Integrate Confidence Ratings in Audience Response Systems in order to help Students to self-regulate their Learning Process

Lucas Braeschke[1], Iris Braun[2], Felix Kapp[3] and Tenshi Hara[4]
[1,2,3]*Technische Universität Dresden, Dresden, Germany*
[1]*Master student Computer Sciences* [2]*Chair of Computer Networks* [3]*Chair of Learning and Instruction*
[4]*University of Cooperative Education – State Study Academy Dresden, Dresden, Germany*
[1,2,3]{*forename.surname*}*@tu-dresden.de* [4]*tenshi.hara@ba-dresden.de*

Keywords: ARS, confidence, learning questions, SRL, AMCS

Abstract: Learning questions are an adequate method to check the knowledge of students in university courses. With the help of audience response systems (ARS), the lecturers can use learning questions during the active lecture to get immediate feedback about the knowledge base of the students. This information can help them to modify the content of the lecture or the kind of presentation of the knowledge. They can discuss the answers with the students and make the lecture more interactive. For the students it is helpful to regulate their learning strategy in the self-regulated learning process (SRL). For a deeper understanding of their own failures in answering the questions it is very important to think about their confidence while answering. Did they only guess or were they sure to have the right answer? In this paper we present an approach to integrate different kinds of confidence ratings in an ARS as well as our results from first user studies.

1 INTRODUCTION AND FOUNDATIONS

Audience Response Systems (ARS) such as AMCS[1] (Kapp et al., 2014) provide the opportunity to introduce more interactivity within university classes. Lecturers provide learning questions (LQ), polls and surveys before, during or after their classes and give students the chance to engage more actively in the learning process. According to models of Self-Regulated Learning (SRL) (e.g., (Zimmerman et al., 2000)), students face various demands during the learning process with regard to motivational, cognitive and meta-cognitive processes. By providing learning questions (LQ) with the help of ARS students are supported in maintaining their learning motivation. Working on learning questions helps them to process relevant information, in order to solve a LQ they are asked to execute cognitive processes such as recall and recognition of prior knowledge and integration of new concepts. Last but not least, LQs provide a unique opportunity to get feedback about their current state of knowledge and skills.

[1]`https://amcs.website`

The feedback obtained during the process of answering the LQ can be used for self-assessment. According to models of SRL, the adequate self-assessment is crucial for successful self-regulation during learning. A student preparing for an exam and repeating over and over content from one part of the knowledge domain while missing another ones in which they are not firm yet, might serve as an example of a failed regulation. Presenting LQs might help to self-assess and to adjust the self-concept of one's competence in a domain. In the next step this serves as the basis of decisions during the learning process such as study time allocation. To sum up, the successful self-regulated learner is both skilled in knowledge-acquisition and self-assessment as a basis for monitoring and other meta-cognitive processes. One way to improve the potential of ARS with regard to self-assessment is to integrate confidence ratings within the LQs.

Confidence ratings ask students to indicate how confident they are that their answer is correct. With regard to the support of the SRL process, the integration of confidence ratings within AMCS has two purposes: 1) to initiate a reflection process about students' state of knowledge/skills during the process of working on the LQ, and 2) to gather information which

can be helpful to regulate subsequent study behaviour at a later point of time (e.g., in restudy phases prior to exams). By storing and providing not only the information whether a LQ was answered correctly but also how certain the student was that the answer was correct, a more adequate picture of possible knowledge gaps or overconfidence can be drawn (e.g., in case of a correct guess, the student would notice that there is still a need to study the subject because of a rather low certainty which normally accompanies a guess).

2 RELATED WORK

We investigated different approaches of confidence acquisition in the context of ARS and learning questions (LQ) in e-Learning settings.

2.1 Confidence Ratings in Audience Response Systems

Starting point of our investigations was an analysis of ARS which are well-established in the market. We checked their features w.r.t. LQs and confidence acquisition (Kubica et al., 2019). From a list of fifty ARS, there is only one using confidence ratings to evaluate the students performance and self-assessment.

