Using the Students' Levels of Preparation and Confidence as Feedback Information in Quiz-Based Learning Activities

Pantelis M. Papadopoulos[1], Antonis Natsis[1], and Nikolaus Obwegeser[2]
[1] Centre for Teaching Development and Digital Media, Aarhus University, Aarhus, Denmark
[2] Department of Management, Aarhus University, Aarhus, Denmark
pmpapad@tdm.au.dk, anatsis@tdm.au.dk, nikolaus@mgmt.au.dk

Abstract. This paper examines ways to enrich the feedback information students receive in closed-type quiz activities that include a revision phase (i.e., students are allowed to change their initial answers after they receive information from their peers, teacher, or system). Typically, in such activities, the information students receive is based on the percentage of students under each possible question choice. The conducted study analyzes the potential of two additional variables, namely the students' level of preparation and confidence. Both variables are self-reported and, therefore, subjective. During the Fall semester 2016, 91 sophomore students enrolled in an Information Systems course participated in the study for five weeks. The activity was taking place during the first 20 minutes of each class. Students had to go through three phases and (a) answer a multiple-choice quiz with 8 questions and 4 options for each question, (b) receive feedback based on the whole classroom population, and (c) see the correct answer and discuss them with the teacher in the lecture that follows. The students were randomly grouped into four conditions, based on the feedback they received. The control group only received information on the percentage of students that selected each choice, the Confidence group received feedback on the percentage and the average level of confidence of students that selected each choice, the Preparation group received feedback on the percentage and the average level of preparation of students that selected each choice, and finally the Both group received feedback on the percentage and both the average level of confidence and preparation of students that selected each choice. Result analysis showed that in the most challenging questions (i.e., the ones where students' answers were diverging) the students in the Confidence, Preparation, and Both groups significantly outperformed the students in the Control group. In addition, both confidence and preparation variables were significantly correlated to students' performance during the initial phase, suggesting that students were accurate and sincere in describing their preparation and confidence levels. This paper is an extended version of [1], presented at the 9th International Conference on Computer Supported Education.

Keywords: Feedback, Group Awareness, Formative Assessment, Quiz, Confidence, Preparation.

1 Introduction

The multiple-choice quiz is a versatile tool that can be used in many different educational contexts for a range of learning purposes. For example, a quiz activity can be used as an assessment tool by the teacher or can be offered as a self-assessment tool to the students. It can be mandatory or optional, taken once or several times. It can be held at any point during a class, or even outside the classroom. A quiz activity in the beginning of the class could provide a useful reference on students' knowledge and this reference can be later used by the teacher in identifying and addressing misconceptions that could affect the lectures to come. On the other hand, short quiz activities during the lecture can maintain students' engagement and attention while assuring the teacher that the lesson taught is indeed understood by the audience [2]. Finally, a quiz near the end of the lecture would allow students to review the day's class and increase their retention.

While pen-and-paper quiz activities are common and easily administered, computer-supported quiz activities significantly increase the learning benefits, by providing personalized and customizable information in real-time [3]. Timely feedback and the ability to repeat the quiz activity multiple times may offer additional opportunities for self-reflection and self-assessment to the students [4][5][6].

The interested teacher can find a range of freely available tools that support the design and implementation of quiz activities in different educational contexts. Even though the functionalities offered by these tools may differ significantly, the underline remains the same, specifically to ask the student to find the correct answer(s) out of a set of possible answers to a question. A typical example of such a tool is Socrative[1]. Apart from offering an easy way to use menus to create and use quizzes, Socrative also supports tracking students' progress through a series of quiz activities, allowing the teacher to assess the progress made during a semester. Following a different approach, PeerWise[2] is based on student-generated questions. In other words, the students have to author interesting, challenging, and well-phrased questions. At the same time, they can answer and review the quality and difficulty of questions submitted by their peers. PeerWise is a widely used tool, and part of its success is arguably based on the fact that it incorporates gamification in the form of leaderboards and badges [7]. Similarly, Kahoot[3] offers a variety of quiz types, apart from multiple-choice questions, such as fill-in-the-blanks, matching words, etc. The design of Kahoot is significantly based on game elements, enhancing also competition between the students.

The present study aims at exploring the potential of multiple-choice quiz activities in formative assessment, in conjunction with the inclusion of both objective and subjective metrics in the feedback information the quiz tool provides to the students. Finally, this paper is an extended version of [1], presented at the 9th International Conference on Computer Supported Education.

---

1 http://www.socrative.com/

2 https://peerwise.cs.auckland.ac.nz/

3 http://getkahoot.com

2 Background

2.1 Quiz and Group Awareness

One element that Socrative, Kahoot, and Peerwise have in common is that students can receive feedback information from both the teacher and fellow students. For example, feedback coming from the teacher can appear as pre-entered hints pointing to the correct choice, or as explanations on why a choice may be wrong/correct. Feedback based on fellow students usually presents information about peers' activity, including average scores of a group of people (e.g., the whole class), distribution of class population into the different choices, etc. Providing feedback information based on peers' activity allows the students to compare their own knowledge to their peers'. According to Bodemer [8], these comparisons support group awareness and are beneficial to students' learning. Yet, the student-based feedback the user receives in all three quiz tools provides information solely on the percentage of students that selected each question choice. We argue that although the percentage metric is widely used, easy to understand, and to a great degree useful for the students, it provides information only on a surface level, without being able to include qualitative information on the groups of people that selected each choice. Metrics that could further describe relevant characteristics of fellow peers could be essential for students, in terms of comparison and self-assessment.

