Computers and Education Open 2 (2021) 100046

Contents lists available at ScienceDirect

Computers and Education Open

journal homepage: www.sciencedirect.com/journal/computers-and-education-open

Concurrent and retrospective metacognitive judgements as feedback in audience response systems: Impact on performance and self-assessment accuracy

Pantelis M. Papadopoulos[a,*], Nikolaus Obwegeser[b], Armin Weinberger[c]

[a] *Department of Instructional Technology, University of Twente, Cubicus (building no. 41), Office B226 PO Box 217, De Zul 10, 7500 AE Enschede the Netherlands*
[b] *Bern University of Applied Sciences, Bern, Switzerland*
[c] *Department of Educational Technology, Saarland University, Saarbrücken, Germany*

ARTICLE INFO

Keywords:
Audience response systems
Calibration
Metacognition
Self-assessment
Peer Instruction

ABSTRACT

Asking questions in classrooms can produce metacognitive judgements in students about their confidence in being able to answer correctly. In audience response systems (ARSs), these judgements can be elicited and used as additional feedback metrics. This study ($n = 79$) explores how online concurrent item-by-item judgments (OCJ) and retrospective composite judgments of performance accuracy (RJPA) can enhance students' performance and self-assessing accuracy (i.e., calibration – as measured by sensitivity, specificity, and absolute accuracy index). In each of eight weeks, the students answered a multiple-choice quiz and had to denote their level of confidence that their answers were correct (OCJ) and estimate their final score (RJPA). The quizzes followed the voting/revoting paradigm according to which students answer all the quiz questions, receive feedback, and answer the same questions again before the correct answers are shown. The students were randomly grouped into two conditions based on the feedback they received in the ARS: the OCJ group ($n = 41$) received the percentage distribution and peers' OCJs as feedback metrics, while the RJPA group ($n = 38$) received the percentage distribution and peers' RJPAs. Data analysis showed a systemic underconfidence that affected students' OCJ judgments. As a result, students in the RJPA group scored significantly higher than the ones in the OCJ one, were more accurate in self-assessing in the revoting phase, and felt overall more confident in the revoting phase. The study also discusses the relationship between the two judgments employed and the calibration variability between the two study phases.

1. Metacognitive judgments in question asking

Asking questions in the classroom serves multiple purposes, from assessing knowledge to fostering reflection, and the qualities and quantity of questions as well as surrounding conditions, such as wait time and turn-taking, would affect learning [66]. The focus of this research was on the impact on the one student being asked, rather than the "collateral impact" on the whole classroom of peers that would engage in (meta-)cognitive processes and potentially learn from peers. This exact question has been brought to the fore with the advent of audience response systems (ARSs) that allow addressing not one, but all students of a class or seminar at once (also appearing in the literature as classroom/student/personal response systems or as "clickers", referring to polling tools that could use special devices, BYOD approaches, or software tools).

ARSs are versatile instructional tools that can be used before, during, and after a lecture to support a range of learning goals including retention [65], engagement [8], and satisfaction [41]. The popularity of ARSs in the classroom is based, in part, on Mazur's seminal work on the Peer Instruction (PI) paradigm (e.g., [42, 43]) according to which students first provide their initial answers to multiple-choice questions (aka "voting phase"), receive aggregated feedback based on class responses through the ARS (i.e., collective feedback; see [46]) and peer discussion, and then answer the same questions for a second time (a.k.a. "revoting phase") before they receive the correct answers and participate in the class discussion that follows. PI aims to set off both cognitive and

[*] Corresponding author.
E-mail addresses: p.papadopoulos@utwente.nl, pmpapad@gmail.com (P.M. Papadopoulos), nikolaus.obwegeser@bfh.ch (N. Obwegeser), a.weinberger@mx.uni-saarland.de (A. Weinberger).

https://doi.org/10.1016/j.caeo.2021.100046
Received 20 April 2021; Received in revised form 25 July 2021; Accepted 15 August 2021
Available online 16 August 2021

meta-cognitive processes that would lead learners to analyze new information, identify gaps in their knowledge, reflect on, and reconsider their initial answer. But different from one-on-one question asking, in which learners would see and know who is answering, students have little way of assessing the adequacy of their peers' understanding in the anonymous polling with an ARS. Provided only with the audience distribution into the available choices, students may feel encouraged to focus more on probabilistic strategies, thus exhibiting conformity bias changing their initial answers to the most popular one [6, 49, 50].

To address this information gap, studies on online assessment and group awareness have suggested including additional feedback metrics that could better describe the characteristics of the population that voted each question choice. The premise of this approach is that a more detailed depiction of the audience characteristics could support students in making more informed decisions in the revoting phase. One example of such a metric is students' level of certitude (e.g., [35, 53, 54, 55]). In previous studies [4, 5], we have explored the potential of using objective and subjective (i.e., self-reported) feedback metrics that would inform students about their peers' scores in previous tests, their self-reported level of preparation and confidence, and their written justifications for voting a particular choice. These studies showed promising results in how learners make use of additional information in judging the quality of peers' answers, potentially also moving beyond what is possible with a one-on-one question asking in the classroom. There is, however, yet little understanding about how learners would benefit from different metacognitive judgements augmenting ARS use. Such an analysis would have to consider both how including these judgments in collective feedback would impact student performance and how accurate students are while making these judgments, which is directly linked to students' calibration, i.e., their ability to accurately monitor their learning.

In the current study, we explore the potential of two types of metacognitive judgments as feedback metrics through a series of ARS activities to facilitate shifts from guessing to knowing. Metacognitive judgements form the basis for self-regulated learning in identifying and addressing discrepancies between desired and monitored levels of knowledge (e.g., [17, 48]). Metacognitive judgements have been classified as prospective, concurrent, and retrospective, depending on when they are being made in the learning process [56]. Prospective judgments are elicited before the task at hand (i.e., studying or testing, depending on the judgment) and they provide insights on how students self-assess their ability to perform, retrieve information from memory, manage the time/effort needed for the task, etc. Concurrent judgments are elicited during the task at hand and because they are recorded during the task, they are usually fine-grained and refer to specific items within the tasks (i.e., item-by-item judgments). Finally, retrospective judgments are elicited after the task at hand and, therefore, usually refer to the whole task (i.e., composite judgments).

In this study, used concurrent and retrospective judgements. Specifically, in the voting phase, we asked students to note their level of confidence while answering each question in item-by-item judgments (i.e., one judgment per question, referred to as "Online Confidence Judgments" – OCJ, e.g., [56]), and provide a composite judgment estimating their overall score once they have answered all questions (i.e., referred as "Retrospective Judgments of Performance Accuracy" – RJPA, e.g., [1]). Their aggregated judgments and the percentage that each question choice received were presented as collective feedback in the revoting phase in which student answered the same question for the second time and submitted their updated judgments. This study design allowed us to evaluate different ways of supporting students during ARS activities, but also to provide empirical evidence on the relation between item-by-item and composite judgments, pertaining to explore the role of metacognitive judgements in question asking [56].

2. Theoretical background: enhancing ARSs with information for metacognitive judgements

2.1. Calibration measures

Even though metacognitive judgements can be predictors of academic success (e.g., [29]), learners are often unable to accurately monitor their knowledge [48, 63]. Studies have shown that the type of judgments used can also affect predictions on academic performance [29, 48].

Calibration describes the relation between metacognitive judgments and actual performance, or in other words, the monitoring accuracy of one's metacognition [57]. Calibration can further refer to absolute or relative accuracy. Absolute accuracy measures the agreement between metacognitive judgments and performance, while relative accuracy measures the relation between correct and incorrect judgments or a set of judgments against a performance set [56]. Literature includes several calibration measures for each type of accuracy, each one focusing on a different aspect. For example, commonly used measures of absolute accuracy are the absolute accuracy index (AAI) (1) that measures the discrepancy between judgments and performance (e.g., [48, 51, 56]), and the bias index that measures the degree students over- or under-judged their performance (e.g., [19]). Similarly, the correlation coefficient that measures the relationship between sets of judgments and performance (e.g., [1]), and the gamma coefficient that measures the dependence between judgments and performance (e.g., [33]) are two of the most commonly used measures of relative accuracy.

$$\text{absolute accuracy index} = \frac{1}{N}\sum_{\iota=1}^{\text{N}}(judgment_i - performance_i)^2 \tag{1}$$

According to Schraw et al., [57], there is no single calibration measure that appears above the others in terms of validity, reliability, and appropriateness (for an overview on calibration measures and their connections, see [24, 56, 57]). The scale used for both the judgments and the performance can be continuous, discrete, or dichotomous. While there are several studies on calibration that use continuous or discrete scales (e.g., [1, 11]), the majority focuses on dichotomous scales and the 2 × 2 contingency table presented in Table 1. Depending on the judgment type used (e.g., JOL, EOL, confidence), a positive judgment means that the student predicted a successful memory retrieval (JOL), expected a high level of ease in learning new information (EOL), felt confident about the performance, etc. Similarly, a positive performance outcome means that the student answered the respective item correctly.