The *arsnova.click*[2] system is part of the ARSNova project and provides easy quizzes for short knowledge evaluations in schools. In *arsnova.click*, the lecturer can choose from seven types of questions, four of them are LQs with correct answers. The questions are added to a collection called 'quiz'. For each quiz, the lecturer can decide whether the confidence should be measured or not. After showing the questions and the possible answers, the students are asked 'How confident are you in your answer?' The confidence can be adjusted with a slider with five levels: 'guessed' (dark read), 'very uncertain' (light red), 'uncertain' (ochre), 'relatively certain' (light green), and 'absolutely certain' (dark green). After all students have answered the questions or when the answering time[3] has ended, an evaluation for all students is shown.

There is a list of all questions with a marker for correct and wrong answers, and a number of overall correct and wrong answers, as well as the confidence of the students. Within a click for each question, a ranking list of all correct answers sorted by shortest answering time is shown. This results presentation is part of *arsnova.click*'s gamification concept. The confidence information is not included in the list.

The problem with this kind of confidence acquisition is the limited answering time available to students during the quizzes. To be good in the ranking, students have to be as fast as possible. Thus, there is insufficient self-reflection time to choose the appropriate confidence level. We can expect that the measured confidence in this case is inaccurate. There are also no other interpretation hints for the measured confidence in relation to the performance of the students.

2.2 Certainty-based Marking

The 'London Agreed Protocol for Teaching' has been developed since 1994 at University College London (UCL) and is freely available for testing[4]. It is used primarily in medicine context, for instance in the 'Medical and Biomedical Students Self-Test' with more than 500 questions. The idea is to stimulate more careful thinking and learning than simple (right/wrong) marking, and to provide more reliable assessment (Gardner-Medwin, 2006).

It is based on Certainty-based Marking (CBM). After each answer the students can indicate their degree of certainty that the given answer will be marked as correct. Thus, a 3-point-scale with 'Unsure', 'Mid', and 'Sure' is used. Students may also choose 'No Idea'. Certainty levels 1 ('Unsure'), 2 ('Mid'), and 3 ('Sure') always give the students marks 1, 2, or 3 when they are correct. If they are wrong and unless they opted for C=1, they will loose marks, namely -2 at C=2, and -6 at C=3.

If the students are very unsure, they can avoid any risk of a penalty by choosing C=1. In contrast, if they are sure, they obviously get best marks with C=3. If they are wrong however, they will loose 6 points (twice the potential gain). The idea is to pay more attention if they make confident but wrong answers, to motivate them to re-think, reflect their strategies and learn more. Additionally, it is more fair: a thoughtful and confident correct answer deserves higher marks than a lucky guess.

2.3 Open Confidence-based Marking

The Open University[5] has developed an approach called 'Open confidence-based marking' (OCBM) as a variant of confidence based marking (CBM). Instead of asking students about their confidence or certainty after answering a question (as in CBM), in this approach only a question without any answer option is

[2] https://arsnova.click/

[3] The remaining answering time is visualised through a count-down clock.

[4] http://www.ucl.ac.uk/lapt/

[5] http://www.open.ac.uk

shown to the students when asking about their confidence. So they have to decide more generally if they have knowledge in the field or not.

The confidence acquisition uses – similar to UCL's approach – a three-level scale with the options 'Low', 'Medium', and 'High'. Besides this, there is an option to choose 'Give up' to not answer the question. For the evaluation, a list of all questions is shown with information about correctness of answer and given points (related to difficulty of the question) and the CBM mark. The confidence mark is calculated as follows: two points for high confidence, one point for medium confidence, and no points for low confidence. The result is added to the achieved mark for the question and then set in relation to the maximum mark that can be achieved.