Group awareness has already been identified as an important design aspect for educational technology tools, with several studies describing the learning benefits that emerge when students are able to compare and analyze peer activity (e.g., [9][10], for a review). One can find two different definitions of the term in the literature. Cognitive group awareness refers to information that allows the students to understand the level of knowledge their peers have attained, while social group awareness refers to information that depicts peers' activity in the group [11]. In the context of the current study, the focus is on cognitive group awareness with the feedback metrics used aiming at providing an aggregated picture of the group knowledge, with the term "group" referring to the whole class population.

The present study combines objective and subjective metrics in an effort to present peer characteristics and support group awareness. Thus, in addition to the objective percentage metric, the feedback that the students receive includes subjective, self-reported, information on peers' level of confidence and preparation. Previous studies highlight the learning gains that such a combination of objective and subjective metrics offers to the students (e.g., [12][13]). For example, previous research shows that asking students to denote how confident they feel about their answers to a quiz can significantly improve their metacognition [5].

We maintain that the inclusion of both objective and subjective metrics in the feedback information provided to the students could offer a better picture on peers' knowledge. We expect that this, in turn, could increase knowledge group awareness and, eventually, students' performance.

2.2 Student Learning and Engagement

Literature abounds with research evidence on the ways the students' engagement and performance are positively affected by quiz activities. For example, Méndez-Coca and Slisko [14] reported that the use of Socrative made students more engaged in the learning activity. In addition, students explicitly expressed positive attitudes towards the approach, underlining that Socrative increased their motivation to actively participate in the class and enhanced their communication with their fellow students. The beneficial impact of quizzes on peer interaction could be linked to the learning gains students reap by externalizing their knowledge. Of course, the process of answering closed-type, multiple-choice, questions does not provide the same opportunities for knowledge externalization as a written task that engages the students into formulating and structuring valid arguments. Nevertheless, the process of making one's opinions explicit can still offer the basis for meaningful peer interaction [15]. To promote dialogue between students, Méndez-Coca & Slisko [14] formed groups of students with different opinions, thus creating the opportunities for meaningful discourse.

Arguably, one of the main advantages of quiz activities is the overall positive attitudes students show towards them. In their study, DiBattista, Mitterer, and Gosse [16] focused on students' attitudes by comparing two multiple-choice testing settings. In the first one, students were receiving immediate system feedback, while in the second the multiple-choice testing was conducted with pen-and-paper. Research data provided overwhelming evidence that students preferred the first setting, even though further data analysis showed that the performance and personal characteristics of students in the two settings were comparable.

Another important advantage of quiz activities is their ability to incorporate game elements in their design. It is common for quiz activities that are administered inside the classroom to showcase aspects of gamification [17]. Typical examples of gamification are complex grading systems (e.g., positive/negative/weighted grades), tier or karma points (that can be used to unlock functionalities or other rewards), badges (that denote significant achievements, such as a streak of correct answers), leaderboard (that show the high-achievers in a group), and so on. It is important to note that these game elements are not linked to the learning process itself. In other words, receiving a badge does not provide additional scaffolding to the student, nor changes the studying conditions. Nevertheless, the positive impact of game elements on student engagement has been reported multiple times in the literature (e.g., [7][18]). It is also worth noting that in order to retain students' engagement in a quiz, the design should extended beyond game elements, since, as Wang [6] suggested, the engagement that is based on the novelty effect of superficial awards should be expected to decrease over time. A rigorous instructional design will discourage students from "gaming the system" [19] or disengaging because of the injection of unproductive peer competition in the learning process [20].

2.3 Study Motivation

The motivation for this study was our effort to improve the feedback information the students receive in closed-type, multiple-choice quizzes that allow for a revision phase. In this study, the feedback information based on the objective percentage metrics is enriched with two self-reported (thus, subjective) metrics that could paint a more detailed picture on the class knowledge. The preparation metric shows students' opinions on how prepared they feel just *before* they participate in the quiz. The confidence metric on the other hand shows how certain students are that they have selected the correct choice *after* answering each question. Thus, the preparation metric is based on a single question answered before the quiz is administered, while the confidence metric is the average of the confidence scores submitted by the student after answering each of the eight questions of the weekly quiz.

It is important to note here that this study is a part of a larger research effort that explores the potential of closed-type formative assessment tools in increasing student engagement and performance in different educational contexts. A necessary requirement for the successful implementation of such an approach is to keep the overhead for the teacher at a minimum level. Furthermore, another aspect of the long-term research effort is to evaluate how a series of quiz activities could eventually provide enough information for the compilation of knowledge profiles for the students and how these profiles could positively affect knowledge group awareness in collaborative learning activities (e.g., group assignments). However, the discussion of the long-term research plan extends outside the scope of the present paper.

3 Method

3.1 Participants and Domain

The subject domain of our study was the "Business Development with Information Systems – BDIS" course. Usually, undergraduate students during their second year (third semester) of studies are enrolled in BDIS. The course offers five credit units (i.e., ECTS) to participants and is part of the "Bachelor's Degree Programme in Economics and Business Administration" in the Department of Management. The course is taught in parallel in both Danish and English. For the purpose of this study, we focused only on the English version. The intended learning objectives of the course are to engage students in the analysis, evaluation, and application of models based on Information Systems, Decision Making, and Business Management domains in a challenging case-study that usually lasts throughout the academic semester. As a common practice in the university the study was conducted in, the lecture material (i.e., relevant literature, lecture notes, external links) was available online to the students, a week prior the respective lecture. Studying the course material beforehand is not mandatory, but encouraged, with students spending time preparing for the upcoming lecture at different degrees. The assessment process of the course includes a case-study group assignment with a written case report as an outcome, and individual final

oral examination in which students are required to elaborate on the case analysis and conceptual knowledge of the domain.