Researchers suggest that, when possible, multiple calibration measures should be employed to represent students' monitoring accuracy, while Schraw et al., [57] compared ten measures, including measures such as simple matching, gamma, g index, odds ratio etc., and concluded that the combination of sensitivity and specificity is the best option as it can explain close to 100% of sample variance. Sensitivity measures the proportion of positive judgments when student's answer is correct (2), while specificity measures the proportion of negative judgments when student's answer is incorrect (3). Statistical analysis showed that these two absolute accuracy measures are orthogonal (i.e., statistically independent) [57]. Feuerman and Miller [24] suggested that the two measures may utilize two independent metacognitive processes, namely judging correct and incorrect performance. It is important to note that both values need to be taken into account when examining the

Table 1
Judgment – performance contingency table.

		PERFORMANCE OUTCOME		
		Positive	Negative	Row marginals
JUDGMENT	Positive	A *(true positive)*	B *(false positive)*	*A* + *B*
	Negative	C *(false negative)*	D *(true negative)*	*C* + *D*
Column marginals		*A* + *C*	*B* + *D*	

diagnostic power of a test as focusing only on one of them may be misleading.

$$\text{sensitivity} = \frac{A}{A + C} \tag{2}$$

$$\text{specificity} = \frac{D}{B + D} \tag{3}$$

2.2. Learning with audience response systems

Audience response systems have been well investigated with regard to several aspects of using quizzes for learning (for an overview, see [16, 32]). A quiz can be an individual or group-based activity [44]. When used before a lecture, a quiz can identify preconceptions and assumptions [2, 30]. Conducting a quiz activity during a lecture allows the teacher to get more information about the audience [34], while the use of a quiz at the end or after a lecture can enhance reflection and retention [60, 65].

ARSs have often been linked to higher student engagement [32], development of critical thinking [28, 67], and increased student satisfaction [41]. Empirical evidence has attributed ARS activities with higher academic performance and better knowledge acquisition [58, 59, 60], improved retention [65], and higher course grades [12]. Compared to other classroom technologies, the success and popularity of ARSs have also been linked to a lower technological barrier for the teacher [9, 23]. As a result, ARSs are widely adopted by teachers with studies reporting on teachers' appreciation of quizzes potential in revealing misconceptions and assisting in organizing lectures [15, 38].

Despite a consensus on the educational value of ARSs, there are studies providing mixed results [22, 27], thus suggesting that the tool itself is not a panacea of fostering learning without an appropriate underlying pedagogy. Therefore, many researchers suggest a need for further studies on ARSs and their underlying pedagogy [16, 60]. As mentioned earlier, perhaps the most effective didactical approach regarding ARSs is Mazur's Peer Instruction model [42, 43] that includes a peer discussion session. PI has been associated with higher academic performance (e.g., [7]) and improved classroom interaction during lectures [10]. Nevertheless, it is also apparent that allocating the necessary time for peer interaction is not always efficient, especially in the cases of large audiences and multiple questions with studies employing variants of the PI method (for an overview, [7, 64]). In addition, quiz participation is often anonymous, thus enhancing psychological safety and acceptance for the students [8, 62]. Peer discussion removes students' anonymity and can make students less willing in discussing their answers with their peers [47].

Mazur [43] argued that ARS tools should be examined in conjunction with the underlying pedagogy stating also that "it is not the technology, but the pedagogy that matters" (ibid., p.51). While we maintain that metrics could not fully substitute the benefits of a peer discussion, we also argue that the feedback offered to students in most ARSs could be greatly improved as the information on student distribution to each question choice (either as a percentage or a sample size) is not adequate and does not provide insights about the audience to the student.

2.3. Study motivation and research questions

Since there is a plethora of design and instructional approaches regarding ARS tools, it needs to be noted at this point that our focus is on tools that are based, at least in part, on the Peer Instruction paradigm (e.g., [42, 43]). This means that systems or approaches that only support a voting phase, do not provide system-based collective feedback between the two voting phases, or include teacher/corrective feedback (e.g., hints) before the revoting phase are outside the scope of this study.

Our approach while working in ARSs is to enrich the system-generated collective feedback the students receive during the voting/revoting paradigm in such a way that would allow students to make more informed decisions during the revoting phase and improve their performance. In a previous study [4], we employed students' self-reported levels of confidence (OCJs) and preparation, and their past performance as feedback metrics, to offer a better view of the audience. Among these feedback metrics, confidence was the most effective one resulting in higher quiz performance and limited conformity bias, as students that had voted the correct, but not most popular answer choice did not change their answer to the most popular one, relying on the confidence levels of their peers (i.e., "confidence by proxy").

Despite the observed value of the confidence metric, the question remained whether this type of feedback is accurate and whether another type of metacognitive judgment could provide better scaffolding to the students. To address this research question, we compare in the current study the item-by-item OCJs to the composite RJPAs in a voting/revoting design over eight weeks. The granularity of metacognitive judgments and the relationship of judgments of different "grain" size has already been marked as an important research question. As Schraw [56] pointed out, most studies investigate item-by-item judgments, with a lower number examining composite judgments (e.g., [20]). Schraw also noted that "there are currently no studies in the metacognitive monitoring literature that attempt to compare the reliability and utility of item-by-item versus composite judgments. This is an important topic for future research" ([56], p. 424), and this still seems to be the case up to this day, especially in the context of ARS research.

Our overarching research goal has two dimensions. One is on tuning ARSs, namely how to improve the learning experience for the students, and in the case of the current study, how to improve the helpfulness of the collective feedback. The other is on understanding the intricate mechanisms that happen when a question is asked to learners. And this is clearly not about testing their knowledge alone, but also about understanding the metacognitive workings of when they are being asked, when judging peers and their answers, getting it right or wrong, etc. The analysis of students' calibration and the use of two metacognitive judgments can provide insights into the latter dimension while using these metacognitive judgments in the collective feedback allows us to examine the link between accuracy and helpfulness for these two metrics. Based on the above, the study investigated the following research questions:

- RQ1: Which of the two metacognitive judgments used as feedback metrics (OCJs/RJPAs) in multiple-choice quizzes is more beneficial for the students in terms of performance, confidence, and calibration?
- RQ2: How does students' calibration as depicted through the two metacognitive judgments (OCJs/RJPAs) changes after feedback is received?
- RQ3: What is the relationship between the two metacognitive judgments (OCJs/RJPAs)?

3. Method

3.1. Participants and domain

Participation in the study was an optional, non-graded, eight-week activity within a second-year course of a Management Bachelor's program. The students were grouped randomly into two study conditions according to the feedback they received during the revoting phase:

- OCJ: 41 students, who received the percentage distribution and peers' online confidence judgments as feedback metric.
- RJPA: 38 students, who received the percentage distribution peers' retrospective judgments of performance accuracy as a feedback metric.

We highlighted how participation in the weekly quizzes may provide cognitive benefits for the students, and to incentivize participation further, the five students with the highest overall scores in each study condition received gift vouchers for the university's bookstore (10 euros). The analysis includes data only from 79 participants that completed the whole eight-week study, as an additional 27 students participated partially and were excluded from the study. Participation in the study quiz was fully anonymized, using a system-generated numerical user-id to trace student activity throughout the study. Only voucher winners had to reveal their identities to receive their rewards.

