Recommendations for Orchestration of Formative Assessment Sequences: a Data-driven Approach

Rialy Andriamiseza, Franck Silvestre, Jean-François Parmentier, Julien Broisin

Recommendations for Orchestration of Formative Assessment Sequences: a Data-driven Approach

Rialy Andriamiseza[1], Franck Silvestre Author[1][0000−0002−1134−8200], Jean-François Parmentier[2], and Julien Broisin[1][0000−0001−8713−6282]

[1] IRIT, Université de Toulouse, 118 Route de Narbonne, 31062 Toulouse cedex 9, France
[2] Toulouse INP, R4 Allée Emile Monso, 31030 Toulouse, France

Abstract. Formative assessment aims to improve teaching and learning by providing teachers and students with feedback designed to help them to adapt their behavior. To face the increasing number of students in higher education and support this kind of activity, technology-enhanced formative assessment tools emerged. These tools generate data that can serve as a basis for improving the processes and services they provide. Based on literature and using a dataset gathered from the use of a formative assessment tool in higher education whose process, inspired by Mazur's Peer Instruction, consists in asking learners to answer a question before and after a confrontation with peers, we use learning analytics to provide evidence-based knowledge about formative assessment practices. Our results suggest that: (1) Benefits of formative assessment sequences increase when the proportion of correct answers is close to 50% during the first vote; (2) Benefits of formative assessment sequences increase when correct learners' rationales are better rated than incorrect learners' ones; (3) Peer ratings are consistent when correct learners are more confident than incorrect ones; (4) Self-rating is inconsistent in peer rating context; (5) The amount of peer ratings makes no significant difference in terms of sequences benefits. Based on these results, recommendations in formative assessment are discussed and a data-informed formative assessment process is inferred.

Keywords: technology-enhanced formative assessment · learning analytics · peer instruction · decision-making

1 Introduction

Formative assessment aims to improve learning by providing teachers and students with feedback designed to help them to adapt their behavior. However, according to Andersson, formative assessment is often used in an informal and approximate way [1]. Ellis also emphasized the difficulty of capturing all learning interactions in a face-to-face context [14]. Providing practitioners and students with meaningful and effective feedback is thus a complex task, especially in large scale settings where the amount of learning interactions to capture increases with the number of learners.

To address this challenge and to support the growing number of students in higher education, Technology-Enhanced Formative Assessment (TEFA) and its interactive voting systems emerged. Such systems implement different processes offering teachers the opportunity to conduct formative assessment sequences. Among them, a group of processes, namely the "two-votes-based processes", requires learners to vote twice during the sequence. Peer Instruction, as described by Mazur [9], is one of the earliest forms of two-votes-based formative assessment processes. Basically, a two-votes-based sequence includes the following phases: (1) Teachers ask a question; (2) Students give their first answer; (3) Students reflect on peers answers and think about their own knowledge; (4) Students give their second answer to the same question; (5) Teachers discuss with students about the results. With two-votes-based processes, the number of students providing the correct answer at the fourth phase is expected to be higher than at the second phase. When this is the case, we qualify such sequence as *beneficial* because it means that students understanding of the topic has been enhanced [34].

These five phases comprise a wide variety of learning interactions. However, due to the lack of data related to two-votes-based processes [3], little work has explored how to use these interactions to bring new knowledge about formative assessment. Hence, in this paper, we address the following research questions: Which meaningful information can be inferred from the analysis of data gathered from a tool implementing a two-votes-based process and used in authentic contexts? How can such information contribute to facilitate two-votes-based process orchestration?

The three main contributions are the followings:

- findings about formative assessment, based on a dataset gathered from the use of a formative assessment tool in authentic learning contexts in higher education;
- recommendations to assist designers of formative assessment systems;
- recommendations to assist teachers when orchestrating two-votes-based sequences.

