Students developing test items: a review

Veronica Bruijns, Amsterdam University of Applied Sciences
Email: v.m.h.bruijns@hva.nl

Abstract

In this review 21 recent articles on students making test items are analyzed. 14 articles are concerned with the generation of multiple choice questions (MCQ's) by students. In 2 cases students made open questions and in 2 cases a mix of MCQ's and open questions was used. In 3 cases students' involvement had the form of a student-written exam. In 18 articles positive results are reported on three main topics: learning strategy, academic achievement and quality of questions. The research design is either qualitative research or case study, except for one experimental design. This means that substantiated inferences by means of causes and effects cannot be drawn. However, there seems to be a strong positive relationship between students' engagement in designing test items and learning strategy, academic achievement and appreciation. Students are able to design questions that meet most of the quality criteria. The positive relationship can be enhanced when students are actively involved in assessment activities in the entire assessment program of a study program.

Key words: student engagement, student-written exams, student-written test items, PeerWise.

Introduction

Students developing test items is widely used as a learning strategy. It has supposed benefits. By making questions students better understand the relation between learning objectives and course content. It strengthens their analytic and problem solving skills. They have to decide which concepts and theories are essential, given the topics of the course. By making questions they have to decide what is crucial about the chosen concept and theories and reason about misconceptions and misunderstandings. It also helps them to improve their ability to take tests.

The assumption regarding learning style are that by generating test items students apply a deep level learning strategy. Regarding academic achievement it is assumed that generating questions leads to applying higher order thinking skills. The assumption regarding quality of questions is that students, with the necessary scaffolding, are able to design higher order questions on the scale of the Bloom's taxonomy. The supposed overall effect is that it results in a better performance on the final, summative test. The aim of this article is to list evidence form recent research articles regarding these assumptions.

Students' active involvement in assessment activities is part of a broader development in (higher) education to engage students in partnership with all aspects of (higher) education (Healey, Flint,

and Harrington 2014). The origin of this development is the deeply felt need to engage students as active partners in education instead of passive consumers, in order to prepare them adequately for working and living in the 21st century. It takes shape in theories on social constructivism, student-centered pedagogy, personalized learning, communities of practice, as such. Active involvement with assessment is broader than the subject of this article: students can be given a choice of assessment forms, can act as their peers' assessors or can collaborate together in assessment activities.

In this article the results of a review study on students developing test items are being presented. The research on students generated test items and the supposed benefits includes, based on the articles in the review, three main topics. The first topic is the relation between students' generating test items and their learning strategy. The second area is the effect on academic achievement, often combined with evaluation of students' experiences. A third well researched topic is the quality of student-generated test items.

Selection of articles

A previous database search on this topic by a graduate student had resulted in 7 relevant articles on this subject. By following the references (snowballing) 4 other relevant articles were identified. Most researchers used PeerWise as a tool. With this search term 9 articles were found (ERIC, Jan. 2000-Oct. 2017, higher education, peer reviewed), from which 5 articles were identified as relevant. Following these references (snowballing) led to another 5 articles .The literature search resulted in twenty one relevant articles (see Table 1).

Categories of student involvement in test item development

The literature on the subject of student involvement in test item development provides, in general, four categories. The first category is students developing MCQ and is most frequently applied. This is often combined with evaluating and rating MCQs developed by fellow students. The second category is developing open questions, mainly short answers or essay questions. The third is a mix of MCQ and open questions. The questions the students generate can be used in the final exam. The fourth category is a student-written exam, whereby the student drafts the exam questions and answers these questions as the final exam. The assessor rates the breadth and depth of the questions and the correctness of the answers.

Mode	First author (year)	Research mode	N	Field of study	Country
MCQ	Fellenz (2004)	Qualitative review	34	Management	Ireland
	Arthur (2006)	Case study	40	Economics	Australia
	Palmer (2006)	Experiment	104	Medicine	Australia
	Denny (2008)	Qualitative review	460	Computer Science	Australia
	Bottemley (2011)	Case study	107	Biochemistry	Australia
	Pittinger (2011)	Qualitative review	89	Pharmacy	USA
	Bobby (2012)	Case study	84	Biochemistry	India
	Rhind (2012)	Qualitative review	442	Vets	New Zealand
	Bates (2014)	Qualitative review	*	Physics	Canada
	Hardy (2014)	Case study	854	Physics, Chemistry, Biology	United Kingdom