3.2. The SAGA tool

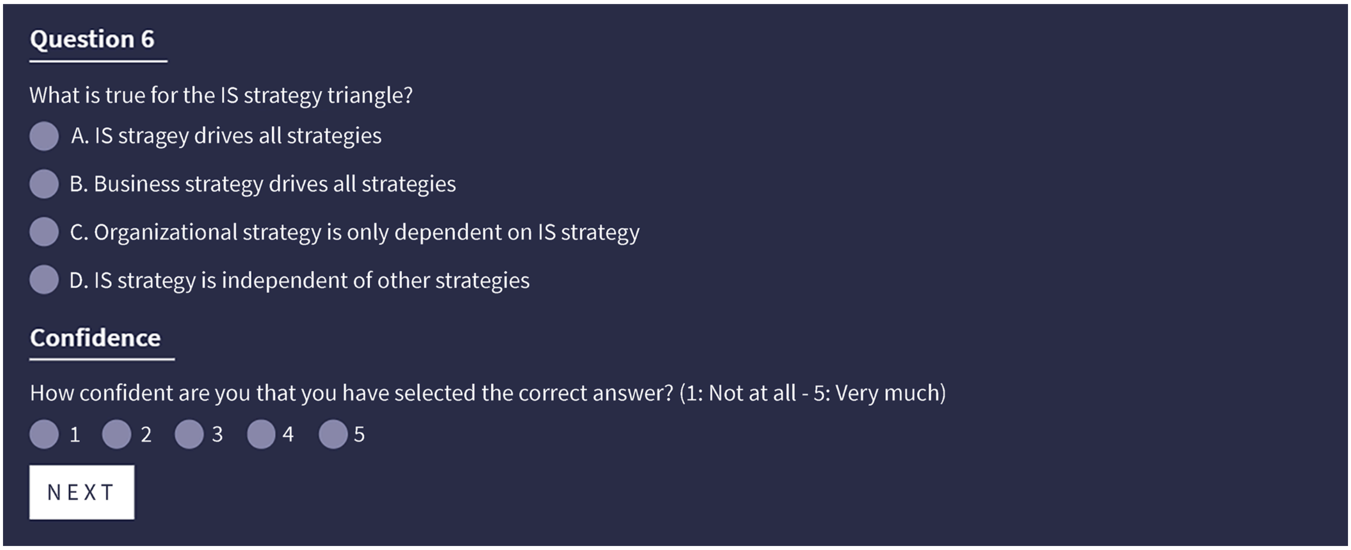
The study used SAGA audience response system, an adaptable online environment developed by our team in which the teacher can customize the feedback metrics the students receive during a quiz. SAGA (acronym for self-assessment/group awareness) is based on the vote/revote paradigm for quizzes and it has been used several times for teaching and research activities.

All participants had to submit OCJs and RJPAs for both phases of a quiz. Each quiz had eight multiple-choice questions accompanied by the OCJ form (titled "Confidence", Fig. 1) that asked students to denote their confidence about their answers using a 1–5 scale (1: Not at all – 5: Very much). At the end of each quiz phase, the students had to self-assess their performance and predict their phase score (titled "Self-Assessment", Fig. 2) using a 0–8 scale (i.e., range of possible scores in the quiz). This procedure was identical for all students on SAGA. To assist students in the revoting phase, SAGA was providing aggregated feedback to them according to their study conditions (i.e., percentage of peers that answered each question choice, plus the respective OCJ/RJPA score) (Figs. 3 and 4).

It is worth noting that the reason for using multiple-choice questions over other closed-type items is that it is one of the most commonly used question types, posing minimum complexity in the answering process, and being widely familiar to all students.

3.3. Design

The study included two groups in vivo conditions of a university course. Students' performance and activity in the two phases (i.e., scores, confidence, self-assessment, calibration, etc.) were the dependent variables of the study, while the feedback condition (percentage + confidence vs. percentage + self-assessment) during the revoting phase was the independent variable.



Question 6

What is true for the IS strategy triangle?

○ A. IS stragey drives all strategies

○ B. Business strategy drives all strategies

○ C. Organizational strategy is only dependent on IS strategy

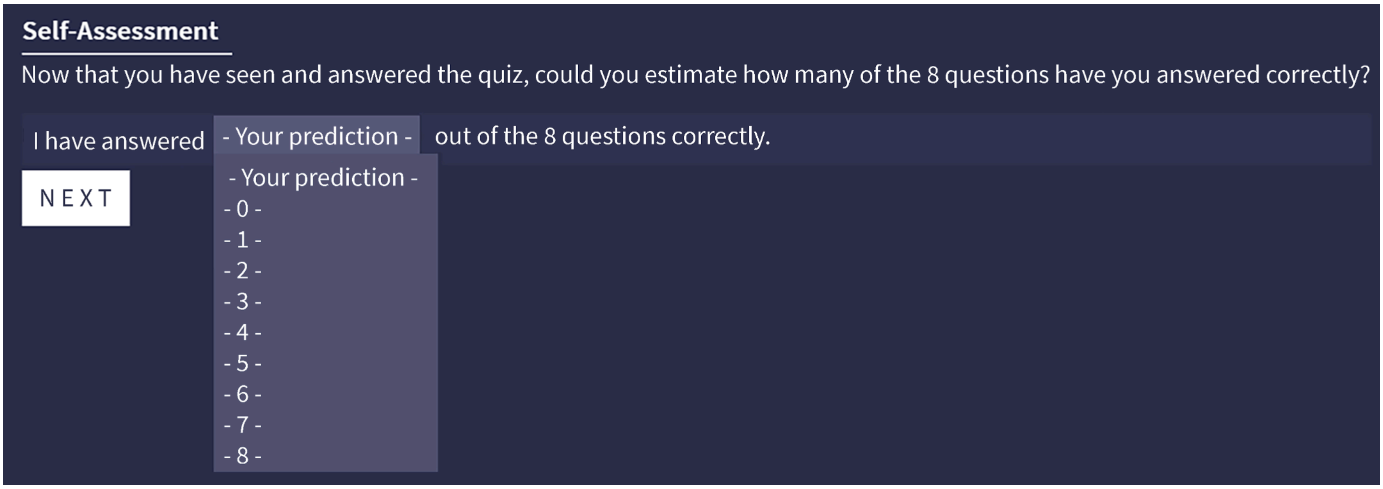
○ D. IS strategy is independent of other strategies

Confidence

How confident are you that you have selected the correct answer? (1: Not at all - 5: Very much)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Fig. 1. SAGA screenshot during the voting phase.

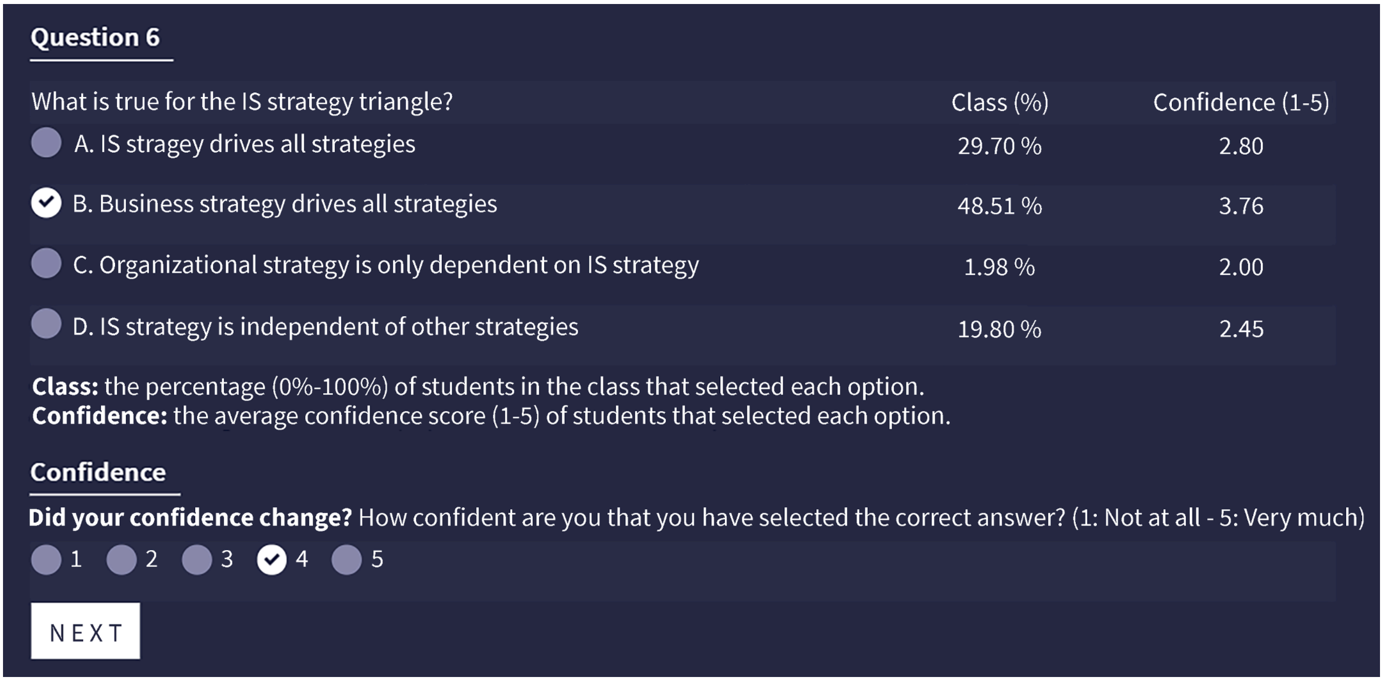


Self-Assessment

Now that you have seen and answered the quiz, could you estimate how many of the 8 questions have you answered correctly?