The course duration is 14 weeks and includes weekly 2-hour lectures in an auditorium. The average number of enrolled students each year is approximately 180. The actual number of students present during lectures, though, fluctuates significantly each week, since attendance is not mandatory. Even though all students were invited to participate in the study, we only used data collected from students that attended all lectures during the study period. Students that were present in only some of the lecture during the study period were still able to participate, but their data were excluded from the analysis. At the end, the findings of the study were based on a sample of 91 sophomore students. The students were distributed randomly into one of four treatment groups (see next Section) by the system at the time of their first login. Student distribution into the four groups was:

- Control: 27 students;
- Confidence: 22 students;
- Preparation: 22 students;
- Both: 20 students.

Students' participation was voluntary, and the activity was not part of students' formal assessment in the course.

3.2 The SAGA System

The study was conducted on the "Self-Assessment/Group Awareness – SAGA" online quiz system that was designed and developed by the research team of this study. SAGA has a long-term scope, being the platform we are going to use to explore the different aspects of quiz activities in a series of studies. Despite the variety of available online quiz tools, no system was able to support the research design of the current study. By creating our own system we were able to tailor its functionality to our research goals and achieve higher degrees of flexibility and customization, using different feedback metrics for different groups and monitoring student activity.

Students in SAGA start their activity by logging in and answering a question regarding the amount of time they spend during the week studying the material of the lecture they are currently attending. The preparation question was stated as follows:

> *Some of the teaching material for today's class became available during the last week.*
> *Using a scale from '1: Not at all' to '5: I have read it thoroughly', how much time did you spent preparing for today's class?*

Students' answer in this question was the only entry point for the preparation metric that was used later in the revision phase. It was not possible to test whether students' answers in the preparation question were accurate. Nevertheless, this metric could provide an estimate on how students self-assess their preparation. In addition,

this metric was used later to analyze whether the level of preparation, as self-reported by the students, was correlated with their actual performance in the quiz.

Next, students moved on to the initial part of the quiz activity, answering a series of eight multiple-choice questions. Each question and the accompanying four choices had been previously inserted in the system by the course instructor. In all questions, there was only one correct answer.

For each question, students had to select one of the available choices and denote their level of confidence, before they would be allowed to continue to the next question. The confidence question appearing under each question was stated as follows:

> *Using a scale from '1: Not at all' to '5: Very confident', note how confident you are that you have selected the correct answer.*

Similarly to preparation, the question regarding confidence was self-reported and as such we were not able to assess its accuracy. However, we were able to analyze whether students' self-assessment was correlated to their performance. The confidence metric was calculated for each student as the mean value of the 8 answers the student provided in the confidence question in the eight quiz questions.

Students had to answer all 8 questions of the quiz to move to the next phase, revision. During the revision phase, the students were able to browse once again all the questions and decide whether or not to revise their initial answers. To assist students in their decisions, the system was providing feedback based on the class activity during the initial phase. The feedback information was different for each group in the study, compiled with a different combination of the percentage, preparation, and confidence metrics (Figure 1):

- Control: the percentage of student in the class that selected each option;
- Confidence: the percentage and the average confidence score of students that selected each option;
- Preparation: the percentage and the average preparation score of students that selected each option;
- Both: the percentage, the average confidence, and the average preparation scores of students that selected each option.

Each metric was calculated against the whole classroom population. Thus, the value of a feedback metric that was appearing in different groups was the same. During revision, the students were also able to change their initial answers to the respective confidence questions. Thus, even in the case where students choose not to revise an initial answer, they can still change their self-reported confidence level. This could be useful, for example, in cases where feedback reinforced (or challenged) a student's perspective on which the correct answer was. After finishing the revision phase, the students were able to see their scores and the correct answers.

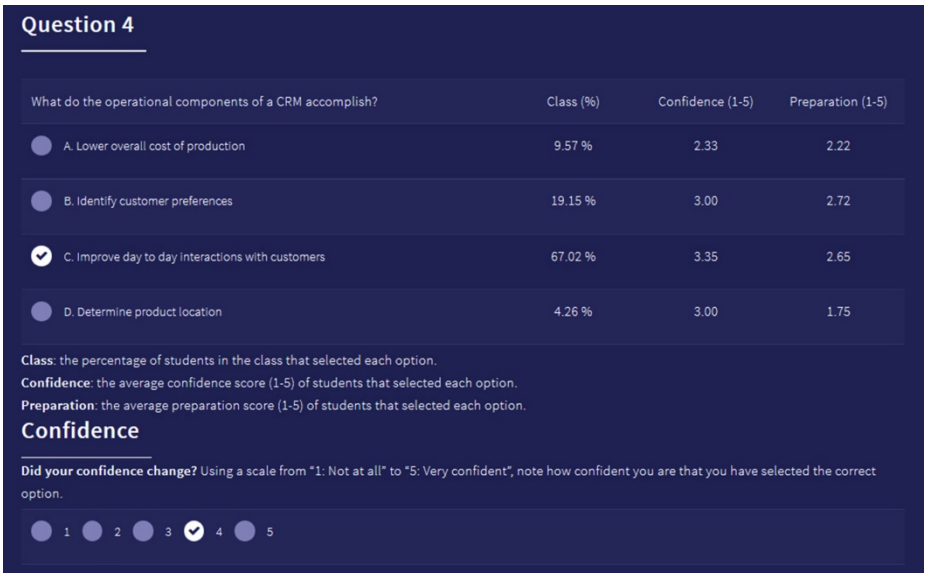

Fig. 1. Screenshot of the SAGA system during the revision phase for students in the Both group – all metrics (percentage, confidence, and preparation) are available [1].

This was the end of the quiz activity. Next, the teacher was able to decide whether to discuss the correct answers with the students right away or revisit them at a later stage during the lecture that was about to start. It is worth noting at this point that throughout the quiz activity, the teacher was able to monitor students' activity in real time and initiate the next phase of the process. In addition, the teacher had access to an aggregated view of the class performance in each phase and could share this view while discussing the correct answers after the quiz had finished.