	Galloway (2014)	Case study	326	Chemistry	United Kingdom
	McQueen (2014)	Qualitative review	750	Medicine	United Kingdom
	Rhodes (2014)	Qualitative review	32	Nursing	New Zealand
	McKenzie (2017)	Case study	20	Psychology	Australia
Open questions	Bobby (2007)	Case study	66	Biochemistry	India
	Luxton-Reilly (2012)	Case study	466	Computer Science	New Zealand
MCQ and open questions	Ahn (2011)	Case study	**	Teacher education	USA
	Papinczak (2012)	Case study	282	Medicine	Australia
Student-written exam	Bearheim (2003)	Case study	90	Medicine	Norway
	Brink (2004)	Case study	15	Engineering	USA
	Corrigan (2013)	Case study	**	Marketing	USA

Table 1: overview of the articles (MCQ= multiple choice questions)

* in this study the object of N were test items instead of students

** not mentioned

The majority of the articles (14) focuses on the generation of MCQs by students, many of them using PeerWise, a digital tool for generating and evaluating MCQs. In 2 cases students made open questions and in also 2 cases a mix of MCQs and open questions was used. In 3 cases students' involvement had the shape of a student-written exam.

Most of the studies are either case studies or qualitative reviews. Case study is defined as: descriptive studies using a mixed methods approach to answer different research questions such as quality of questions, academic performance or student appreciation. Qualitative reviews use questionnaires or focus group discussions to evaluate student appreciation of generating, reviewing and rating test items. Only in one case an experimental setting was used with two randomized groups (Palmer and Devitt 2006). The number of students involved ranges from 15 participants (Brink, Capps, and Sutko 2004) to 854 (Hardy et al. 2014). The fields of study of the students involved are predominantly natural sciences, medicine and computer science. Almost all the articles - 19 out of 21 - stem from Anglo-American countries. A reason for this could be the use of English web based tools as a dominant instrument.

In the following two paragraphs a brief description is given of the two main categories found in the literature: first, developing MCQ's and open questions or a combination of both and, second, the student-written exam.

Developing multiple choice and open question

The activity students undertake by making questions is first choosing the relevant idea, concept or method out of the course material as topic of a question. In the case of a MCQ they have to formulate the stem and the distractors. The distractors have to be reasonable and must relate to misconceptions or incorrect thinking directions. Students also have to substantiate the right and wrong answers. Luxton-Reilly et al. (2012) point out that: 'Generating the alternatives is perhaps more interesting than the question stem, since the best alternatives are ones that reveal a possible misconception' (151). In case of open questions the student formulates the questions and the answer key. Developing questions is not the only activity students undertake. They are

also invited to answer questions from their fellow students and to review and rate these questions. This reviewing process results in peer feedback.

Most authors state the importance of considering designing questions as part of a more extensive process. It is equally important that students answer, evaluate, discuss and rate other questions. Some authors facilitate the evaluation and rating by organizing small group discussions. Bobby et al. (2007) see this as an efficient review exercise. Students can use questions authored by other students also as a learning resource (Luxton-Reilly et al. 2012).The interesting concept behind this form of student engagement is that students are not only stimulated to be actively involved with the course content but also practice with evaluating and giving peer feedback.

Web based tools are developed to optimize this process. For instance, many researchers facilitate students making questions by using PeerWise. PeerWise is an online, open source tool developed by the department of computer science at the University of Auckland, New Zealand. It provides three main activities: authoring, answering, and commenting MCQs. The tool provides the author of a MCQ insight in how fellow students answered the question. Students can rate other's questions on criteria such as difficulty and quality and provide and discuss comments. The tool is designed for use by students with a minimum of monitoring by staff.

In order to prepare students it is first of all important to present the reason behind this approach and explain what the benefits are (Luxton-Reilly et al. 2012). Moreover, it is important to provide scaffolding by means of guidelines, instructions and good examples (Luxton-Reilly et al. 2012; Pittenger and Lounsbery 2011) and opportunity to exercise and discuss questions in class (Fellenz 2004; Luxton-Reilly et al. 2012; Palmer and Devitt 2006). Luxton-Reilly et al. (2012) add an extra gain to discussion in class: 'This may also serve the purpose of focusing student attention on the learning outcomes of a course, discussing in detail the aims and goals of the course and how they might be assessed' (156). Pittenger and Lounsbery (2011) point out that scaffolding is effective and necessary in order to make students authoring questions successful.

Most authors state that, also to communicate and discuss benefits, students should be motivated to participate. Also intrinsically motivated students benefit from an external incentive. Different strategies are used to stimulate students. First, the activity of making questions is part of the workload of the course, and not an extra added-on activity. Second, the final test contains a certain amount of high quality questions made by students. Third, students receive a bonus on the final mark when they have done their share of pre-arranged activities.