I have answered - Your prediction - out of the 8 questions correctly.

Fig. 2. SAGA screenshot after a phase has finished and students are asked to self-assess and predict their overall score in the phase.



Question 6

What is true for the IS strategy triangle?	Class (%)	Confidence (1-5)
○ A. IS stragey drives all strategies	29.70 %	2.80
✓ B. Business strategy drives all strategies	48.51 %	3.76
○ C. Organizational strategy is only dependent on IS strategy	1.98 %	2.00
○ D. IS strategy is independent of other strategies	19.80 %	2.45

Class: the percentage (0%-100%) of students in the class that selected each option.
Confidence: the average confidence score (1-5) of students that selected each option.

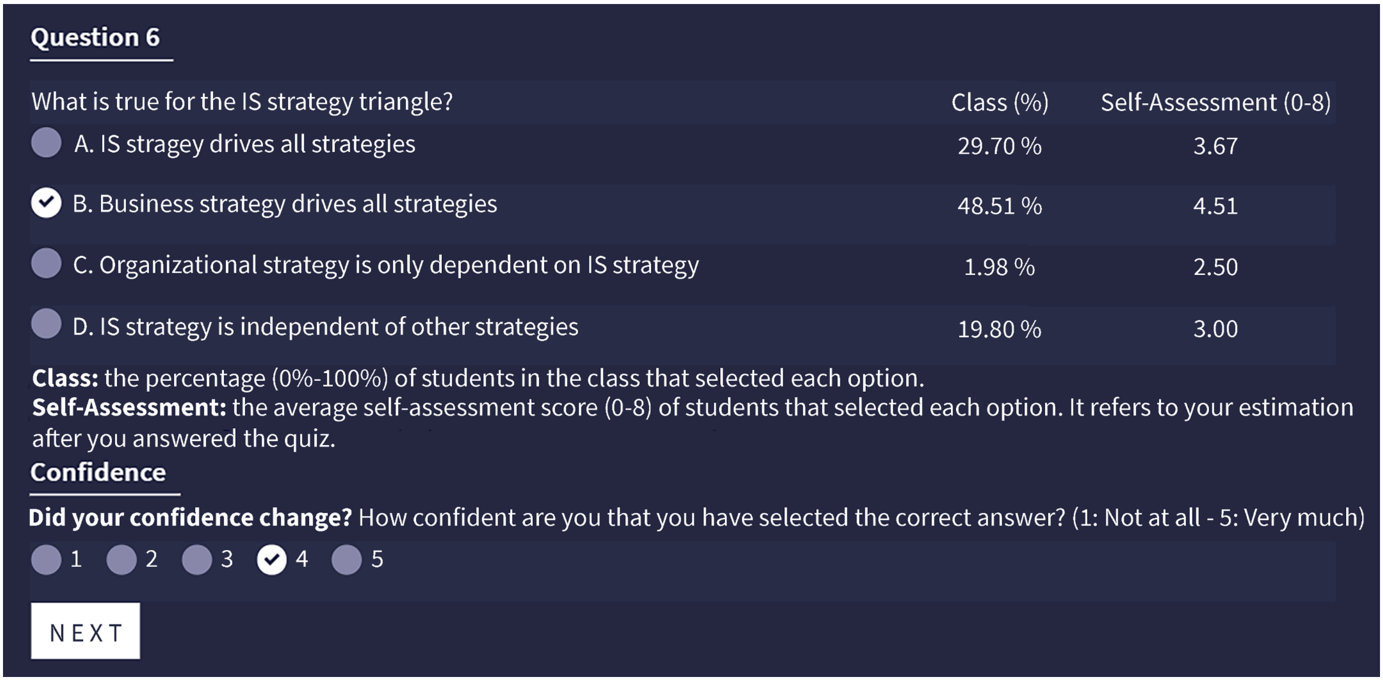
Confidence

Did your confidence change? How confident are you that you have selected the correct answer? (1: Not at all - 5: Very much)

○ 1 ○ 2 ○ 3 ✓ 4 ○ 5

NEXT

Fig. 3. SAGA screenshot during the revoting phase for OCJ group. The aggregated percentage and confidence scores (OCJs) for each question choice are presented as feedback metrics.



Question 6

What is true for the IS strategy triangle?	Class (%)	Self-Assessment (0-8)
○ A. IS stragey drives all strategies	29.70 %	3.67
✓ B. Business strategy drives all strategies	48.51 %	4.51
○ C. Organizational strategy is only dependent on IS strategy	1.98 %	2.50
○ D. IS strategy is independent of other strategies	19.80 %	3.00

Class: the percentage (0%-100%) of students in the class that selected each option.
Self-Assessment: the average self-assessment score (0-8) of students that selected each option. It refers to your estimation after you answered the quiz.

Confidence

Did your confidence change? How confident are you that you have selected the correct answer? (1: Not at all - 5: Very much)

○ 1 ○ 2 ○ 3 ✓ 4 ○ 5

NEXT

Fig. 4. SAGA screenshot during the revoting phase for RJPA group. The aggregated percentage and self-assessment scores (RJPAs) for each question choice are presented as feedback metrics.

3.4. *Procedure*

The study lasted eight weeks. The instructional approach adopted elements of the flipped classroom paradigm, as the material of the lecture (i.e., slides, texts, online resources) was available a week in advance and the students were encouraged to study it before coming to class. The quiz was taking place during the first 15 min of the lecture (voting: 10′; revoting: 5′). At the end of the quiz, SAGA was revealing the correct answers and students' scores in the two phases, while the teacher was able to discuss the answers further in the ensuing lecture. In the end-of-the-study questionnaire, the students were asked to evaluate the overall activity and the helpfulness of the feedback metrics they received.

3.5. *Data collection and analysis*

The *t*-test was used to compare students' activity in the voting phase, while analysis of covariate (ANCOVA) was used for the revoting phase, using the respective voting values as covariates. Paired-samples *t*-test was used to analyze differences between voting and revoting phases within the groups. Pearson's correlation coefficient was used for the bivariate analysis of feedback metrics, calibration, and performance. One-way repeated analysis of variance (repeated ANOVA) was used to analyze students' performance and calibration over the period of eight weeks. For all statistical tests, a level of confidence at 0.05 was used, while none of the test assumptions of the tests used was violated.

Regarding calibration measures, the students provided eight OCJs and one RJPA for each phase of a quiz. Even though students denoted their confidence levels using a 1–5 scale, we transformed their judgments into "confident/not confident" using the mean value of their responses (M = 3.14, SD = 1.03). Eventually, confidence levels of 1–3 were marked as "not confident" and 4–5 as "confident". By doing so, we allowed students to use a scale that was more convenient for them than a dichotomous one and we were able to perform the calibration analysis using the sensitivity and specificity measures, thus following Schraw et al. [57] suggestion. In addition, accepting the suggestion that judging correct/incorrect performance may involve two independent metacognitive processes [24] the use of this measure combination allowed us to analyze students' calibration while answering correctly/incorrectly and how their calibration changed when they revise their answers in the revoting phase.

Since there was only one RJPA per student in each quiz phase (voting/revoting), students' accuracy regarding RJPAs was calculated for each phase by using the AAI for the judgments made within a group.

4. Results

4.1. Student performance

Table 2 shows the aggregated average performance of students in the two study conditions for all eight weeks. T-test results showed that the two groups were comparable in the voting phase ($p > .05$), while ANCOVA results, using the voting score as a covariate, revealed a statistically significant difference in the revoting scores with a minimal effect size in favor of the RJPA group ($F(1, 76) = 4.28$, $p = .03$, $\eta_p^2 = .01$). Paired-samples t-test results revealed that both groups increased their performance significantly during the revoting phase (OCJ: $t[40] = 4.06$, $p < .01$, $d = .78$; RJPA: $t[37] = 4.11$, $p < .01$, $d = .92$).

4.2. Student judgments

Table 3 shows the aggregated average confidence judgments for all eight weeks. Once again, t-test results showed that the two groups were comparable in the voting phase ($p > .05$), while ANCOVA results, using the confidence values in the voting phase as a covariate, showed a statistically significant difference in the revoting phase in favor of the RJPA group ($F(1, 76) = 6.07$, $p < .01$, $\eta_p^2 = .04$). Paired-samples t-test results revealed that students in both groups felt significantly more confident in the revoting phase (OCJ: $t[40] = 5.01$, $p < .01$, $d = 1.12$; RJPA: $t[37] = 5.70$, $p < .01$, $d = 1.33$).

Table 2
Student performance.