3.3 Process and Study Conditions

The study took place during the Fall semester of 2016 and lasted five weeks in total, with the first four weeks being used for typical weekly quizzes related to the respective lectures, and the last week used for a retention quiz and the survey questionnaire of the whole activity.

In a typical weekly quiz, the students had to login to the SAGA tool and go through the three phases of the process, namely provide initial answers to the eight multiple-choice questions, revise their initial answers based on the feedback they receive, see the correct answers and discuss them with the teacher. We informed participating students of the research scope of the activity, clarifying that their placement into a study group is random, that their performance will not have an impact on their assessment in the course (this was also guaranteed by the course formal regulations), and that they may receive different information by the SAGA tool during the activity.

The weekly quiz was administered during the first 20 minutes of the 2-hour lecture session. Keeping the quiz activity short was an important requirement in our design,

since the goal was not to disrupt the teacher's lecture plan. In addition, the quiz activity should support students' learning and engagement in the course, without requiring significant additional effort or resources from the teacher (e.g., preparation and administration of the quiz). The 20 minutes used for the quiz activity were allocated to the three phases of the process as such: ten minutes for the initial phase, five minutes to reflect on the received feedback and perform the revision, and five minutes to check the correct answers and discuss them with the teacher. All students were at the same phase at any point during the activity, while the next phase could be activated earlier than planned, in case all students had finished the current activity.

The retention quiz took place during the fifth week of the activity and was not previously announced to the students. On the contrary, the students were expecting the typical weekly quiz for the lecture that was about to start. The retention quiz included a selection of 16 questions that had been previously used in the weekly quizzes. The quiz was followed by a questionnaire that included both open and closed-type questions, asking students to share their opinions regarding the usefulness of the different feedback metrics, the impact of the quiz activities on the amount of effort they put in preparing for the lectures, and the overall improvement of the SAGA tool and the quiz process. Lastly, the students were asked to fill in the Scale for Social Comparison Orientation (SSCO) ([21][22]). The SSCO instrument contains 11 statements focused on how often students compare themselves with others. Comparison could refer to students' feelings, opinions, abilities, etc. and it is not characterized as "good" or "bad" by the instrument. Since the research goal during the fifth week of the activity was to measure students' retention, opinions, and SSCO profile, the quiz phase did not have the revision phase. Instead, students submitted their initial answers to the retention quiz, then filled in the survey and the SSCO instrument, and saw their scores and the correct answers at the end of the activity.

The students participated in every aspect of the activity individually. Participation was anonymous and no personal information was recorded on maintained by SAGA or the research team. Finally, the study conditions were identical for all students in the study, apart from the different feedback information the four groups received during the revision phase.

3.4 Research Design

The study followed a between-subjects 2x2 factorial design with the combination of feedback metrics in each group being the independent variable of the study (Table 1).

Table 1. Levels of independent variables and student groups [1].

		Confidence Feedback	
		No	***Yes***
Preparation Feedback	***No***	Control	Confidence
	Yes	Preparation	Both

The dependent variables of the study were students' performance on the quizzes (i.e., initial and revision scores for weekly, and initial score for retention) and their responses on the survey questionnaire and on the SSCO instrument.

3.5 Data Collection and Analysis

For all the statistical tests performed in the study, the level of significance was chosen at 0.05. For the analysis of students' performance in the quizzes (weekly and retention) and their responses in the SSCO instrument, we used parametric tests. On the contrary, non-parametric tests were used for the analysis of students' responses in the survey questionnaire. The reason for this was that test assumption analysis showed that the normal distribution criterion was violated for that dataset.

Each week, we compared students' performance in the four groups, analyzing both the dataset of each week and the aggregated dataset of all weeks up to that point. Thus, at the end of the fourth week, our dataset included information on all 32 questions used. The retention quiz dataset was analyzed separately from the weekly quizzes, since the research goals and the study design was different. In addition, at the end of the fourth week, another analysis took place, taking into account only a subset of the total 32 weekly questions. This subset included only questions that students found challenging. In other words, we were not able to know at the beginning of the study which of the questions submitted by the teacher would be difficult for the students. As expected, there were some questions in which the vast majority of the students got the correct answer even during the initial phase of the quiz. This meant that the additional feedback provided in the revision phase (i.e., confidence and preparation metrics) had little effect on students' decision, since the percentage metric was showing clearly that there was a strong consensus in the classroom on the correct answer. So, the subset analysis focused only on the questions in which the percentage information alone was not able to direct the students to the correct answer. To phrase it differently, the challenging questions subset included the questions where there was a somewhat balanced distribution of students into the four question choices.

We mentioned earlier that the percentage metric is widely used in quiz tools. We expected that the students in this study would rely primarily on this percentage information before checking how confident or prepared the students under each question choice were. Thus, we argue that the impact of the two newly introduced metrics could be visible in cases in which the percentage information alone could not "clearly" point to the correct choice.

To maintain a level of objectivity, we agreed on a definition of what constitutes a "clear" case and we examined our dataset after the fourth week to identify the challenging questions. Our definition for "clear" cases for the percentage metric included two mandatory conditions:

- The correct choice was selected by at least 50% of the students;
- The correct choice had a least 20 points difference from the second most selected choice.

These two conditions ensured that the majority of the students selected the correct choice during the initial phase and that there was no clear alternative. For example, if in a question the choice A was selected by 55% of the students and choice B was selected by 40% of the students, then the question would be identified as challenging, since student population appeared split. As such, the subset included questions in which the percentage metric was either pointing to a wrong choice (e.g., the most favorite choice was wrong) or the distribution of selected answers was ambiguous (e.g., students were divided between two or more choices).

Using this definition, we identified 13 challenging questions. Since we did not anticipate which of the questions might be challenging during the design phase, the distribution of these questions in the four weekly quizzes was unbalanced: four challenging questions in the first week, five in the second, one from the third, and three from the fourth. Eventually, the impact of the confidence and preparation metrics on student performance was analyzed using this subset.