Student-written exam

The concept of a student-written exam differs from the idea of authoring MCQs and open questions as described in the previous paragraph. A student-written exam is an individual activity whereby the student develops the questions for the final exam and completes the answers. The instructor specifies the content to be covered, the learning objectives, the number of questions and the cognitive level (Corrigan and Craciun 2013). The instructor rates the student-written exam to the extent the course content is covered, level of the questions - on the basis of Bloom's taxonomy - and accuracy and completeness of the answers. Corrigan and Craciun (2013) added

a feedback session in class where students presented one example of their questions for discussion in class.

The authors list several potential benefits of this exam method (Corrigan and Craciun 2013). First, by writing their own exam questions and correct answers students can take responsibility for their own learning and evaluation, which provides a challenging learning experience. Second, this method allows students to be more creative about problem formulation and solution development. Third, students can experience less exam-related stress than with traditional exams and cheating can be minimized. And finally, the "role reversal" may strengthen students' trust in teaching and evaluation methods and improve the teacher–student relationship. Brink, Capps, and Sutko (2004) state that 'the creation of an exam and answer key by the student forces the students to do a "role reversal" and to consider what could be covered on an exam' (269).

Research topics and findings

Table 2 gives an overview of the research objectives and main findings of the articles on MCQs and open questions generated by students and student-written exams. The assumptions regarding learning style is that by generating test items students apply a deep level learning strategy. Regarding academic achievement it is assumed that generating questions leads to applying higher order thinking skills and to higher results on the final test. The assumption regarding quality of questions is that students, with the necessary scaffolding, are able to design higher order questions on the scale of the Bloom's taxonomy.

Form	First author (year)	Research object(s)	Main finding(s)
MCQ	Fellenz (2004)	-student appreciation	-valuable in learning -increased time-on-task -better understanding mc items -more ownership
	Arthur (2006)	-student appreciation	-more active approach to learning -link between learning objectives and course material -identify relevant topics/methods -improve test taking strategies
	Palmer (2006)	-student appreciation	-before: no stimulating method; afterwards more favorable -no apparent gain in academic achievement
		-academic achievement -quality questions	-high standard questions
	Denny (2008)	-academic achievement	-significant correlation between performance and PeerWise activity
	Bottemley (2011)	-academic achievement	-significant correlation between end of semester mark and PeerWise score
		-quality questions	-most questions on first three levels of Bloom
	Pittenger (2011)	-student engagement	-agreed or strongly agreed meaningful learning activity -allowed for personalized learning
	Bobby (2012)	-student appreciation -academic achievement	-useful revision exercise -significant higher marks post-test
	Rhind (2012)	-support student learning	-majority agreed or strongly agreed that both authoring and answering was helpful for their studies
		-academic achievement	- significant correlation between number of questions answered in PeerWise and the final exam results
	Bates (2014)	-quality questions	75% of the questions met all the quality criteria
	Hardy (2014)	-academic achievement	-Modest statistically significant positive correlation

	Galloway (2014)	-academic achievement	-significant correlation between PeerWise activity and performance end-of-year examinations
		-student contribution	-high performing student contribute more questions
		-quality questions	-86% classified as high quality questions
	McQueen (2014)	-Academic achievement	-statistically significant association between engagement with PeerWise and academic performance
		-Quality questions	-25% Category 3 [Bloom] or above, the majority of questions fell into Category 2
	Rhodes (2014)	-student appreciation	-provided opportunities for knowledge building and consolidating understanding
	McKenzie (2017)	-student engagement	-authoring and answering questions in PeerWise equally benefited students' learning
Open questions	Bobby (2007)	-academic achievement	-highly effective in improving understanding - mean post-test scores significantly higher than mean pre-test scores
	Luxton-Reilly (2012)	-student engagement	-useful learning activity
MCQ and open questions	Ahn (2011)	-learning style	-facilitated social constructivist learning
	Papinczak (2012)	-student appreciation	-supported students to work collaboratively -no direct improvement in metacognitive capacity
		-quality questions	- Many questions of satisfactory to good quality
Student-written exam	Bearheim (2003)	-student appreciation	-60% had changed their learning strategies -5-7% higher marks on their own questions
	Brink (2004)	-academic achievement	-students who develop good model tests and answer keys make higher final exam grades
	Corrigan (2013)	-student appreciation	-improved the relevance of exam questions increased student involvement with learning and self-evaluation -helped students manage exam stress

Table 2: Overview student generated questions, research object and main findings.