	OCJ group *(n = 41)* M *(SD)*	RJPA group *(n = 38)* M *(SD)*
Voting (0–8)	4.79 *(1.83)*	4.70 *(1.85)*
Revoting (0–8)*	6.11 *(1.58)*	6.26 *(1.59)*

* $p < .05$.

Table 3
Students' online confidence judgments.

	OCJ group *(n = 41)* M *(SD)*	RJPA group *(n = 38)* M *(SD)*
Voting (1–5)	2.97 *(0.96)*	2.99 *(1.02)*
Revoting (1–5)*	3.20 *(1.01)*	3.43 *(1.04)*

* $p < .05$.

Table 4
Students' retrospective judgments of performance accuracy.

	OCJ group *(n = 41)* M *(SD)*	RJPA group *(n = 38)* M *(SD)*
Voting (0–8)	3.61 *(1.96)*	3.62 *(1.82)*
Revoting (0–8)	5.18 *(1.77)*	5.22 *(1.71)*

Table 4 shows the aggregated average judgments the students made while self-assessing their performance after the voting and revoting phases. Students in the two conditions self-assessed themselves similarly ($p > .05$), while paired-samples t-test results showed that both groups predicted significantly higher scores in the revoting phase (OCJ: $t[40] = 4.78$, $p < .01$, $d = 1.07$; RJPA: $t[37] = 4.91$, $p < .01$, $d = 1.14$).

Pearson's bivariate correlation test also showed that students' OCJ and RJPA judgments were positively correlated in both quiz phases (voting: $r = .71$, $p < .01$; revoting: $r = .72$, $p < .01$).

4.3. Calibration measures

Table 5 shows the three calibration measures for the two groups in the two quiz phases. Note that for sensitivity and specificity, high values suggest high accuracy, while high accuracy is denoted by low values for the AAI.

In the voting phase, t-test results showed that the two groups were comparable in all calibration measures ($p > .05$). The specificity and AAI measures showed high accuracy, while sensitivity values suggested that the students were underconfident in questions they answered correctly.

In the revoting phase, the groups were comparable regarding the AAI ($p > .05$), while ANCOVA results revealed statistically significant differences in favor of the RJPA group for sensitivity ($F(1, 76) = 4.21$, $p < .01$, $\eta_p^2 = .02$) and the OCJ group for specificity ($F(1, 76) = 4.45$, $p < .05$, $\eta_p^2 = .02$).

Paired-samples t-test results showed significant differences in all three calibration measures, but only for the RJPA group (sensitivity: $t[37] = 2.23$, $p < 0.01$, $d = .52$; specificity: $t[37] = 1.66$, $p < 0.01$, $d = .39$; AAI: $t[37] = 1.88$, $p < 0.01$, $d = .44$). On average, the values of sensitivity increased during the voting phase and decreased for specificity and AAI (Fig. 5).

Figs. 6 and 7 show the distribution of students in the 2 × 2

Table 5
Calibration metrics.

		OCJ group *(n = 41)* M *(SD)*	RJPA group *(n = 38)* M *(SD)*
Sensitivity (online confidence judgments)	Voting	.44 *(0.15)*	.48 *(0.17)*
	Revoting*	.48 *(0.16)*	.57 *(0.18)*
Specificity (online confidence judgments)	Voting	.81 *(0.15)*	.80 *(0.16)*
	Revoting*	.78 *(0.19)*	.73 *(0.20)*
Absolute accuracy index (retrospective judgments of performance accuracy)	Voting	.09 *(0.13)*	.10 *(0.12)*
	Revoting	.08 *(0.13)*	.05 *(0.11)*

* $p < .05$.

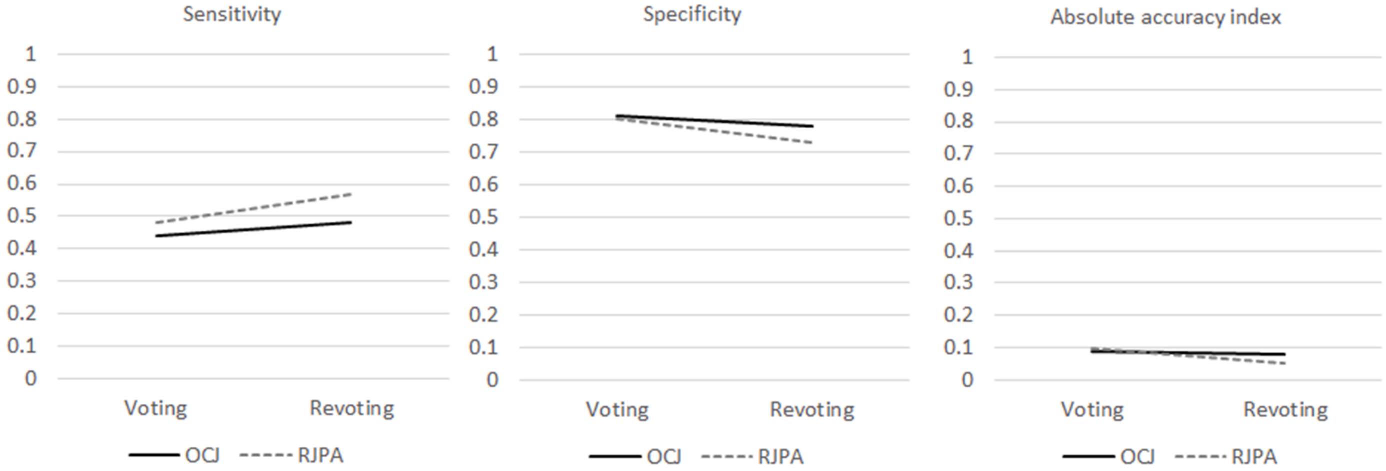


Fig. 5. Changing of calibration metrics in the two quiz phases.

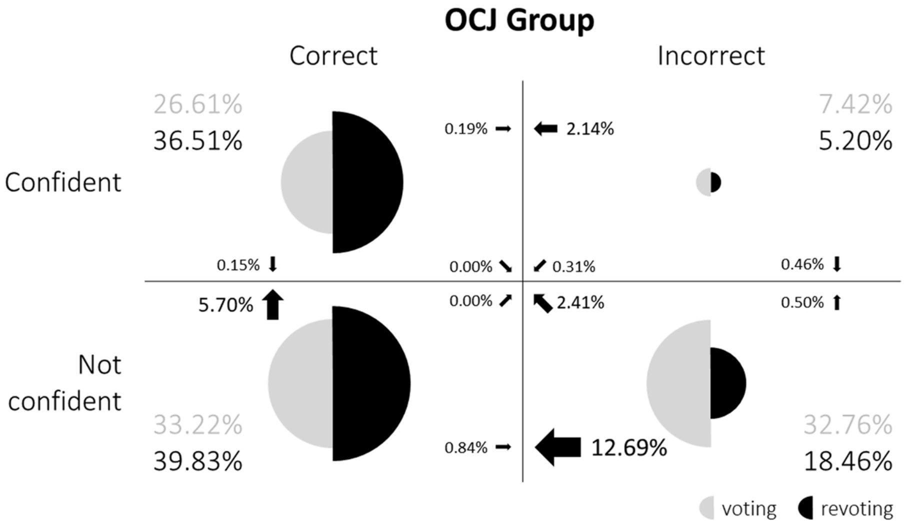


Fig. 6. OCJ group movements from voting to revoting.

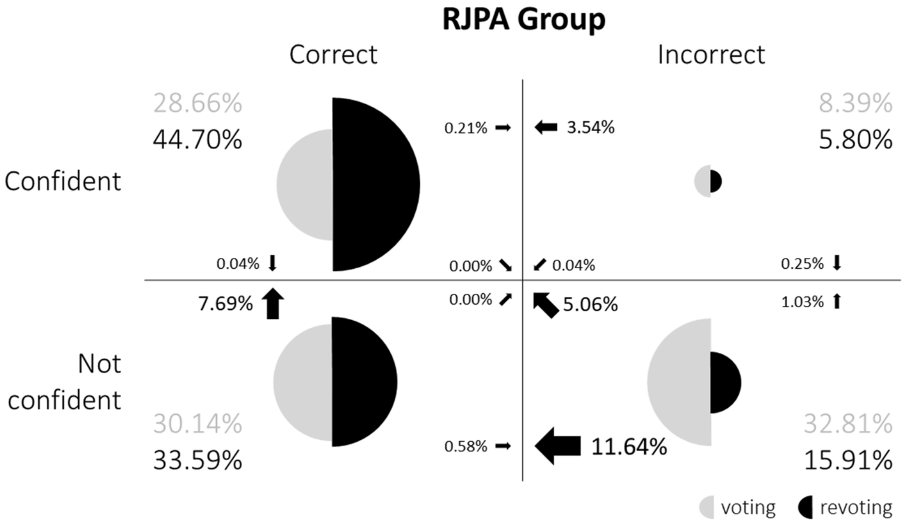


Fig. 7. RJPA group movements from voting to revoting.

contingency table for the two phases for the OCJ and RJPA groups, respectively. In both cases, there are two distinctive population shifts during the revoting phase. Students from the bottom right cell (true negative) move to the bottom left cell (false negative) and students from the bottom left cell move to the top left one (true positive).