We used these 13 challenging questions and we added three more that were close to be categorized as challenging to compile the list of the 16 questions used in the retention quiz. Adding these three questions allowed us to create a longer and somewhat balanced quiz that included four questions from the first week, five from the second, three from the third, and four from the fourth.

4 Results

4.1 Weekly Quizzes

Table 2 presents students' performance in the four weekly quizzes.

Table 2. Students' performance in the four weekly quizzes [1].

	Control			Confidence			Preparation			Both		
Week 1	M	*SD*	n	M	*SD*	n	M	*SD*	n	M	*SD*	n
Initial	4.58	*(1.53)*	27	4.48	*(1.19)*	22	4.15	*(2.34)*	22	4.85	*(1.73)*	20
Revision	6.25	*(1.32)*	27	6.40	*(1.29)*	22	5.62	*(2.22)*	22	6.46	*(1.39)*	20
Week 2												
Initial	3.50	*(1.27)*	27	3.64	*(1.17)*	22	4.13	*(1.14)*	22	4.06	*(1.43)*	20
Revision	4.35	*(0.87)*	27	4.01	*(1.34)*	22	4.69	*(0.94)*	22	4.50	*(1.04)*	20
Week 3												
Initial	5.52	*(1.64)*	27	5.19	*(1.74)*	22	5.43	*(1.59)*	22	5.09	*(1.63)*	20
Revision	6.87	*(1.10)*	27	7.08	*(1.38)*	22	7.00	*(1.00)*	22	7.05	*(1.25)*	20
Week 4												
Initial	3.73	*(1.98)*	27	3.52	*(1.37)*	22	4.14	*(1.83)*	22	4.05	*(1.43)*	20
Revision	5.76	*(1.04)*	27	5.26	*(1.05)*	22	6.14	*(1.15)*	22	6.06	*(1.21)*	20

It is obvious that there is no apparent pattern in students' performance. On the contrary, the differences observed on the initial scores of a group in different weeks indicate different levels of difficulty for the weekly test, or different levels of preparation.

Despite our effort to have quizzes of similar difficulty each week, it is possible that students' understanding of the topics covered in the associated reading material available to them during the week differed. Similarly, students' preparation level refers to their perceived readiness to answer questions on the lecture's topics. The question about preparation refers to the amount of effort spent by the students, but not on whether this effort was spent effectively on understanding the learning material.

According to the two-way analysis of variance (two-way ANOVA) we performed, the four groups were comparable in the initial phase of the quiz in all four weekly quizzes ($p > 0.05$). Similarly, the results of the two-way analysis of covariance (two-way ANCOVA), when using students' initial phase scores as covariate, showed that the four groups also performed the same in the revision phase, over all four weeks ($p > 0.05$). Regarding revision, the paired-samples t-test showed that all student groups improved significantly from the initial to the revision phase, in all four quizzes.

4.2 Subset Performance

Table 3 presents students' performance in the subset of the 13 challenging questions.

Table 3. Students' performance in the 13 challenging questions.

	Control			Confidence			Preparation			Both		
Challenging	M	*SD*	n	M	*SD*	n	M	*SD*	n	M	*SD*	n
Initial	4.44	*(4.34)*	27	3.82	*(3.59)*	22	5.27	*(3.98)*	22	4.40	*(2.87)*	20
Revision	4.00	*(4.29)*	27	4.90	*(3.00)*	22	6.36	*(4.22)*	22	6.60	*(3.73)*	20

*$p < 0.05$

Performing a question-by-question qualitative analysis, we found out that students relied first and foremost on the percentage metric. Specifically, the percentage of the most popular choice during the initial phase of the quiz was increasing during the revision phase. Notably, this was happening even in cases in which the most popular choice during the initial phase was wrong. By using our definition of "clear" cases, we found out that the percentage metric was pointing to a specific correct choice in 24 out of the 32 questions available in the four weekly quizzes. However, in five cases, the percentage metric was pointing at a wrong choice. Thus, the eight questions in which students' distribution in the four question choices was split, and the five questions in which the percentage metric was pointing at a wrong choice, formed the subset of the 13 challenging questions we mentioned earlier.

We argue that the additional feedback metrics would be more useful to the students for this subset of questions. By transferring the "clear" case definition to the confidence and preparation metrics, we discovered that the confidence metric was pointing at the correct choice in eight of the challenging questions, while the preparation metric was doing the same for seven of the challenging questions.

We analyzed groups' improvement from the initial to the revision phase of the study, using paired-samples t-test. Results showed that only the Control group did not manage to improve its performance. On the contrary, Confidence ($t[21] = 2.324$,

$p = 0.030$, $d = 0.720$), Preparation ($t[24] = 2.027$, $p = 0.046$, $d = 0.630$), and Both ($t[19] = 2.979$, $p = 0.008$, $d = 0.970$) groups improved significantly. Two-way ANCOVA, using initial phase scores as a covariate, showed a significant main effect for the confidence ($F(1,86) = 4.115$, $p = 0.046$, $\eta^2 = 0.046$) and preparation ($F(1,86) = 7.153$, $p = 0.009$, $\eta^2 = 0.077$) metrics, but not for their interaction ($p > 0.05$).

4.3 Retention Test

Table 4 presents students' performance in the retention quiz. The results of the two-way ANOVA revealed no significant difference in performance of the four groups in the retention test ($p > 0.05$).

Table 4. Students' performance in the retention quiz.