Deep level learning strategy

One of the supposed benefits is that students through the authoring of MCQs and open questions apply a deep level learning strategy as opposed to a surface approach of learning. Pittenger and Lounsbery (2011) found that making questions required students to more thoroughly study and understand the course materials. Students reported making questions, formulating distractors, and justifying the correct and incorrect answers as a challenging activity. The authors stated that making questions demands an active involvement of students, which stimulates deep level learning. According to Pittinger and Lounsbery (2011) it is different from answering questions posed by the teacher. By making questions students have to provide explanation for the right answer and explain why other options are incorrect**.** Fellenz (2004) describes that providing justification 'requires students to make explicit their understanding of the complexities of the subject matter' (711).

Students, in the research that Rhind and Pettigrew (2012) conducted, reported: 'An increased depth and breadth of knowledge and understanding; development of a useful learning and revision resource; and the benefit and enjoyment of the peer discussion focused on the questions' (376). Important to note is that participation in generating questions was on a voluntary basis. Luxton-Reilly (2012) report that next to content specific knowledge students also acquire metacognitive knowledge, for instance standards and expectations.

According to some authors making questions can heighten the awareness of one's own learning style. Ahn and Class (2011) researched how students' generated exams challenged the students to think about their approach to learning. They concluded that students become 'active agents of their own learning' (269). Arthur (2006) concluded on the basis of his research that students developed a more active approach to learning and Baerheim and Meland (2003) reported that 60% of the respondents changed their learning style as a consequence of making their own written exam questions. The students in Fellenz's (2004) research reported increased reflection on their own learning approaches.

Two researchers reported negative effects of student generated items. Pittenger and Lounsbery (2011) found that 15% of the respondents made negative comments regarding the student-generated question assignment. About half of these students said the assignment negatively impacted their learning, explaining that instead of engaging more with the content, they focused on finding details around which they could construct questions. Papinczak et al. (2012) found a negative effect regarding learning style; students tended to memorize the student generated questions, which led to a surface approach of learning. They concluded thus that an activity designed to enhance metacognition can have a paradoxical effect on students' learning styles.

Academic achievement

Applying a deep level learning strategy stimulates the development of higher order thinking skills, one of the aspects of academic achievement. According to Bottomley and Denny (2011) these skills can be summarized as: research, evaluate, give and receive feedback. Denny et al. (2008) define higher order learning activities as: 'help to consolidate, reinforce and deepen understanding, by engaging students in cognitively demanding tasks: reviewing, summarizing, clarifying, giving feedback, diagnosing misconceptions, identifying missing knowledge, and considering deviations from the ideal' (51). They also mention reflection and social and professional skills. Luxton-Reilly et al. (2010) add self-assessment skills**:** 'The main benefits of authoring questions are to engage students in activities that focus on the content of a course, and to encourage them to develop self-assessment skills so they can evaluate their own understanding of course content with respect to the defined learning outcomes' (151).

Several authors emphasize that, although most of the MCQs are on the first three levels of Bloom revised taxonomy (see next paragraph) the construction of MCQs invokes higher cognitive levels. Fellenz (2004) mentioned that students have to justify why answer options are correct or incorrect. This demands 'evaluation and the construction of compelling arguments' (711). Students learn to use clear and unambiguous language (Bottomley and Denny 2011). Bottomley and Denny (2011) also state that students have to use higher order thinking skills even if they construct questions on the first three levels. Bates et al. (2014) suggest 'that students are substantially more creative than we might have previously given them credit for, and this creativity might be usefully harnessed in meaningfully developing core skills (such as problem solving) within the discipline' (10).

The second assumption regarding academic achievement is that students' involvement leads to a better performance on the final test. In 11 cases the relation between students authoring

questions and academic achievement was researched (see table 3). In 9 cases the authors used final marks on the summative test as a measure for academic achievement.

Form	First author (year)	finding(s) academic achievement
MCQ	Palmer (2006)	-no apparent gain in academic achievement
	Denny (2008)	-significant correlation between performance and PeerWise activity
	Bottemley (2011)	-significant correlation between end of semester mark and PeerWise score
	Bobby (2012)	-significant higher marks post-test
	Hardy (2014)	-Modest statistically significant positive correlation
	Galloway (2014)	-significant correlation between PeerWise activity and performance end-of-year examinations
	McQueen (2014)	-statistically significant association between engagement with PeerWise and academic performance
	Rhind (2014)	- significant correlation between number of questions answered in PeerWise and the final course examination results
Open questions	Bobby (2007)	-highly effective in improving understanding -mean post-test scores significantly higher than mean pre-test scores
Student-written exam	Bearheim (2003)	-5-7% higher mark on their own questions in the final exam
	Brink (2004)	-students who develop good model tests and answer keys make higher final exam grades

Table 3: Overview findings academic achievement

With the exception of the research conducted by Palmer and Devitt (2006) the researchers found a significant difference between pre- and post-test scores. This is an interesting finding because Palmer and Devitt (2006) were the only researchers who conducted an experiment, with a random intervention and control group.