Pearson's bivariate correlation analysis showed that students' performance in the voting phase was positively significantly correlated to sensitivity ($r(78) = .28$, $p < .01$) and negatively to AAI ($r(78) = .−20$, $p < .01$) and to specificity ($r(78) = −.18$, $p < .01$).

4.4. *Student opinions on feedback metrics*

Table 6 presents students' responses in the end-of-study questionnaire. The students were positive about the activity and appreciated the anonymity provided in it. The percentage information was deemed highly helpful in the revoting phase for both groups, while the metacognitive judgments received mostly positive opinions, although paired-samples *t*-test results showed that these were significantly less positive than the ones for the percentage (OCJ-percentage: $t[40] = 2.23$, $p < .01$, $d = .50$; RJPA-percentage: $t[37] = 2.49$, $p < .01$, $d = .56$).

5. Discussion

5.1. *RQ1: the potential of using OCJs/RJPAs as feedback metrics: systemic underconfidence and advantage for RJPAs*

Eliciting metacognitive judgments can affect learning [61]. In this study, however, the focus was not on the impact of eliciting judgments but on the potential of these judgments as useful feedback metrics. Therefore, we elicited the same metacognitive judgments from all participants and the two treatment groups differed only on the types of metacognitive judgments we used as feedback in the revoting phase. Results showed that students had a positive opinion about the feedback metrics they received, with the percentage metrics being deemed significantly more helpful. This was expected as students were more familiar with that metric.

Both groups improved their performance and their level of confidence going from the voting to the revoting phase. This was expected and has been regularly reported as the effect that feedback in an ARS can have on learning (e.g., [4, 31, 39]). Despite the small effect sizes recorded, the repeated differences observed between the two groups in scores, confidence levels, and calibration during the revoting phase suggest that, indeed, the two metacognitive judgments had a different effect on students when used as feedback metrics. Students in the RJPA group scored higher and felt more confident about their answers than their counterparts in the OCJ group in the revoting phase. Regarding calibration, the three measures used showed three different outcomes, but we can also note that in all three cases, the values for the RJPA group changed significantly, while they remained largely unchanged for the OCJ group. This also suggests an effect of the feedback treatment on students' calibration. A possible explanation for the small effect sizes could be linked to the scales used in the study (1–5 for OCJs; 0–8 for quiz performance and RJPAs). This may have limited variability, especially since any learning or confidence gains recorded in the revoting phase were dependent on the remaining range of the scale left by the scores and confidence of the voting phase. An alternative explanation could be found on what Hubbard and Couch [31] reported regarding the positive effect of clicker questions being based on students' initial performance in the voting phase. In other words, while collective feedback can help students improve their performance, their final scores may still be correlated to their voting scores.

Table 6
Students' opinions.

		OCJ ($n = 41$) M *(SD)*	RJPA ($n = 38$) M *(SD)*
Q1	I liked the quiz activities.	3.87 *(1.11)*	3.99 *(1.20)*
Q2	Answering anonymously was important for me.	4.12 *(0.88)*	4.07 *(0.94)*
Q3	I found the percentage information helpful in revising my answers.	3.88 (1.06)	3.94 *(1.00)*
Q4$_{\text{(OCJ)}}$	I found the confidence information helpful in revising my answers.	3.39 *(0.94)*	n.a.
Q4$_{\text{(RJPA)}}$	I found the self-assessment information helpful in revising my answers.	n.a.	3.42 *(0.87)*

Sensitivity values were low for both groups in both phases, suggesting that the students were underconfident when they were answering correctly. The RJPA group had significantly higher sensitivity values in the revoting phase, but a high number of students that answered correctly remained underconfident. Contrary to the sensitivity measure, the specificity values were high for both groups in the voting phase, and remained moderately high in the revoting phase, despite the significant drop observed for the RJPA group. High specificity values imply that it was easy for the students to identify in which questions they have an incorrect answer. However, an alternative explanation could be that students were constantly underconfident throughout the study, and therefore, their specificity values were high whenever students' general feeling of uncertainty coincided with incorrect answers. Finally, the absolute accuracy index showed that both groups were highly accurate in their judgments and their accuracy improved in the revoting phase, with the RJPA group achieving a significant improvement.

The mixed picture painted by the three calibration measures can be simply explained when looking closer at the metacognitive judgments the measures were based on. On one hand, the values of both sensitivity and specificity converged to the same finding, that students reported low OCJs in the activity. Our results are in line with the systemic underconfidence reported in previous studies that included multiple cycles and eliciting judgments right after a study cycle (e.g., [25, 26, 37]). This phenomenon is called underconfidence with-practice [37]. In addition, underconfidence is common also when unit judgments are employed [52]. This means that students' confidence was misplaced and that OCJs were not always effective in directing students to the correct answer. On the other hand, as the high AAI values suggested, students were accurate in their RJPAs. We hypothesize that the fact that students made only one RJPA judgment per quiz phase may have diminished the attrition caused by multiple cycles. Eventually, students that saw RJPAs as feedback in the revoting phase (i.e., RJPA group) were able to identify the correct question choice more easily than students that saw OCJs as feedback (i.e., OCJ group).

Based on the above, we argue that RJPAs were indeed more helpful feedback for the students, even though the low observed effect sizes suggest only a modest improvement.

5.2. *RQ2: calibration changes during study cycles: stepwise improvement toward true positive*

Results showed that calibration measures and students' performance were significantly correlated (positively for sensitivity and AAI and negatively for specificity). Bivariate correlation analysis showed that sensitivity and AAI measures suggested that lower performance results in lower accuracy. This is in line with studies that have identified task difficulty as a moderator of students' accuracy, with harder tasks resulting in lower accuracy [13, 40]. Specificity, however, painted the opposite picture. As we discussed earlier though, this finding can be explained when examining the overall underconfidence of the students in the activity. As Figs. 6 and 7 showed, the majority of students in both groups were in the two bottom cells of the 2 × 2 contingency table (i.e., cells C and D; true/false negative judgments, respectively on Table 1) during the voting phase. Consequently, a harder quiz means more incorrect answers, a higher population of students in the D cell (true negative), and therefore a higher specificity value. The different outcomes of the three calibration measures used support vividly what Schraw et al., [57] suggested that there is not a single calibration measure to fully present student activity and when possible, a combination should be employed. The analysis of how students' calibration changes during the revoting phase shows an even clearer picture of the potential and limitations of sensitivity and specificity to depict students' calibration.

Figs. 6 and 7 revealed how students move across the 2 × 2 contingency table. Ideally, students should end up in cell A (true positive), meaning getting the correct answer and being confident about it. Cell B (false positive) was the least populated one and most of the students moving out of it landed on cell A. This resembles the "hypercorrection effect" according to which corrective feedback on test items in which the students were confident but answered incorrectly is more likely to make these students answer correctly the same item in a future test [14, 45]. A second movement recorded was from cell D to cell A; students that were not confident about their (incorrect) answer, answered correctly in the revoting phase and felt confident about their answer. As cells A and D are indicative of students' accuracy according to the formulas of sensitivity (2) and specificity (3), this movement suggested that students' overall accuracy remained the same.

Arguably, the most prominent and most important movement revealed was that the majority of students that leave cell D land in cell C and the majority of students that move out of cell C end up in cell A. However, the ᴅ-C-A sequence results in a temporary drop in students' accuracy as depicted through sensitivity and specificity. As many studies have already reported, higher accuracy may affect self-regulation, and eventually performance (e.g., [17, 48]). However, examining students' populations in the 2 × 2 contingency table and their movements during the revoting phase, we argue that despite the temporary drop of calibration, moving into cell C represents a positive impact of the feedback approach as cell C represents for many students their interim state during their eventual transition to cell A. This finding can be grounded in Bloom's taxonomy [3] since for students it is one thing to attain the right answers, and another, second step to become gradually more confident about it. Therefore, feedback acts in two levels affecting students' performance and their metacognitive feelings, which according to Efklides [21] are affectively charged and inform the students on specific aspects of cognitive processing such as the feeling of familiarity, knowing, and confidence. We argue that the analysis of calibration and feedback strategy in an ARS should take into account both levels of student activity and consider students' trajectories within the 2 × 2 contingency table.