	Control			Confidence			Preparation			Both		
Retention	M	*SD*	n	M	*SD*	n	M	*SD*	n	M	*SD*	n
Initial	10.00	*(3.23)*	27	10.86	*(2.14)*	22	10.68	*(3.24)*	22	10.80	*(3.2)*	20

4.4 Student Opinions and Behavior

Analysis of the internal consistency of the SSCO instrument used to record students' social comparison treats showed that the reliability of the instrument was mediocre (Cronbach's $\alpha = 0.635$) and lower than it is usually reported in the literature. Students' SSCO score in all groups were comparable. In addition, the SSCO score was not correlated to the total number of revisions performed, the total number of correct/wrong revisions performed, the initial/revised performance, nor the initial/revised confidence values.

Table 5 shows students' opinions on the most important questions of the survey questionnaire administered at the end of the study. The results of the Kruskal-Wallis and Mann-Whitney tests revealed no significant differences between the four groups ($p > 0.05$). Corroborating our question-by-question analysis, students stated that the percentage metric was the most useful metric for them during the revision phase of the study ($M = 3.62$, $SD = 1.01$). The second most useful feedback metric, according to students, was the level of confidence ($M = 3.32$, $SD = 1.20$), and the third was the level of preparation ($M = 2.64$, $SD = 1.43$).

We also asked students to suggest additional feedback metrics that could have been useful for them. Students mentioned past performance ($M = 3.20$, $SD = 1.14$), argumentation ($M = 3.15$, $SD = 1.15$), and peer communication ($M = 2.87$, $SD = 1.19$). The level of confidence ($M = 3.35$, $SD = 1.11$) and preparation ($M = 3.15$, $SD = 1.19$) were also included in the list, by students that did not have access to these metrics during the current study. Regarding the three new metrics suggested by the students, past performance refers to the average score a student received in previous weekly quizzes, argumentation refers to students' ability to read/write anonymous justifications for the choices, and peer communication refers to chatting anonymously with peers online for a brief period of time.

Table 5. Students‘ responses in the questionnaire. Scale – 1: Not at all; 5: Very much [1].

Control *n = 27*		Confidence *n = 22*		Preparation *n = 22*		Both *n = 20*		Total *n = 91*	
M	*SD*	M	*SD*	M	*SD*	M	*SD*	M	*SD*
Q1. Has the quiz made you spend more time preparing during the week for each lecture?									
2.17	*(1.04)*	2.90	*(1.17)*	2.68	*(1.39)*	2.17	*(1.37)*	2.49	*(1.28)*
Q2. Do you find the percentage values you see useful in choosing your final responses?									
3.72	*(0.89)*	3.43	*(0.87)*	3.73	*(1.12)*	3.61	*(1.15)*	3.62	*(1.01)*
Q3. Do find the confidence values you see useful in choosing your final responses?									
-	-	3.33	*(1.19)*	-	-	3.30	*(1.25)*	3.32	*(1.21)*
Q4. Do find the preparation values you see useful in choosing your final responses?									
-	-	-	-	2.59	*(1.53)*	2.70	*(1.39)*	2.64	*(1.44)*
Q5. How useful do you think the confidence level (confidence level of fellow students that selected each option) would be for you in choosing your final answers?									
3.61	*(0.85)*	-	-	3.14	*(1.28)*	-	-	3.35	*(1.12)*
Q6. How useful do you think the preparation level (average preparation level of fellow students that selected each option) would be for you in choosing your final answers?									
3.28	*(1.22)*	3.05	*(1.20)*	-	-	-	-	3.15	*(1.20)*
Q7. How useful do you think the past performance (average past scores – based on previous weeks – of fellow students that selected each option) would be for you in choosing your final answers?									
3.83	*(0.85)*	2.95	*(1.28)*	3.14	*(1.28)*	2.00	*(0.95)*	3.20	*(1.14)*
Q8. How useful do you think argumentation (a short argument for each option, written by a fellow student – anonymity remains) would be for you in choosing your final answers?									
2.72	*(1.36)*	3.05	*(0.97)*	3.18	*(1.25)*	3.22	*(1.04)*	3.06	*(1.15)*
Q9. How useful do you think peer communication (opportunity to briefly text anonymously with fellow students) would be for you in choosing your final answers?									
2.78	*(1.06)*	2.95	*(1.16)*	2.95	*(1.49)*	2.78	*(1.08)*	2.87	*(1.20)*

Students were asked to estimate whether their participation in the weekly quizzes motivated them to increase the amount of time they spent preparing each week for the upcoming lecture (Q1). Result analysis showed no significant different between the four groups ($p > 0.05$), with students having diverging opinions ($M = 2.49$, $SD = 1.28$). Nevertheless, students' answers in the preparation question in the four weekly quizzes and the retention test revealed a significant increase of the preparation time during the study. The results of the analysis of variance, with repeated measures with a Greenhouse-Geisser correction (sphericity assumption was violated), showed

that the mean value for the preparation level were statistically significantly different ($F(3.306, 247.966) = 44.128$, $p = 0.00$, $\eta^2 = 0.370$). Figure 2 presents the average preparation score for 91 participating students in each week of the study.

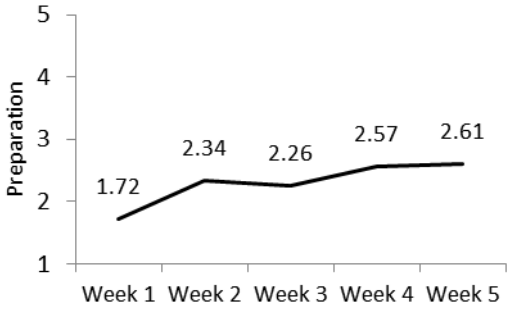

Fig. 2. Student preparation [1].

As expected, students' confidence increased significantly from the initial to the revision phase in all four weeks, for all four study groups ($p < 0.05$). In addition, correlation analysis showed that confidence, preparation, and initial performance scores were all significantly correlated (Pearson's bivariate correlation coefficient) ($p < 0.01$). This indicates that the students were accurate in estimating their levels of preparation and confidence, since the most prepared were also the most confident students, in addition to having higher initial phase scores.