An interesting aspect of the research on academic achievement is differences in high and low performing students. This was part of the research done by Bobby et al. (2007, 2012), Hardy et al. (2014) and McQueen et al. (2014). Bobby et al. (2007) combined individual formulation of questions with group discussion on the answers. In the 2007 research they found that high achievers gain more from the formulation of questions as opposed to the group discussion. The gain from group discussions was higher among low and medium achievers. In the 2012 research they found no significant differences in gain among high, medium and low achievers.
Hardy et al. (2014) and McQueen et al. (2014) divided the students in four quartiles, based on prior learning abilities. Hardy et al. (2014) concluded that 'no immediately obvious pattern emerges in terms of which types of PeerWise activity had most impact on which student quartiles' (2190). They found little evidence that high performing students benefit more from writing questions compared to low performing students. In two cases they found that students in the lowest quartiles showed a significant effect between writing questions and examination marks. Overall they concluded that students in the lower/intermediate part of the classes may have benefitted most. McQueen et al. (2014) found that high ability and low/intermediate ability students gained the most benefit from PeerWise activities. They concluded that high ability (quartile 4) students benefit because they are expected to engage well with all aspects of their course and have a firm knowledge base on which to build. The low/intermediate ability students (quartile 2) benefit because PeerWise offers 'an activity that is challenging enough to be useful,

but not so challenging as to be beyond their grasp' (378) For the high/intermediate group (quartile 3) their explanation is that 'this group might take a "strategic" learning approach, putting in the minimum effort required to gain marks' (378). Brink, Capps and Sutko (2004) found different effects in their research on student-written exam. Their results indicate that the student-written exam seemed more effective for the above average students than for those below average. Their explanation is that 'the above average student are better at predicting what the instructor will ask on a final exam' (268).

Quality of the questions

In 6 cases (see Table 4) the quality of the questions was object of research (Bates et al. 2014; Bottomley and Denny 2011; Galloway and Burns 2015; McQueen et al. 2014; Palmer and Devitt 2006; Papinczak et al. 2012). The quality is rated on the following aspects: format aspects (language, format), cognitive level, accurate answer and quality of the distractors, quality of the justification of the right answer and the distractors and compliance with the course content.

Form	First author (year)	Finding(s) quality questions
MCQ	Palmer (2006)	-high standard questions clearly displaying understanding of the topic concerned -75% remember/understand; 25% higher levels.
	Bottemley (2011)	-56.2% remember, 34.8% understand, 8.8% apply
	Bates (2014)	-75% of the questions met all the quality criteria -all taxonomic levels, with a majority of the questions in the middle categories of apply and analyze
	Galloway (2014)	-86% classified as high quality questions
	McQueen (2014)	-25% apply or above, 75% on level understand
MCQ and open questions	Papinczak (2012	- Many questions of satisfactory to good quality

Table 4: Overview findings quality questions.

Although there is criticism on Bloom's taxonomy with regard to the order of its hierarchy and mismatch with actual process of learning, for the evaluation of the cognitive level of MCQs, Bloom's revised taxonomy is used as the dominant framework. This framework comprises 6 ascending levels of complexity: 1) remembering, 2) understanding, 3) applying, 4) analyzing, 5) evaluating, 6) creating. Designing MCQs on the higher levels of Bloom's taxonomy is challenging, also for instructors. A study of Momsen et al. (2010) examining almost 10,000 assessment items from 50 instructors in the USA reported that 93% of the examination questions in introductory biology modules were at the lowest two levels of Bloom's taxonomy: remembering and understanding. They noticed that the field of study can have some influence on the level of the questions.

All researchers report that students are capable of designing questions that meet most of the quality criteria. However, there are clear differences in the findings on the levels of the questions according to Bloom's taxonomy. Palmer and Devitt (2006) found that although the students were capable of producing high quality MCQs, the majority of the MCQs created by them were on the two lowest levels of Bloom's taxonomy: 'remembering' and 'understanding'. Bottemley and

Denny (2011) reported that 56.2% of the questions were on level 1 'remembering', 34.8% on level 2 'understanding', 8.8% on level 3 'applying'. Bates et al. (2014) concluded that 'not only do students produce, on the whole, very good questions, but also the appropriately detailed and useful explanations to accompany them' (9). 75% of the questions met all the quality criteria. There was a broad distribution across all cognitive levels, with a majority of the questions in the middle categories of 'applying' (level 3) and 'analyzing' (level 4) of Bloom's taxonomy. Galloway and Burns (2014) found the same results as Bates et al. (2014) with one notable difference in the relative proportion of level 4 'analyzing' and level 5 'evaluating' questions. In the case of the chemistry study there are significantly more questions on the level of 'evaluating than 'analyzing', whereas the situation is reversed for the results of the physics course in Bates' et al. (2014) research. This can be explained by the nature of the two subjects, according to Galloway and Burns (2014): 'Chemistry is more focused on qualitative or conceptual evaluation questions, whereas physics has more defined multi-step problems that may require a sequence of mathematical treatments' (87). This led Galloway and Burns (2014) to the following observation: 'Indeed, one must be cautious in comparisons due to the differences in degree course structures between subjects, institutions and countries. There is also a possibility of some variation in the particular interpretation of the taxonomy levels used for classification by different researchers' (87).