The paper is structured as follows. Section 2 introduces formative assessment and emphasizes limits of prior TEFA initiatives. Section 3 describes the formative assessment system used as the data provider of our study, as well as the dataset. Section 4 details the analysis we conducted and gives the main results. Starting from these results, Section 5 proposes an orchestration model of formative assessment sequences implementing the two-votes-based process. Section 6 discusses the limitations of our study. Section 7 concludes and discusses future work.

2 Related works

2.1 Formative Assessment

Although assessment is often used as assessment *of* learning, it can also be used as assessment *for* learning [22]. On one hand, summative assessment is used to evaluate student's level of achievement at the end of an instructional unit. On the other hand, formative assessment is crucial to make teachers able to evaluate students' understandings and adapt their lessons [11]. Hattie highlighted formative assessment as one of the most efficient methods to improve student achievement [17]. In 1998, Black and William suggested the following definition: "Formative assessment is to be interpreted as encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" [4]. This definition emphasizes the importance of collecting data to provide feedback designed to improve learning and teaching.

For instance, in face-to-face settings, Meltzer and Mannivan reported on the usage of visual artefacts (such as pieces of papers or cardboards) to allow students to answer questions asked by teachers [25]. Thanks to this feedback, teachers can collect learners' answers at a glance and adapt their teaching. However, this method hardly fits large scale educational settings since collecting and processing several answers is time consuming. Technology is then needed to collect and process interaction data efficiently, making Learning Analytics relevant for improving formative assessment.

2.2 Technology-Enhanced Formative Assessment

TEFA is one of the emerging solutions for delivering formative assessment with immediate feedback [33]. Since questioning an audience enters in the frame of formative assessment [4], Classroom Response Systems (CRS) are one of the most commonly used systems supporting TEFA in face-to-face context [2].

A generic formative assessment process of CRS is implemented by web-based platforms such as Poll Everywhere [8]. It allows teachers to ask a question, and learners to vote for the correct answer. Histograms or pie charts are then immediately displayed as feedback in order to show the distribution of votes and help teachers and learners engage in a debriefing phase. Several platforms such as Kahoot [18] support the same process. However, beyond the overview of learners' vote for the question, they propose a feedback providing teachers with the answers of each learner regarding all the formative assessment sequences she has been involved in.

Activating learners as instructional resources is an efficient way to implement formative assessment [5]. Student performance over a course of an academic programme can be significantly affected and positively influenced through a series of feedback processes handled by peers [26]. Hence, a richer formative assessment process implemented by ComPAIR [31] lets teachers ask open-ended question, while learners provide textual answers. Afterwards, learners engage in a peer review loop. They are asked to give a textual feedback about two peers answers, but also to justify why one answer is more relevant than the other. During and after this phase, teachers are provided with a feedback about each learner interaction such as her chosen answer, the textual feedback she provided, and the comparisons she submitted for the presented pair of answers.

Elaastic [13,30] and myDalite [6] offer even richer processes with even more interactions. Both systems implement the two-votes-based process illustrated in Figure 1. The processes proposed by Elaastic and myDalite consist in asking learners to vote a first time and to provide a written explanation (also called "rationale") to justify their choices. Then the process allows learners to vote a second time. At this point, both platforms differ. On one hand, myDalite allows learners to select one rationale as their second vote. Then, it provides teachers with a feedback detailing how many learners went from being wrong to right, right to wrong, wrong to wrong and right to right. On the other hand, Elaastic engages learners in a peer rating phase before they submit their second answer, as they are asked to rate several peers rationales. At anytime of the sequence, Elaastic can display first and second votes of learners and provide teachers with each learner written explanation and the mean rate attributed by peers (see Section 3.1).

This section showed that advanced technology-enhanced formative assessment processes such as two-votes-based processes offer a wide variety of interactions. Previous quantitative studies emphasized the benefits of such interactivity-rich processes [23,29,34]. Furthermore, qualitative works about the usage of a two-votes-based process emphasized learners' growing sense of self-regulation and awareness of their own explanation [6]. According to Crouch and Mazur [9] and to the ICAP framework [7], this process cognitively engages students at different levels. Finally, based on Black and William's theory of formative assessment [5], we argue that two-votes-based processes have a very satisfying coverage of formative assessment requirements [32]. Consequently, we tackle our research questions by (i) identifying hypotheses based on a review of literature, and (ii) applying various data mining techniques to evaluate these hypotheses and infer relevant information about formative assessment.