Student statements in the open-ended questions of the survey painted a strongly positive picture, regarding students' attitudes towards the activity. Some of the statements recorded are the following:

Nice program design, well-put questions.

The quiz is a good starting point for the lectures. However it should be kept short.

I really like that you asked us about these things. I am a huge fan of giving feedback and striving for improvement. I am a highly competitive person and the quizzes are compelling to me.

Regarding suggestions for improvement in future versions of the SAGA tool, students suggested (student's statement in quotation marks):

- Gamification: "Maybe a leaderboard/high score list."
- Information on the wrong answers: "It might be nice to know which answers we already got wrong."
- Additional information on peers: "How many lectures the persons have participated in."

- Feedback from a specific group of people: “I would like to see my study-group’s feedback.”
- Splitting the two phases of the quiz before and after the lecture: “Reading the actual curriculum before the class OR repeat the second phase [i.e., revision] of the quiz at the end of the class to actually see if we are taking something out of the lecture.”

Finally, analysis of the SAGA log files showed that although the total allocated time for a weekly quiz was 20 minutes (i.e., 10 minutes for the initial phase, five minutes for the revision, and five minutes for the discussion), the actual time students needed on average was significantly lower. Specifically, students used approximately six minutes for the initial phase and four minutes for the revision. This finding is important, since one of the activity requirements was to keep it short and avoid disrupting the lecture plan. In addition, this allowed more time to the teacher to discuss the questions and students’ performance.

5 Discussion

The weekly quiz analysis showed that the four student groups were comparable throughout the activity. As we have already mentioned, this should have been expected up to a point, since the additional feedback information provided by the confidence and preparation metrics becomes more useful in cases where the percentage metric alone is not enough to guide the students. A certain challenge for designing the current activity was to predict the number of cases where the additional feedback would be important. Despite the effort to have weekly quizzes of similar difficulty, factors such as the complexity of the topics covered each week, students’ preparation, etc. affected students’ scores in the initial phase, thus determining to a great degree the need for additional feedback. As such, the low number of challenging questions identified each week was not enough to create a significant difference between student groups in the weekly quizzes.

At first glance, one could expect that the difference observed in the subset performance would also be evident in the retention quiz. However, the lack of any significant difference in the retention test can be easily explained. The last phase of the activity each week was the discussion between the teacher and the students on the correct answers and students’ performance. In other words, the teacher had plenty of time during the lecture to revisit the questions used in the weekly quiz and provide further explanations to students about the correct choice. The weekly quizzes are a snapshot of students’ knowledge just before the day’s lecture. We expect that this picture was significantly different at the end of the two-hour lecture that followed our activity. Thus, it makes sense that by the fifth week of the activity, all students had acquired the same level of knowledge. Groups’ performance in the retention test was not just comparable; it was overwhelmingly satisfactory, with more than 10% of the student population achieving a perfect score (i.e. 16/16).

By analyzing each question separately, it was easy to figure out that students relied primarily on the percentage metric during the revision phase. It was clear that students

changed their initial answers to the most popular question choice, in cases where the percentage metric was pointing to a specific choice. Students' trust on the percentage metric was so strong that in several cases, students that had selected the correct choice during the initial phase, revised their answers to the most popular, but incorrect, choice during the revision phase. However, despite some misleading cases, the percentage metric is still a very useful way to provide feedback on class knowledge. In our study, the percentage feedback was pointing at the correct choice in 19 out of the 32 total questions of the weekly quizzes. Additional benefits of using the percentage metric are the fact that it is objective and commonly used – thus familiar and easily understood by the students. However, our argument is that the percentage metric cannot provide any qualitative information on the student population under each question choice.

This qualitative information could be offered by metrics such as the level of confidence and preparation. The shortcoming of these two metrics is that they are self-reported (thus subjective) and their validity is affected by students' metacognitive ability to self-assess their level of confidence and preparation. In the current study, preparation, confidence, and initial performance were correlated, suggesting that our students were accurate in their assessment. This may not be the case in a different context, in which student metacognition or their engagement (i.e., time spent on preparation) in the course is low. One question that arises is how students view these two metrics. In our survey, students expressed a positive opinion about the percentage and confidence metrics, while they had diverging views on the usefulness of the preparation metric. A possible explanation could be that students relied more on confidence because it depicted students' understanding after answering a question, while the preparation question was answered before the quiz had started.

Statistical analysis showed clearly that students that received any combination of the two additional feedback metrics were able to significantly outperform the Control group that received only percentage feedback. Thus, the current study provides empirical evidence on the potential of integrating simple subjective metrics, such as the levels of confidence and preparation, in quiz activity, in order to provide a more detailed picture on class knowledge.

Regarding students' attitudes towards the activity, analysis of the SSCO instrument showed that social comparison was not an issue in this study, since no correlation was found between the SSCO score and any other major study variable (e.g., number of revisions performed, initial/revised performance, etc.). However, the low recorded reliability of the tool (i.e., Cronbach's $\alpha = 0.635$) may also be the reason for this. It can only be hypothesized that, in a different context, students' social comparison traits could also affect their behavior in an activity that engages them in comparing their knowledge with that of the whole class.

A very encouraging finding, in favor of the quiz activity, is that despite students' responses in the survey, it seems that their engagement increased significantly during the study. This finding is linked to the learning benefits quiz activities may have on a course in general and it is not attributed to a certain study condition, since it is evident in all four groups.

Overall, students expressed a positive opinion towards the activity, offering also useful suggestions for improvement. Out of these suggestions, students in the Control and Preparation groups asked for the inclusion of the confidence metric in the feedback. Past performance metric was also a popular suggestion, indicating that students are in favor of objective metrics. Finally, it is worth noting that, although they were the least desirable, reading/writing anonymous arguments for each question choice and direct anonymous peer texting were both evaluated positively.