Student engagement and appreciation

Besides the main three areas of research described above many researchers included students' engagement and appreciation as a research objective. How inspiring designing your own questions can be is illustrated by a student's reaction reported by Ahn and Class (2011): 'I could not help feeling a moment of delight, when I first heard we would be writing our own example questions and allowed to use our notes and textbook during the midterm. "Could anything be easier?" What ensued was instead one of the most interesting and difficult tasks I have undertaken as a credential candidate' (272).

In 14 cases student engagement or appreciation in relation to developing questions was object of research (see Table 5). In most cases the vast majority of students reacted positively on the activity, except for Papinczak et al. (2012). In some cases high scores on appreciation were reported (Bobby et al. 2012; Rhind 2012, Rhodes 2013).

Form	First author (year)	finding(s) student engagement/appreciation
MCQ	Fellenz (2004)	-valuable in learning -increased time-on-task -better understanding mc items -more ownership
	Arthur (2006)	-more active approach to learning -link between learning objectives and course material -identify relevant topics/methods -improve test taking strategies
	Palmer (2006)	-before: no stimulating method; afterwards more favorable
	Bottemley (2011)	-70% agreed PeerWise helped learning biochemistry, 60% liked the rating system

	Pittenger (2011)	-agreed or strongly agreed meaningful learning activity -allowed for personalized learning
	Bobby (2012)	-95% wanted to have such sessions in future, from which 78% preferred to have such exercises at the end of each chapter, 21% once in a week and 1% once in 6 months -gain from the exercise: 81% very much; 17% little; 2% nothing -99% did not feel that the exercise was a burden
	Rhind (2012)	-majority agreed or strongly agreed that both authoring and answering was helpful for their studies
	Galloway (2014)	-high performing student contribute more questions
	Rhodes (2014)	-90% would like to use PeerWise again for knowledge building
	McKenzie (2017)	-authoring and answering questions in PeerWise equally benefited students' learning
Open questions	Luxton-Reilly (2012)	-useful learning activity
MCQ and open questions	Papinczak (2012)	-32.8% found the activity supported their learning; 29.9% their problem-solving ability -77.3% supported the continuation of the project for ensuing cohorts
Student-written exam	Bearheim (2003)	-60% had changed their learning strategy
	Corrigan (2013)	-improved the relevance of exam questions, increased student involvement with learning and self-evaluation, and helped students manage exam stress

Table 5: Overview findings 'student engagement/appreciation'

The PeerWise tool offers different activities that students find helpful for their learning: review content, reflect on content, review answers of other students, self-evaluation, evaluating others, receiving feedback, compare with peers (Luxton-Reilly et al. 2012). In the case of student-written exams the students experienced the following benefits: in-depth learning, higher involvement, greater responsibility and less stress (Corrigan and Craciun 2013). Despite the mixed results on academic achievement, McKenzie and Roodenburg (2017) found that students considered authoring and answering questions in PeerWise equally benefitting their learning.

One of the relevant factors in the appreciation and engagement of students is the fact that high quality questions generated by students are part of the final exam. Fellenz (2004) used more than half of the MCQs on the exam authored by students. His motivations is that: 'this increases student ownership of the assessment procedures used and motivates students to participate and submit high quality multiple choice items' (706). This also applies to the incentive used in order to stimulate participation. Students receive a 1% to 5% bonus on the final mark if they have constructed and reviewed the requested amount of questions (Hardy et al. 2014). However, this can also lead to an unhealthy competition, according to McQueen et al. (2014). They have set a maximum number of ten questions that students can contribute and restricted the bonus to 3%, 4%, and 5% depending on the PeerWise score.

Corrigan and Craciun (2013) also included experiences of teachers. They stated that grading time was similar to that required by teacher-written exams. Their experience showed 'that it was less tedious to grade different questions than to read and grade the same answers multiple times' (34). They emphasize that the method works well in small groups and that it may be too time consuming in classes over 50 students.