3 Design of the Dataset

We present here the formative assessment platform used for our study, together with the dataset gathered from its usage in authentic learning contexts in higher education.

3.1 Elaastic, a Technology-Enhanced Formative Assessment Tool

Elaastic is a web platform [30] used since 2015 in different higher education contexts across various disciplines such as computer science, physics or project management.

During phase 1, teachers ask learners to answer a question. If the question is closed-ended, it can be either a multiple- or exclusive-choice question. Phase 2 requires learners to answer the question and provide a written rationale to justify their choice(s). They are also asked to provide their confidence degree about their answer on a four-items Likert scale (see Figure 2). This scale has 4 items because a neutral value would be difficult to interpret [27] regarding confidence degree. Phase 3 engages learners in a peer rating activity. As shown in Figure 3, they are provided with peers' rationales or their own and are asked to evaluate each of them by reporting their level of agreement using a five-items Likert scale (1="Strongly disagree", 2="Disagree", 3="Not agree and not disagree", 4="Agree", 5="Strongly agree"). To avoid middle response bias [19], learners can also select a null response option ("I'm not giving my opinion"). Teachers can configure the number of rationales (up to 5) evaluated by each learner. Then, phase 4 begins and learners have the opportunity to vote a second time for the correct answer(s). Finally, teachers can start the phase 5. The distribution of learners scores, the rationales and their mean rate are displayed for a debriefing.

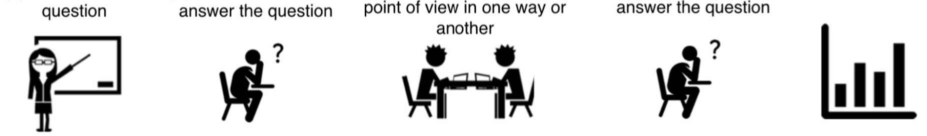
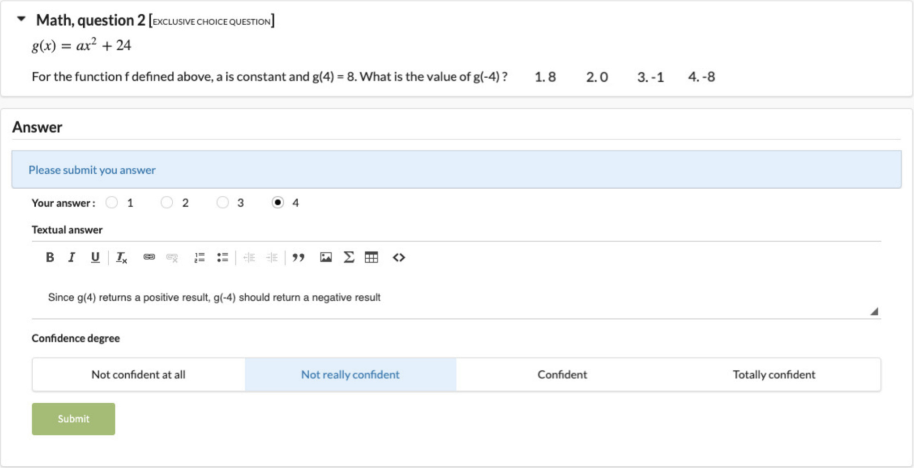

(1) Teachers ask a question
(2) First vote: Learners answer the question
(3) Learners confront their point of view in one way or another
(4) Second vote: Learners answer the question
(5) Discussion
?
?

Fig. 1. The 5 phases of the two-votes-based process.