3 Confidence Ratings in AMCS

As in all Audience Response Systems, polls represent the main functionality of AMCS (Kapp et al., 2014). Each poll consists of a few questions the students can answer. These polls are accessible for the whole course or under certain circumstances. For example, when a specific slide is shown during the lecture, making it a slide poll (SP). Other types of polls include 'global' course polls (CP) that are always accessible as well as lecture polls (LP) that can only be answered during a specific lecture[6]. All types of polls are further summarised in Table 1.

Table 1: Different types of polls in AMCS

Poll Type	Explanation
preparation poll (PP)	Active before a lecture. Can be utilised to prepare a certain topic or to inquire the students' previous knowledge.
lecture poll (LP)	Active during a lecture. Questions that should not be missed by attendees arriving late.
slide poll (SP)	Active when specific slide is shown. Can be used for quizzes during the lecture, in particular at a predefined time.
post processing poll (PPP)	Active after a lecture. Usable to check gained knowledge, for homework with automated feedback, etc.
course poll (CP)	Active during a course, especially during all of the courses lectures. Can be used for questions about the students degree programmes or their interests.

Polls in AMCS can consist of various types of questions, including single-choice survey (SC), multiple-choice survey (MC), single-best choice (SBC), multiple-best choice (MBC), free text, correct-assignment (CA) and scaled questions (SQ). In the cases of answering SBC, MBC and CA questions, immediate textual feedback is shown to students for their given answers. In contrast to other ARS, we implemented a two-step feedback algorithm. When answering wrongly on a first attempt, students will get a second attempt to redeem themselves. Regardless of the correctness of the second given answer, students will again get textual feedback that consists of an explanation on why exactly the given answers were wrong/correct. In the case of SBC and MBC questions, the answers are highlighted in colours corresponding to the correctness or incorrectness respectively.

We chose these LQs with correct answers for confidence ratings because the performance of the students can be measured directly with the tool. Thus, the students will receive feedback on their performance in relation to their own confidence ratings after finishing the polls. This can provide them with hints about the divergence of their self-assessment and their learning achievements. In subsection 3.3 we describe the design of different presentations of the results to the students as well as to the lecturers.

In literature, different options to integrate the confidence ratings into the workflow of answering LQs are discussed. The approaches differentiate in the time of asking about the confidence of giving the correct answer. When only the question text is shown without the possible answers, it is called *Open Confidence Rating*. If all possible choices are shown, students can describe their *certainty* to know the right answer. How we integrated these options of confidence ratings is described in the following sections.

3.1 Learning Questions with Certainty Rating

For the certainty rating, students will be asked after or while answering the question how confident they are that the chosen answer is right. In this case, they can see all choices to decide how certain they are. In related work we have analysed some other systems using certainty ratings. Most of them use a multi-level Likert scale with three or five items such as 'Low', 'Medium', and 'High'. In a first user study, we tried out different scales and representations for confidence ratings. The results are described in section 4. We decided to use the percentage of certainty to define the

[6]In our nomenclature, a course consists of one or multiple lectures. Each lecture must be part of a course.

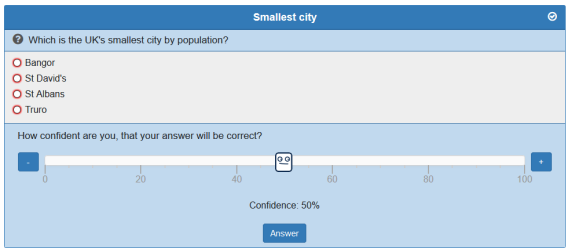

Figure 1: Learning questions with Certainty Rating

confidence more accurately. Students can share their confidence with a slider between 0 and 100% arranged under the question and possible answers. After selecting the answer(s) and the corresponding certainty in their answer, students can submit both results to AMCS (Figure 1).