As mentioned earlier PeerWise is a tool that requires little staff involvement. According to Hardy et al. (2014) this means low costs and independence of course, institution, instructor or student. Rhind and Pettigrew (2012) found that students have different views regarding the importance of staff monitoring or reviewing of the process. They found 'that students perhaps in earlier years have a greater need for reassurance and this is supported by the qualitative data, indicating that staff presence in the system and the ability to review gives the student necessary confidence in the quality of the resource being generated. This, however, did not seem to be an issue for the students who had progressed further in the curriculum; in this group there was a strong theme around confidence in the ability of the peer group to perform the role more than adequately' (379).
An other important finding on students' appreciation is that it helped managing exam stress (Corrigan 2013) and reduced anxiety (Bearheim and Meland 2003; Papinckzak et al. 2012).

The exception to the positive experiences are the results of Papinczak et al. (2012): 'On the whole, many participants did not consider that the process of generating examination questions improved their problem-solving abilities or assisted in regulation of understanding' (448). In this case students developed questions in small groups; this condition was not part of the evaluation. Their impression is that the intensity with which the different groups engaged in the activity has influenced the appreciation.

Conclusions and discussion
Most of the studies in this review were either case studies or qualitative reviews. Only Palmer and Devitt (2006) can be identified as experimental research with the inclusion of a control group. Therefore it is difficult to draw substantiated inferences by means of causes and effects. It can only be noted that there is a positive relationship between student engagement in designing test items on the one hand, and learning style and academic achievement on the other. As far as the quality of the questions is concerned it can be concluded that different researchers found that students are well capable of authoring good quality questions and that the level of the questions, according to Bloom's taxonomy depends on the subject matter and the support students have received. PeerWise is a solid tool to enhance the quality of the questions, because it offers the opportunity to discuss, review and rate questions. MCQs seem to have an advantage to open questions. The generation of distractors adds an extra dimension because students have to elaborate on possible misconceptions and misunderstandings.
Research on student engagement in assessment can be conducted in experiments. The default setting in most programs is teacher-directed item construction. In the experiment researchers can use this default setting as the control situation and design an experimental setting with student generated questions.

The advantage of engaging students in test item construction is the possibility to vary with the intensity of the engagement. This can range from students exercising with authoring questions to students writing and answering their own exam questions. It provides students with the opportunity to integrate the learning of the course content with demonstrating the acquired knowledge and skills. The learning objectives can be achieved by students developing questions

and rating, evaluating and answering questions made by fellow students. It can develop in a meaningful learning activity with a formative function, instead of preparation for the final exam. In the most far reaching situation students making and answering their own exams could make summative testing unnecessary and obsolete.

The above mentioned advantage, variety of intensity, is also a pitfall. The cases in this review give the impression of isolated small-scale initiatives by enthusiastic teachers/researchers motivated in giving students an active role and believing in its potential benefits. The studies are based on results of a particular course or a selected set of courses. They do not give the impression that the initiative is rooted in an enacted educational concept or as a characteristic of an overall testing program. The effect can be more powerful if is an integral part of the testing program of the whole curriculum and based on a learning concept of active involvement of students.
Another delicate matter is the power issue which is hardly mentioned by the authors of the studies in this review. The reason might be that students were not involved in the assessment and rating of other students. Sadler (1989) and Taras (2008) have addressed this issue. They claim that involving students means allowing students to play a role in an area where teachers were used to be the authority (Sadler 1989). It gives students access to power that teachers traditionally have (Taras 2008). Sadler (1989) states that it gives teachers the idea that it undermines their authority as a teacher. It is their specific professionalism to design test items and it is not the students' domain. They see it as a prerogative of being a teacher. It gives them a specific position. On the other hand, Sadler (1989) argues that teachers should teach students how to evaluate and assess their learning. Bobby et al. (2012) illustrates this issue by stating: 'Inside of every student there is a teacher and inside of every teacher there is a student' (171).

Biographical note

Veronica Bruijns is a staff member at the Amsterdam University of Applied Sciences. Her main topics are assessment policy and curriculum design. She is chief editor of www.score.hva.nl, a Dutch website on assessment in higher education.

References

Ahn, R., and M. Class. 2011. Student-centered pedagogy: Co-construction of knowledge through student-generated midterm exams. *International Journal of Teaching and Learning in Higher Education,* 23(2): 269-281.

Arthur, N. 2006. Using student-generated assessment items to enhance teamwork, feedback and the learning process. *Synergy,* 24: 21-23.

Baerheim, A., and E. Meland. 2003. Medical students proposing questions for their own written final examination: Evaluation of an educational project. *Medical Education,* 37(8): 734-738.