6 Conclusions

This study provided empirical evidence on the potential of combining subjective metrics with widely used objective metrics, such as the percentage, in order to support students in closed-type quiz activities that include a revision phase. The implication for designers and teachers is that subjective metrics can be used effectively to get a better picture of class knowledge and assist students in improving their performance. The three feedback metrics used in this study provide information on three different questions a student may ask in a quiz activity: What do the others say (percentage)? How much have they studied (preparation)? How confident are they of their answers (confidence)?

Future research will focus on including additional metrics and addressing some of the limitations of this study. As such, future studies are planned with larger audiences, different subject matters, and multimodality in representation of the metric information (e.g., combination of text with graphs and color schemes). Finally, as we have repeatedly mentioned, another goal of this series of studies is to analyze the effect of short quizzes on student engagement and performance. As such, the goal of a future study will be to compare classes with and without the quiz activities.

Acknowledgements

This work has been partially funded by a Starting Grant from AUFF (Aarhus Universitets Forskningsfond), titled “Innovative and Emerging Technologies in Education”.

References

1. Papadopoulos, P. M., Natsis, A., & Obwegeser, A. (2017). Improving the Quiz: Student Preparation and Confidence as Feedback Metrics. In the Proceedings of the 9th International Conference on Computer Supported Education – CSEDU 2017, Porto, Portugal. (in press).
2. Buil, I., Catalán, S., & Martínez, E. (2016). Do clickers enhance learning? A control-value theory approach. Computers & Education, 103, 170-182.
3. Sosa, G.W., Berger, D. E., Saw, A. T., &Mary, J. C. (2011). Effectiveness of computer-assisted instruction in statistics: A meta-analysis. Review of Educational Research, 81(1), 97–128.

4. Bransford, J. D., Brown, A., & Cocking, R. (2000). How people learn: Mind, brain, experience and school. Washington, DC, National Academy Press.
5. Kleitman, S., & Costa, D. S. J. (2014). The role of a novel formative assessment tool (Stats-mIQ) and individual differences in real-life academic performance. Learning and Individual Differences, 29, 150-161.
6. Wang, T.-H. (2008). Web-based quiz-game-like formative assessment: Development and evaluation. Computers & Education, 51(3), 1247-1263.
7. Denny, P. (2013). The effect of virtual achievements on student engagement. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13). ACM, New York, NY, USA, 763-772.
8. Bodemer, D. (2011). Tacit guidance for collaborative multimedia learning. Computers in Human Behavior, 27(3), 1079–1086.
9. Janssen, J., & Bodemer, D. (2013). Coordinated computer-supported collaborative learning: Awareness and awareness tools. Educational Psychologist, 48(1), 40–55.
10. Lin, J. -W., Mai, L. -J., & Lai, Y.-C. (2015). Peer interaction and social network analysis of online communities with the support of awareness of different contexts. International Journal of Computer-Supported Collaborative Learning, 10(2), 139-159.
11. Buder, J. (2011). Group awareness tools for learning: Current and future directions. Computers in Human Behavior, 27(3), 1114–1117.
12. Erkens, M., Schlottbom, P., & Bodemer, D. (2016). Qualitative and Quantitative Information in Cognitive Group Awareness Tools: Impact on Collaborative Learning. In Looi, C.-K., Polman, J., Cress, U., & Reimann, P. (Eds.), Transforming Learning, Empowering Learners: 12th International Conference of the Learning Sciences (pp. 458-465). Singapore: International Society of the Learning Sciences.
13. Schnaubert, L., & Bodemer, D. (2015). Subjective Validity Ratings to Support Shared Knowledge Construction in CSCL. In O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), Exploring the Material Conditions of Learning: The Computer Supported Collaborative Learning (CSCL) Conference 2015 (Vol. 2) (pp. 933-934). Gothenburg: International Society of the Learning Sciences.
14. Méndez-Coca, D., & Slisko, J. (2013). Software Socrative and smartphones as tools for implementation of basic processes of active physics learning in classroom: An initial feasibility study with prospective teachers. European Journal of Physics Education, 4(2), 17-24.
15. Papadopoulos, P. M., Demetriadis, S. N., & Weinberger, A. (2013). “Make It Explicit!”: Improving Collaboration through Increase of Script Coercion. Journal of Computer Assisted Learning, 29(4), 383 – 398.
16. DiBattista, D., Mitterer, J. O., & Gosse, L. (2004). Acceptance by undergraduates of the immediate feedback assessment technique for multiple-choice testing. Teaching in Higher Education, 9(1), 17-28.
17. Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments. ACM, New York, 9-15.
18. Wang, A.I. (2015). The wear out effect of a game-based student response system. Computers & Education, 82, 217-227.
19. Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why Students Engage in “Gaming the System” Behavior in Interactive Learning Environments. Journal of Interactive Learning Research. 19(2), 185-224.
20. Papadopoulos, P. M., Lagkas, T. D., & Demetriadis, S. N. (2016). How Revealing Rankings Affects Student Attitude and Performance in a Peer Review Learning Environment.

Communications in Computer and Information Science (CCIS): Computer Supported Education 2015. Vol. 583 Springer Verlag, 2016. p. 225-240.
21. Gibbons, F.X. & Buunk, B.P. (1999). Individual differences in social comparison: The development of a scale of social comparison orientation. Journal of Personality and Social Psychology, 76(1), 129-142.
22. Buunk, A.P., & Gibbons, F.X. (2006). Social comparison orientation: a new perspective on those who do and those who don't compare with others. In Guimond, S. (Ed.) Social Comparison and Social Psychology: Understanding cognition, intergroup relations and culture (pp. 15-33). Cambridge: Cambridge University Press.