3.2 Learning Questions with Open Confidence Rating

Another option we implemented in AMCS are LQs with open confidence rating. The students can only see questions without any answers to choose. Hence, they have to decide more generally if they have enough knowledge in this field to answer the question correctly (Figure 2). After selecting the confidence with a slider (between 0 and 100%), the possible answers are shown to the student in order for them to select therefrom. Afterwards, the confidence and answer selection are sent to the AMCS server.

3.3 Evaluation of Confidence Ratings

The main intent of the confidence rating is to show the students how their performance differs from their self-assessment. In a first step, these values for each question are shown to the students directly after answering the question. In a second step, the confidence ratings for all questions of a lecture or course should be aggregated so that the students can get feedback on their assessment accuracy.

As basis for the visualisation in the confidence graph (Figure 3), we used the average performance and confidence of all answered questions. The values for the performance can vary within the different LQ types.

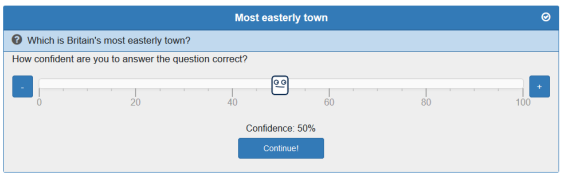

Figure 2: Learning questions with Open Confidence Rating

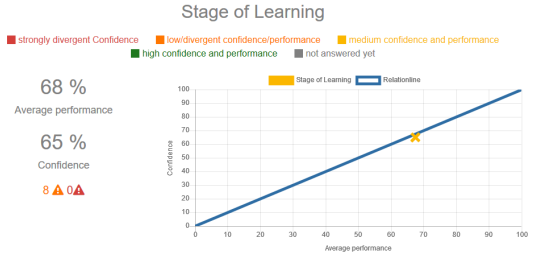

Figure 3: Confidence Rating Visualisation for Students

For MBC and CA questions they can achieve values *between* 0 and 100%, for SBC only 0 *or* 100% are possible. Nevertheless, the total value of summarising all questions is representative for the overall performance of the students. We only include the first answers of LQs in our aggregation, for the second answers the confidence can differ because of the given hints for first wrong answer and more time for re-thinking.

In the confidence graph (Figure 3) the relation between confidence and performance is visualised. The blue line marks the ideal self-assessment, the confidence is exactly as high as the performance. The cross marks the achieved values of a student, so it is easy to recognise if they are overconfident or too insecure. The overall goal of the students in the self-regulated learning process is to reach a high confidence and always choose the right answers. In literature this state is referred to as *informed*. In general, it is better when confidence and performance are in the same range as it shows that the student has a good self-concept of their competence in the domain and is aware of their impairments. On the other hand, it is critical when confidence and performance differ too much, so a student's picture of their own knowledge/skills is not accurate.

Confidence is mostly categorised in four levels (Curtis et al., 2013; Hunt, 2003; Burton, 2002):

- *informed* - high confidence and high performance (both over 75%),
- *partial knowledge* - medium confidence and medium performance (both between 42% and 75%),
- *uninformed* - low divergent confidence and low performance (both under 42%), and
- *misinformed* - strongly divergent confidence (difference between both more than 40%).

These levels should be visualised to the student in the overall results on the confidence graph as well as for the single answers in the questions list. For a better recognition value, we decided to use colours to mark the different levels of confidence. We based our decision for the colours on (Tak and Toet, 2014). The authors recommend to use a traffic light metaphor. Be-

cause we need four instead of three states, we decided to add orange between *red* and *yellow*. Beside this, we added grey as a marker for unanswered questions. A user study confirmed our decision (cf. section 4) and has shown how comprehensible this categorisation is.

3.4 Question Pool

After the lectures or at the end of semester, lecturers can choose a couple of LQs to be placed in a question pool. Students can use the question pool for exam preparations and repeat the included questions as often as they desire. They can choose a question subset or answer the entire set. However, with numerous questions, students are often overwhelmed and do not know which to choose. For example, they may ask themselves whether they should select all incorrectly answered questions or the ones where the confidence mostly differs from their performance?

