org/10.3102/00346543066002181

Ryan, B. J. (2013). Line up, line up: using technology to align and enhance peer learning and assessment in a student centred foundation organic chemistry module. *Chemistry Education Research and Practice*. http://doi.org/10.1039/C3RP20178C

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London: Sage Publications.

Sykes, A., Denny, P., & Nicolson, L. (2011). PeerWise - The Marmite of Veterinary Student Learning. In S. G. and A. Rospigliosi (Ed.), *Proceedings of the 10th European Conference on E-Learning, Vols 1 and 2*.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Boston: Pearson.

Vygotsky, L. (1978). Interaction between learning and development. In *Mind and Development* (pp. 79–91). Cambridge: Harvard University Press.

Wilson, E. V. (2004). ExamNet asynchronous learning network: augmenting face-to-face courses with student-developed exam questions. *Computers & Education*, *42*(1), 87–107. http://doi.org/10.1016/S0360-1315(03)00066-6

Yu, F.-Y., & Chen, Y.-J. (2014). Effects of student-generated questions as the source of online drill-and-practice activities on learning. *British Journal of Educational Technology*, *45*(2), 316–329. http://doi.org/10.1111/bjet.12036

Yu, F.-Y., & Liu, Y.-H. (2005). Potential Values of Incorporating a Multiple-Choice Question Construction in Physics Experimentation Instruction. *International Journal of Science Education*, *27*(11), 1319–1335. http://doi.org/10.1080/09500690500102854

Yu, F., & Wu, C.-P. (2012). Student Question-Generation: The Learning Processes Involved and Their Relationships with Students ' Perceived Value. *Journal of Research in Education Sciences*, *57*(4), 135–162.

Supplementary Material

Table S1: PeerWise requirements and marking scheme by course in each academic year studied.

	2011–2012	2012–2013	2013–14
Physics 1A Edinburgh	**3 deadlines**	**2 deadlines**	**2 deadlines**
Author	3	2	2
Answer	15	10	1
Comment	9	6	6
% mark	6%	4%	4%
Scoring	If students meet minimum requirements, those with the with lowest PW score get 40%. Below the minimum, marks are linearly fitted between 0 and 40%. Student with highest PW score whilst only completing minimum requirements. gets 70% students below this get a mark between 40 and 70%. The student with the highest PW score gets 100% - linear interpolation between those students who scored 70% and the maximum score students	At least 1 question submitted but other minima not satisfied: 25% All minima satisfied and scoreboard score below class median: 75% All minima satisfied and scoreboard score above class median: 100%	As in Physics 1A 2012–13
Physics 1B Edinburgh	**1 deadline**	**1 deadline**	**1 deadline**
Author	1	1	1
Answer	5	5	5
Comment	3	3	3
% mark	1%	1%	1%
Scoring	As in Physics 1A 2011–12	As in Physics 1A 2012–13	As in Physics 1A 2012–13
Chemistry 1B Edinburgh	**2 deadlines**	**2 deadlines**	**1 deadline**
Author	2	2	2 (max of 10)
Answer	20	20	20
Comment	6	6	6
% mark	3%	5% for fulfilling requirements	5%
Scoring	As in Physics 1A 2011–12		2% for meeting minimum requirements. 3, 4, 5% for scores above 3000, 4000 and 5000 respectively
GGA Edinburgh	**1 deadline**	**1 deadline**	**1 deadline**
Author	2	2	2 (max of 10)
Answer	20	20	20
Comment	5	5	5
% mark	4%	5%	5%
Scoring	2% for meeting minimum requirements, 3–4% depending on PW score, relative to cohort. Marked similarly to Physics 1A 2011–12	2% for fulfilling requirements, 3–5% depending on PW score, relative to cohort. Marked similarly Physics 1A 2011–12	1% if below minimum requirements. 2% for meeting minimum requirements 3, 4, 5% for scores above 3000, 4000 and 5000 respectively
Physics2 Glasgow.	**2 deadlines**	**2 deadlines**	**2 deadlines**
Author	4	4	4
Answer	8	8	8
Comment	4	4	4
% mark	1.67%	1.67%	1.67%
Scoring	As in Physics 1A 2011–12	As in Physics 1A 2011–12	As in Physics 1A 2011–12
Foundations of Chemistry Nottingham.	**1 deadline**	**1 deadline**	**1 deadline**
Author	1	1	1
Answer	5	5	5
Comment	3	3	3
% mark	5%	5%	5%
Scoring	As in Physics 1A	2% meeting min. 3% Exceeding min. engagement & below median score 5% Exceeding min. engagement and above median score	As in Foundations of Chemistry 2012–13

Table S2: Number of students in each course

	Course					
	Physics 1A (%)	**Physics 1B (%)**	**Chemistry 1B (%)**	**GGA (%)**	**Glasgow Physics. (%)**	**Nottingham. Chemistry. (%)**
2011–12	172 (83.1)	90 (50.6)	155 89.6)	213 (76.6)	138 (90.8)	162 (95.9)
2012–13	245 (84.2)	131 (59.5)	136 (78.6)	232 (79.5)	151 (87.8)	167 (92.3)
2013–14	269 (88.2)	138 (55.6)	164 (77.7)	220 (76.9)	133 (90.5)	155 (85.6)

Coding of student comments: reliability testing

All the coding of student comments was carried out by AK. As a check of reliability, over 10% of Physics 1A 2012–13 comments were coded by JH, with minimal discussion about how to apply the scheme. The Spearman correlation between the original and recoded samples was .845, $p < 0.01$, indicating a high correlation between the application of the scheme by both coders. Cohen's Kappa was calculated as .783, $p < .001$ – indicating strong agreement between coders. The coding scheme was therefore deemed to be sufficiently reliable.

Table S3: Comment coding scheme

Code	Description
1	Symbols Nonsensical/off topic comments Reply to another comment without deep engagement Where student states they clicked the wrong button by mistake, or just reiterates what answer they chose Where student just states they got the question correct/incorrect **Examples:** : "*HAHAHA*" "*Yeah, thanks*" "*Got it wrong*"
2	Non-specific comment about ease/difficulty; whether good/bad; whether helps understanding Non-specific expressions of thanks for previous feedback Non-specific statement of own understanding/whether the question tripped the answerer up or whether it clarified matters **Examples:** : "*Very interesting question*" *"Good question"* *"Easy question"*
3	Specific mention of distractors/traps/explanation/ Specific evaluation of why the question is good/bad; difficult/easy; why they like/dislike it Specific suggestions how to improve question/other options for distractors or solutions Specific evaluation of their own ability and understanding Specific evaluation of how the explanation/question has helped improve understanding Specific recognition that the question combines different aspects of the course/sheds new insight into a topic Specific expressions of thanks stating how writing question helped own understanding; agreeing/disagreeing with previous commentators Specific request for further assistance because of lack of understanding **Examples:** : "*Good question with a good, clear explanation. Challenging as well*" "*Would the one sliding down not be the fastest as all the potential energy is converted into linear motion i.e. 1/2mv^2 and none rotational motion?*" "*I did it just using conservation of energy. Work done by pulling a rope W=F*d=5*8=40N. Work is a change in cylinders Kinetic energy, so W=0.5*I*w^2, and from here we get, that w=141.4rad/s. answer is the same, and it is much faster this way. Nice problem though, thanks*".