Bates, S. P., R. K. Galloway, J. Riise, and D. Homer. 2014. Assessing the quality of a student-generated question repository. *Physical Review Special Topics-Physics Education Research,* 10(2): 1-11.

Bobby, Z., B.C. Koner, M.G. Sridhar, H. Nandeesha, P. Renuka, S. Setia, S. Senthil Kumaran, and S. Asmathulla. 2007. Formulation of questions followed by small group discussion as a revision exercise at the end of a teaching module in biochemistry. *Biochemistry and Molecular Biology Education,* 35(1): 45-48.

Bobby, Z., M. R. Radhika, H. Nandeesha, A. Balasubramanian, S. Prerna, N. Archana, and D.N. Thippeswamy. 2012. Formulation of multiple choice questions as a revision exercise at the end of a teaching module in biochemistry. *Biochemistry and Molecular Biology Education,* 40(3): 169-173.

Bottomley, S., and P. Denny. 2011. A participatory learning approach to biochemistry using student authored and evaluated multiple-choice questions. *Biochemistry and Molecular Biology Education,* 39(5): 352-361.

Brink, J., E. Capps, and A. Sutko. 2004. Student exam creation as a learning tool. *College Student Journal,* 38(2): 262-373.

Corrigan, H., and G. Craciun. 2013. Asking the right questions: Using student-written exams as an innovative approach to learning and evaluation. *Marketing Education Review,* 23(1): 31-36.

Denny, P., J. Hamer, A. Luxton-Reilly, and H. Purchase. 2008. PeerWise: Students sharing their multiple choice questions. Paper presented at the *Proceedings of the Fourth International Workshop on Computing Education Research,* 51-58.

Fellenz, M. R. (2004). Using assessment to support higher level learning: The multiple choice item development assignment. *Assessment and Evaluation in Higher Education,* 29(6): 703-719.

Galloway, K. W., and S. Burns. 2015. Doing it for themselves: Students creating a high quality peer-learning environment. *Chemistry Education Research and Practice,* 16(1): 82-92.

Hardy, J., S.P. Bates, M.M. Casey, K.W. Galloway, R.K. Galloway, R. K., A.E. Kay, P. Kirsop, and H.A. McQueen. 2014. Student-generated content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education,* 36(13): 2180-2194.

Healey, M., A. Flint, and K. Harrington. 2014. Engagement through partnership: students as partners in learning and teaching in higher education. *The Higher Education Academy.*

Luxton-Reilly, A., P. Denny, B. Plimmer, and R. Sheehan. 2012. Activities, affordances and attitude: How student-generated questions assist learning. Paper presented at the *Proceedings of the 17th ACM Annual Conference on Innovation and Technology in Computer Science Education,* 4-9.

McKenzie, W., and J. Roodenburg. 2017. Using PeerWise to develop a contributing student pedagogy for postgraduate psychology. *Australasian Journal of Educational Technology,* 33(1): 32-47.

McQueen, H.A., C. Shields, D.J. Finnegan, J. Higham, and M.W. Simmen. 2014. Peerwise provides significant academic benefits to biological science students across diverse learning tasks, but with minimal instructor intervention. *Biochemistry and Molecular Biology Education,* 42(5): 371-381.

Momsen, J. L., T.M. Long, S.A. Wyse, and D. Ebert-May. 2010. Just the facts? introductory undergraduate biology courses focus on low-level cognitive skills. *CBE-Life Sciences Education,* 9(4): 435-440.

Palmer, E., and P. Devitt. 2006. Constructing multiple choice questions as a method for learning. *Annals-Academy of Medicine Singapore,* 35(9): 604-608

Papinczak, T., R. Peterson, A.S. Babri, K. Ward, V. Kippers, and D. Wilkinson. 2012. Using student-generated questions for student-centred assessment. *Assessment and Evaluation in Higher Education,* 37(4): 439-452.

Pittenger, A. L., and J.L. Lounsbery. 2011. Student-generated questions to assess learning in an online orientation to pharmacy course. *American Journal of Pharmaceutical Education,* 75(5): 94.

Rhind, S. M., and G.W. Pettigrew. 2012. Peer generation of multiple-choice questions: Student engagement and experiences. *Journal of Veterinary Medical Education,* 39(4): 375-379.

Rhodes, J. 2013. Using PeerWise to knowledge build and consolidate knowledge in nursing education. *Southern Institute of Technology Journal of Applied Research (SITJAR),* http://sitjar.sit.ac.nz.

Sadler, R. 1989. Formative assessment and the design of instructional systems, *Instructional Science* 18:119-144.

Taras, M. 2008. Issues of power and equity in two models of selfassessment, *Teaching in Higher Education*, 13(1): 81-92.