5.3. RQ3: the relationship between OCJs and RJPA: correlated but different

Despite the greater benefit that RJPAs seemed to have when used as feedback in the ARS activity, both judgments were helpful for the students as both groups improved their performance, confidence, and accuracy in the revoting phase. In addition, both judgments were lower in relation to students' performance, but as discussed above, students were more accurate in the RJPAs. Pearson's correlation analysis showed strong positive correlations between the two judgments in both phases of a quiz, suggesting that students' item-by-item judgment informed their later composite judgment. Since RJPAs were more accurate, we can hypothesize that students rounded up their score predictions expecting to have some correct answers in the subset of questions in which they did not feel confident. As mentioned earlier, eliciting multiple confidence judgments may lead to underconfidence (e.g., [37, 52]). As this process adds cognitive load for the students which also results in prolonged activity duration, the implication for the instructional design would be to focus on composite judgments. However, this needs to be tested, as studies have also suggested that eliciting judgments may have a positive effect on learning. As Soderstorm et al. ([61], p. 558) stated "just the simple act of making such assessments may have some power to make the assessed material more memorable and thus, under the right conditions, increase students' later exam performance". Therefore, more evidence is needed to determine under which conditions eliciting OCJs in the context of ARSs is beneficial for the individual making the judgments and for the peers receiving these judgments as feedback.

6. Limitations

The study followed an in vivo design, conducting an ARS activity in the context of a course over eight weeks. This means that certain constraints had to be respected in terms of quiz content and duration. For example, the activity had to be completed within the first 15 min of a lecture, while no recap questions were used to measure retention during the weeks. Furthermore, since the activity lasted eight weeks, course attendance, and consequently activity attendance, fluctuated with several students missing one or more quiz sessions. We excluded these students from the analysis, thus making the sample sizes smaller. Finally, since our overarching goal was to improve students' experience in ARS activities, we decided to include the percentage metric as standard feedback for all students, alongside the two metacognitive judgments we examined. This provided a more realistic experience for the students as it is expected that an ARS tool that utilizes the voting/revoting paradigm would offer this information to the students.

7. Conclusions and future research

ARSs allow for advancing the activity of question-asking in the classroom by augmenting it not only with anonymity and asking everyone, but with collecting and displaying additional information about students' metacognitive judgments. These metacognitive judgements add significantly to the effectiveness of the use of an ARS and show how learners become continuously more correct and more confident in answering questions. The study compared the helpfulness and accuracy of two different types of metacognitive judgments, concurrent item-by-item OCJs, and retrospective composite RJPAs. The analysis of these judgments in assisting students when used as collective feedback in PI was based on student scores and confidence, while their accuracy in depicting student performance was based on three calibration measures: sensitivity, specificity, and absolute accuracy index. As such, this study draws from research on metacognition and ARSs and provides empirical evidence on the relation of OCJs and RJPAs in a context that is not adequately covered in the literature.

In ways of understanding and facilitating learning, this study underlines how advances in learning are intertwined, but not identical with gaining confidence and how composite feedback of metacognitive judgements may be at a more adequate grain size for students to productively calibrate their learning in such a development of understanding and confidence. The suggestion that composite retrospective judgments may be more effective can have pedagogical, practical, and design implications. First, composite judgments appear more accurate and therefore helpful in providing a better picture of the audience. Second, eliciting composite judgments may be less time-consuming than eliciting judgments for each item. And third, since composite retrospective judgments are elicited after a voting phase, they can be easier to integrate into a polling tool.

Future research could inquire how these tools could be further advanced and create scenarios of eliciting additional information from peers, integrate options for help-seeking among peers, and thereby enhance the overall learning experiences in the classroom as well as in the large lecture.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Alin Panainte for his work in developing and working with the SAGA environment.

References

[1] Allwood CM, Jonsson AC, Granhag PA. The effects of source and type of feedback on child witnesses' metamemory accuracy. Appl Cogn Psychol 2005;19(3): 331e344. https://doi.org/10.1002/acp.1071.

[2] Anderson LS, Healy AF, Kole JA, Bourne Jr LE. The clicker technique: cultivating efficient teaching and successful learning. Appl Cogn Psychol 2013;27(2):222–34.

[3] Anderson LW, Krathwohl DR. A taxonomy for learning, teaching, and assessing: a revision of bloom's taxonomy of educational objectives. New York: Addison Wesley Longman, Inc; 2001.

[4] Papadopoulos PM, Natsis A, Obwegeser N, Weinberger A. Enriching feedback in audience response systems: Analysis and implications of objective and subjective metrics on students' performance and attitudes. Journal of computer assisted learning 2019;35(2):305–16. https://doi.org/10.1111/jcal.12332.

[5] Papadopoulos, P. M., Obwegeser, N., & Weinberger, A. (2021). Let Me Explain! The Effects of Writing and Reading Short Justifications on Students' Performance, Confidence, and Opinions in Audience Response Systems. Journal of computer assisted learning. (in press).

[6] Baker RSJ, et al. Modeling and Studying Gaming the System with Educational Data Mining. In: Azevedo R., Aleven V. (eds). International handbook of metacognition and learning technologies. springer international handbooks of education. New York, NY: Springer; 2013. https://doi.org/10.1007/978-1-4419-5546-3_7. vol 28.

[7] Balta N, Michinov N, Balyimez S, Ayaz M. A meta-analysis of the effect of Peer Instruction on learning gain: identification of informational and cultural moderators. Int J Educ Res 2017;86:66–77.

[8] Barr ML. Encouraging college student active engagement in learning: student response methods and anonymity. J Comput Assist Learn 2017;33(6):621–32.

[9] Blackwell CK, Lauricella AR, Wartella E, Robb M, Schomburg R. Adoption and use of technology in early education: the interplay of extrinsic barriers and teacher attitudes. Comput Educ 2013;69:310–9.

[10] Blasco-Arcas L, Buil I, Hernández-Ortega B, Sese FS. Using clickers in class. The role of interactivity, active collaborative learning and engagement in learning performance. Comput Educ 2013;62:102–10.

[11] Boekaerts M, Rozendaal JS. Using multiple calibration measures in order to capture the complex picture of what affects students' accuracy of feeling of confidence. Learn Instr 2010;20(4):372e382. https://doi.org/10.1016/j.learninstruc.2009.03.002.

[12] Brady M, Seli H, Rosenthal J. "Clickers" and metacognition: a quasi-experimental comparative study about metacognitive self-regulation and use of electronic feedback devices. Comput Educ 2013;65:56–63.

[13] Burson KA, Larrick RP, Klayman J. Skilled or unskilled, but still unaware of it: perceptions of difficulty drive miscalibration in relative comparisons. J Pers Soc Psychol 2006;90:60–77.

[14] Butterfield B, Metcalfe J. Errors committed with high confidence are hypercorrected. J Exper Psychol 2001;27:1491–4. 10.1037//0278-7393.27.6.1491.

[15] Chen JC, Whittinghill DC, Kadlowec JA. Classes that click: fast, rich feedback to enhance students' learning and satisfaction. J Eng Educ 2010;99(2):158–69.

[16] Chien Y-T, Chang Y-H, Chang C-Y. Do we click in the right way? A meta-analytic review of clicker-integrated instruction. Educ Res Rev 2016;17:1–18.

[17] de Bruin ABH, van Gog T. Improving self-monitoring and selfregulation: from cognitive psychology to the classroom. Learn Instr 2012;22(4):245–52. https://doi.org/10.1016/j.learninstruc.2012.01.003.

[19] Dunlosky J, Rawson KA. Overconfidence produces underachievement: inaccurate self-evaluations undermine students' learning and retention. Learn Instr 2012;22(4):271–80.

[20] Dunlosky J, Rawson KA, Middleton EL. What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypothesis. J Mem Lang 2005;52:551–65.

[21] Efklides A. Metacognitive experiences in problem solving: metacognition, motivation, and self-regulation (Eds.). In: Efklides A, Kuhl J, Sorrentino RM, editors. Trends and prospects in motivation research. Dordrecht, The Netherlands: Kluwer; 2001. p. 297–323.

[22] Elicker J, McConnell N. Interactive learning in the classroom: is student response method related to performance? Teach Psychol 2011;38(3):147–50.

[23] Ertmer PA, Ottenbreit-Leftwich AT, Sadik O, Sendurur E, Sendurur P. Teachers beliefs and technology integration practices: a critical relationship. Comput Educ 2012;59(2):423–35.

[24] Feuerman M, Miller AR. Relationships between statistical measures of agreement: sensitivity, specificity and kappa. J Eval Clin Pract 2008;14(4):930–3. https://doi.org/10.1111/j.1365-2753.2008.00984.x.