The possible options for selecting questions from the question pool are shown in Figure 4. Students' first stimulus could be to repeat questions they answered incorrectly. However, another useful option could be to first answer questions where the self-assessment most obviously differs from the performance (*Divergent Confidence*). With the help of the aggregated confidence ratings, students are able to identify areas in which they are overconfident or too insecure. This additional information helps them to successfully regulate their learning process.

Students must not solely rely on the generated suggestions: they can choose questions directly from the question pool, too. In these individual lists, the questions are coloured differently in correspondence to the different confidence levels (Figure 5).

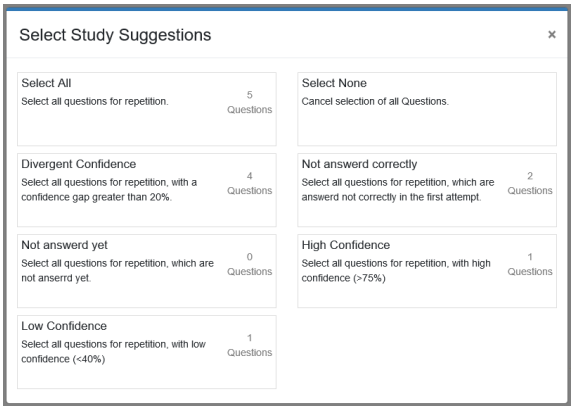

Figure 4: Suggestion of learning questions for repetition

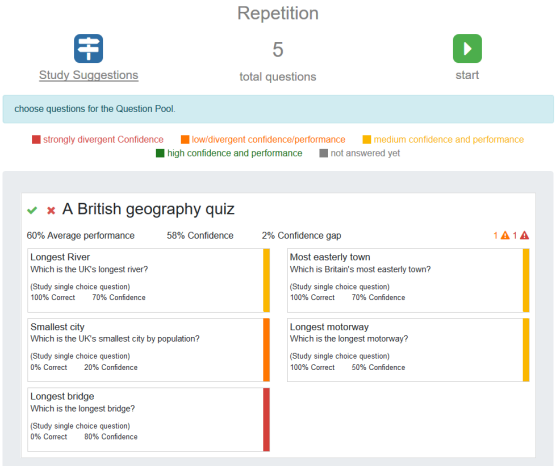

Figure 5: Question pool

4 USER STUDIES

For the design process of our AMCS prototype it is imperative that the developed systems fulfils a majority of user requirements, ideally all requirements. Hence, we conducted user studies in order to identify core expectations as well as necessities of the targeted users, namely students willing to revisit LQs to improve their understanding. The main focus was on visualisation and usability (e.g., the workflow).

We showed test users paper-based design prototypes, later let them use a prototypical implementation of our system, and used a paper questionnaire based on the User Experience Questionnaire[7] (UEQ). Additionally, we asked a few questions relating to the colour selection as well as symbol selection. Further, we asked if and how the workflow could be improved.

In a first study, we wanted to determine how to best visualise confidence in given answers. We suggested three designs as paper-based prototypes:

- *Oxford Cap* (a.k.a. square academic cap, graduate cap, or mortarboard):
  The more caps a student selects, he more confident they are.
- *Numerical Scale* (0 through 100%):
  The more confident a student is, the higher the percentage of confidence to select.
- *Emojis*
  With possible selections ranging from a Frowny to a Smiley with gradually changing facial expressions, students select the Emoji most suitable to describe their confidence in the given answer.

Within a controlled group setting, the five test users were then handed a UEQ in order to compare the three prototypes (Table 2).