[25] Finn B, Metcalfe J. The role of memory for past test in the underconfidence with practice effect. J Exper Psychol 2007;33:238–44. https://doi.org/10.1037/0278-7393.33.1.238.

[26] Finn B, Metcalfe J. Judgments of learning are influenced by memory for past test. J Mem Lang 2008;58:19–34. https://doi.org/10.1016/j.jml.2007.03.006.

[27] Fortner-Wood C, Armistead L, Marchand A, Morris FB. The effects of student response systems on student learning and attitudes in undergraduate psychology courses. Teach Psychol 2013;40(1):26–30.

[28] Ghanaat Pisheh EAZ, NejatyJahromy Y, Gargari RB, Hashemi T, Fathi-Azar E. Effectiveness of clicker-assisted teaching in improving the critical thinking of adolescent learners. J Comput Assist Learn 2019;35(1):82–8. https://doi.org/10.1111/jcal.12313.

[29] Gyllen JG, Stahovich TF, Mayer RE, Darvishzadeh A, Entezari N. Accuracy in judgments of study time predicts academic success in an engineering course. Metacogn Learn 2019;14:215–28. https://doi.org/10.1007/s11409-019-09207-6.

[30] Hoekstra A, Mollborn S. How clicker use facilitates existing pedagogical practices in higher education: data from interdisciplinary research on student response systems. Learn Media Technol 2012;37(3):303–20.

[31] Hubbard JK, Couch BA. The positive effect of in-class clicker questions on later exams depends on initial student performance level but not question format. Comput Educ 2018;120:1–12. https://doi.org/10.1016/j.compedu.2018.01.008.

[32] Hunsu NJ, Adesope O, Bayly DJ. A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. Comput Educ 2016;94:102–19.

[33] Jemstedt A, Kubik V, Jönsson Fredrik U. What moderates the accuracy of ease of learning judgments? Metacogn Learn 2017;12(3):337–55. https://doi.org/10.1007/s11409-017-9172-3.

[34] Kay RH, LeSage A. Examining the benefits and challenges of using audience response systems: a review of the literature. Comput Educ 2009;53(3):819–27.

[35] Kleitman S, Costa DSJ. The role of a novel formative assessment tool (Stats-mIQ) and individual differences in real-life academic performance. Learn Individ Differ 2014;29:150–61.

[37] Koriat A, Sheffer L, Ma'ayan H. Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. J Exper Psychol 2002;131:147–62. https://doi.org/10.1037/0096-3445.131.2.147.

[38] LaDue ND, Shipley TF. Click-on-diagram questions: a new tool to study conceptions using classroom response systems. J Sci Educ Technol 2018;27(6):492–507. https://doi.org/10.1007/s10956-018-9738-0.

[39] Lasry N, Charles E, Whittaker C. Effective variations of peer instruction: the effects of peer discussions, committing to an answer, and reaching a consensus. Am J Phys 2016;84(8):639–45. https://doi.org/10.1119/1.4955150.

[40] Lin LM, Zabrucky KM, Moore D. Effects of text difficulty and adults' age on relative calibration of comprehension. Am J Psychol 2002;115:187–98.

[41] Marshall L, Valdosta G, Varnon A. An empirical investigation of clicker technology in financial accounting principles. Learn High Educ 2012;8(1):7–17.

[42] Mazur E. Peer instruction: a user's manual series in educational innovation. Upper Saddle River, NJ: Prentice Hall; 1997.

[43] Mazur E. Farewell, lecture? Science 2009;323:50–1.

[44] McDonough K, Foote J. The impact of individual and shared clicker use on students' collaborative learning. Comput Educ 2015;86:236–49.

[45] Metcalfe J, Finn B. Hypercorrection of high confidence errors in children. Learn Instr 2012;22:253–61. https://doi.org/10.1016/j.learninstruc.2011.10.004.

[46] Michinov N, Anquetil E, Michinov E. Guiding the use of collective feedback displayed on heatmaps to reduce group conformity and improve learning in Peer Instruction. J Comput Assist Learn 2020;36(6):1026–37.

[47] Michinov N, Morice J, Ferrières V. A step further in Peer Instruction: using the Stepladder technique to improve learning. Comput Educ 2015;91:1–13.

[48] Mihalca L, Mengelkamp C, Schnotz W. Accuracy of metacognitive judgments as a moderator of learner control effectiveness in problem-solving tasks. Metacogn Learn 2017;12:1–23. https://doi.org/10.1007/s11409-017-9173-2.

[49] Nielsen KL, Hansen-Nygård G, Stav JB. Investigating peer instruction: how the initial voting session affects students' experiences of group discussion. ISRN Educ 2012;2012. article id 290157.

[50] Perez KE, Strauss EA, Downey N, Galbraith A, Jeanne R, Cooper S, et al. Does displaying the class results affect student discussion during peer instruction? CBE Life Sci Educ 2010;9:133–40.

[51] Pieger E, Mengelkamp C, Bannert M. Metacognitive judgments and disfluency-does disfluency lead to more accurate judgments, better control, and better performance? Learn Instr 2016;44:31–40.

[52] Rawson KA, Dunlosky J. Retrieval-Monitoring-Feedback (RMF) Technique for Producing Efficient and Durable Student Learning. In: Azevedo R., Aleven V. (eds). International Handbook of Metacognition and Learning Technologies. Springer International Handbooks of Education, vol 28. New York, NY: Springer; 2013. https://doi.org/10.1007/978-1-4419-5546-3_5.

[53] Schnaubert L, Bodemer D. Subjective validity ratings to support shared knowledge construction in CSCL. In: Lindwall O, Häkkinen P, Koschmann T, Tchounikine P, Ludvigsen S, editors. Exploring the Material Conditions of Learning: The Computer Supported Collaborative Learning (CSCL) Conference 2015. International Society of the Learning Sciences; 2015 *(Vol. 2)* (pp. 933-934). Gothenburg:.

[54] Schnaubert L, Bodemer D. Prompting and visualising monitoring outcomes: guiding self-regulatory processes with confidence judgments. Learn Instr 2017;49:251–62.

[55] Schnaubert L, Bodemer D. Providing different types of group awareness information to guide collaborative learning. Int J Comput-Support Collab Learn 2019;14:7–51.

[56] Schraw G. Measuring metacognitive judgments. In: Hacker DJ, Dunlosky J, Graesser AC, editors. Handbook of metacognition in education; 2009. p. 415–29.

[57] Schraw G, Kuch F, Gutierrez AP. Measure for measure: calibrating ten commonly used calibration scores. Learn Instr 2013;24:48–57. https://doi.org/10.1016/j.learninstruc.2012.08.007.

[58] Shapiro AM, Gordon L. Classroom clickers offer more than repetition: converging evidence for the testing effect and confirmatory feedback in clicker-assisted learning. J Teach Learn Technol 2013;2(1):15–30.

[59] Shapiro AM, Gordon LT. A controlled study of clicker-assisted memory enhancement in college classrooms. Appl Cogn Psychol 2012;26:635–43.

[60] Shapiro AM, Sims-Knight J, O'Rielly GV, Capaldo P, Pedlow T, Gordon L, Monteiro K. Clickers can promote fact retention but impede conceptual understanding: the effect of the interaction between clicker use and pedagogy on learning. Comput Educ 2017;111:44–59.

[61] Soderstrom NC, Clark CT, Halamish V, Bjork EL. Judgments of learning as memory modifiers. J Exper Psychol 2015;41(2):553–8.

[62] Stowell J, Oldham T, Bennett D. Using student response systems ("clickers") to combat conformity and shyness. Teach Psychol 2010;37(2):135–40.

[63] In Veenman MVJ. Learning to self-monitor and self-regulate (Eds.). In: Mayer R, Alexander P, editors. Handbook of research on learning and instruction. 2nd ed. New York: Routledge; 2017. p. 233–57.

[64] Vickrey T, Rosploch K, Rahmanian R, Pilarz M, Stains M. Research-based implementation of Peer Instruction: a literature review. CBE Life Sci Educ 2015;(1):14.

[65] Wu Y-CJ, Wu T, Li Y. Impact of using classroom response systems on students' entrepreneurship learning experience. Comput Human Behav 2019;92:643–5.

[66] Kayima F, Jakobsen A. Exploring the situational adequacy of teacher questions in science classrooms. Research in Science Education 2020;50(2):437–67. https://doi.org/10.1007/s11165-018-9696-9.

[67] Mollborn S, Hoekstra A. "A meeting of minds": Using clickers for critical thinking and discussion in large sociology classes. Teaching Sociology 2010;38:18–27.