[7] https://www.ueq-online.org

Table 2: User Experience Questionnaire scales

	Oxford Cap		**Scale**		**Emojis**	
	Avg	Var	Avg	Var	Avg	Var
Attractiveness	1.30	1.41	0.83	2.33	1.93	.43
Perspicuity	1.40	1.64	1.95	.39	1.70	1.23
Efficiency	1.85	.68	2.00	.59	1.35	1.27
Dependability	1.60	1.21	1.80	.73	1.30	.61
Stimulation	1.40	1.52	.65	1.58	1.70	.42
Novelty	.85	1.02	−.05	4.08	1.80	.98

The Oxford cap prototype averages between the other two. For any of the six dimensions of the UEQ, one of the other two prototypes achieves better scores. Based on additional feedback provided, the lack of originality combined with a lack of identification, namely association of graduation/finalisation of studies rather than continued studies, disqualifies the Oxford caps.

The numerical scale lacks on an aesthetic level, scoring roughly a point weaker compared to Emojis in the attractiveness dimension, and even scoring negatively in the novelty dimension. Further, stimulation falls far behind the other prototypes. We did ask test users about an exact scale (in 1% steps) as well as more coarse scales (in 10% and 20% steps). However, we advise caution with these results as scales are a known design element in our AMCS system, which might have had a deprecatory influence on the UEQ score of this prototype. Nevertheless, the high efficiency score shows that the fine-grained scale obviously allows students the most precise confidence rating.

Finally, the Emoji-based prototype convinces due to high attractiveness, dependability and novelty. Additionally, the variance is significantly lower compared to the other two prototypes. Taking these advantages into consideration, the obvious choice for the visualisation of confidence are Emojis – at least for AMCS.

We conducted a second evaluation after implementing an Emoji-based confidence slider into AMCS (Figure 1 and Figure 2). The goal was to identify input problems as well as to assess the workflow of answering questions *and* providing confidence ratings for these questions. For this, we prepared another UEQ which was handed out to fourteen students in an actual lecture[8]. The students used AMCS on their own devices[9] at the end of the lecture to answer four questions. With an averaged perspicuity score of 2.3, we strongly believe to have found a suitable input metaphor for levels of certainty in the context of AMCS and other audience response systems. Based on the results of the questionnaire as well as oral feedback, the Emoji-based input scale seems to be self-explanatory and was well received. From the user interviews we also received feedback to improve our slider prototype by combining the advantages of the Emoji-based solution with the numerical approach (as can be seen in the final prototype in Figure 2).

[8]A graduate lecture on Service and Cloud Computing.

[9]As we follow a bring-your-own-device policy, various device types as well as operating systems were used. The students experienced no technical difficulties.

Table 3: Resulting exam preparation data

Test User	**Percentage**	**Confidence**	**Representation**	**Used Learning Set**	**Utility**
1	45	55	6	1	7
2	69	65	6	4	7
3	57	53	5	1	6
4	50	69	6	1	7
5	65	74	6	4	6
6	77	82	6	4; 5	n/a
7	64	51	3	3; 5; 6	6
8	84	54	6	5	7
9	74	62	5	4	5

Once the confidence ratings are ascertained by the system, we utilise these to suggest sets of LQs for strongly adaptive quizzes during exam preparation. Hence, we wanted to find out if and how the suggested question sets are perceived as helpful by the students. Thus, we conducted a third evaluation in which we asked ten test users to answer twenty questions from varying areas of expertise[10] and provide confidence rating through the Emoji scales. Due to technical difficulties we did not received data from one test user. Afterwards, we asked the test users to simulate an exam preparation session by using the question pool section of AMCS. Table 3 show the resulting data by user, namely the percentage of correct answers (second column) and their provided average confidence rating (third column). Additionally, we asked the test users how well the system represented their perceived learning level (fourth column) and how they used the provided learning question sets (fifth column). Finally, we asked the test user whether they thought of the confidence rating and provided suggestions as being useful or not (sixth column).

Notably, the largest deviation from the actual score was 30%, which is 10% below the critical insecurity threshold. None of the test users totally agreed with the representation of the learning level; the average

[10]The AMCS test questionnaire consisted of questions from politics, geography and history of Germany.

score is 5.4 on a scale ranging from 'totally disagree' (1) to 'totally agree' (7). Finally, the test users were asked to selected a subset from the learning questions for targeted repetition. Possible subsets were:

1. all learning questions,
2. no repetition,
3. diverging assessment,
4. incorrectly answered,
5. low confidence,
6. high confidence, and
7. not yet answered.

As can be seen in the fifth row of Table 3, no test user opted for options 2 and 7. In a real scenario we would expect option 7 ('not yet answered') to be favourable as this would obviously be a bad choice to skip during exam preparation; not knowing what kind of questions one missed is fatal. Option 4 ('incorrectly answered' learning questions) was a favoured subset for repetitions. This is noteworthy as we expected 'diverging assessment' as well as 'low confidence' to be favourable as they would produce more individualised learning quizzes.

In the end, our test users agreed that the provided tool is a useful addition to AMCS. On a scale from 'totally disagree' (1) to 'totally agree' (7), an average score of 6.3 could be achieved (final row in Table 3).

5 CONCLUSION AND FUTURE WORK

Our study presents an approach to integrate confidence ratings within an existing ARS in order to support students and lecturers further while learning self-regulated in university settings. The theory based concept and first user studies with a focus on the design revealed that the additional feature is well accepted and is able to provide further information for reflection during answering the questions/polls and upcoming exam preparation.

In the future, we wish to further investigate the suitability of the different options of confidence ratings in different use-cases, the design of the feedback provided based on the confidence ratings and how the additional information from the ratings affects decisions during the learning process of students. With the prototype, we will be able to conduct experiments with students to find out how they use the suggestions of LQs for exam preparation and how the confidence information helps them to regulate their learning strategy in the SRL process.

Another open issue is to integrate the confidence results in the evaluation centre for lecturers we presented in (Braun et al., 2018). How can the ratings be aggregated and visualised to get a fast overview of the self-assessment of all students in the course?

REFERENCES

Braun, I., Hara, T., Kapp, F., Braeschke, L., and Schill, A. (2018). Technology-enhanced self-regulated learning: Assessment support through an evaluation centre. In *Proceedings of 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Symposia CELT: Computer Education & Learning Technologies*, Tokyo.

Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education*, 36(9):805–811.

Curtis, D. A., Lind, S. L., Boscardin, C. K., and Dellinges, M. (2013). Does student confidence on multiple-choice question assessments provide useful information? *Medical Education*, 47(6):578–584.

Gardner-Medwin, A. (2006). *Confidence-Based Marking - towards deeper learning and better exams*, pages 141–149. Routledge.

Hunt, D. P. (2003). The concept of knowledge and how to measure it. *Journal of Intellectual Capital*, 4(1):100–113.

Kapp, F., Braun, I., Körndle, H., and Schill, A. (2014). Metacognitive Support in University Lectures Provided via Mobile Devices - How to Help Students to Regulate Their Learning Process during a 90-minute Class. In *Proceedings of the 6th International Conference on Computer Supported Education (CSEdu 2014)*, pages 194–199. ScitePress.

Kubica, T., Braun, I., Hara, T., Kapp, F., and Schill, A. (2019). Choosing the appropriate audience response system in different use-cases. In *Proceedings of the 10th International Conference on Education, Training and Informatics (ICETI 2019)*, Orlando. (provisionally published as of 13 March 2019).

Tak, S. and Toet, A. (2014). Color and Uncertainty: It is not always Black and White. In Elmqvist, N., Hlawitschka, M., and Kennedy, J., editors, *EuroVis - Short Papers*. The Eurographics Association.

Zimmerman, B. J., Boekarts, M., Pintrich, P., and Zeidner, M. (2000). Attaining self-regulation: a social cognitive perspective. *Handbook of self-regulation*, 13.

URLs in this paper were last accessed on 13 March 2019.