Math, question 2 [EXCLUSIVE CHOICE QUESTION]

$g(x) = ax^2 + 24$

For the function f defined above, a is constant and g(4) = 8. What is the value of g(-4) ? 1. 8 2. 0 3. -1 4. -8

Answer

Please submit you answer

Your answer : ○ 1 ○ 2 ○ 3 ● 4

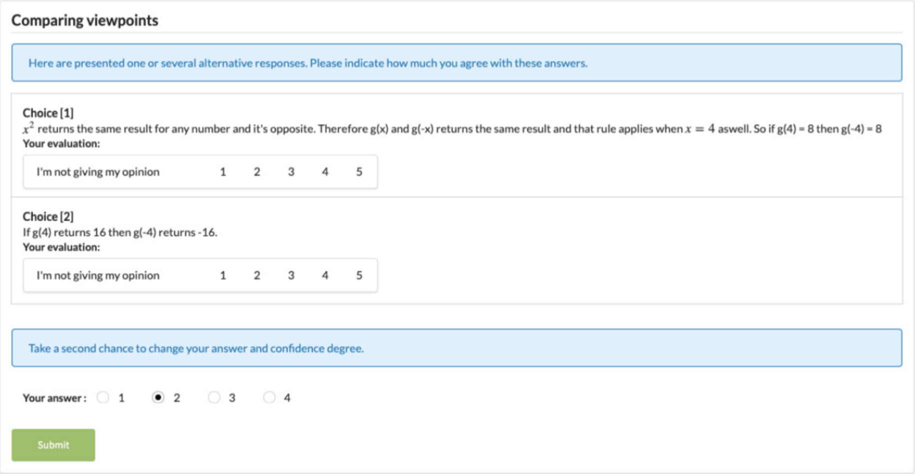
Textual answer

Since g(4) returns a positive result, g(-4) should return a negative result

Confidence degree

Not confident at all	Not really confident	Confident	Totally confident

Fig. 2. Elaastic: submission form of first vote.



Comparing viewpoints

Here are presented one or several alternative responses. Please indicate how much you agree with these answers.

Choice [1]
x^2 returns the same result for any number and it's opposite. Therefore g(x) and g(-x) returns the same result and that rule applies when $x = 4$ aswell. So if g(4) = 8 then g(-4) = 8
Your evaluation:

I'm not giving my opinion 1 2 3 4 5

Choice [2]
If g(4) returns 16 then g(-4) returns -16.
Your evaluation:

I'm not giving my opinion 1 2 3 4 5

Take a second chance to change your answer and confidence degree.

Your answer : ○ 1 ● 2 ○ 3 ○ 4

Fig. 3. Elaastic: submission form of second vote.

3.2 The Dataset

We conducted our analysis on data gathered from the use of Elaastic in higher education from 2015 to 2019. Until now, we collected 623 sequences conducted by 53 teachers where 1769 learners provided 8757 answers and performed 9256 peer ratings.

A sequence is characterised by a learning context (i.e. face-to-face, distant or hybrid), the answers of the first and second votes, as well as the number of participants. For each answer, the following data are collected: the learner identifier, the content of the rationale, the score and the selected choice(s) when applicable. If the answer is a first vote, it is characterised by additional data such as the mean grade assigned by peers to the rationale associated with the answer, and the confidence degree of the learner who provided the answer. Questions are described by their statement, their type (e.g. open ended, multiple- or exclusive-choice) and, in case of choice questions, by the number of different choices proposed to learners. Finally, for each evaluation resulting from the peer rating activity, the following data are collected: the rated rationale, the identifier of the rater, and the rate she assigned.

4 Data Analysis

The whole dataset has been filtered in order to reduce influential external factors and outliers. First, we only considered choice questions so as to be able to evaluate correctness of answers. In our analysis, in order to classify an answer as right or wrong, we considered answers as incorrect if the score is lower than the maximum score that can be obtained (i.e. 100). Also, since the asynchronous nature of distant and hybrid execution contexts in Elaastic doesn't require full orchestration from teachers [30], we kept face-to-face sequences only. Then we removed sequences where there were less than 10 participants because we wanted to focus on large scale settings. Finally, we considered the variables $p1$ and $p2$ which are the proportion of learners who answered correctly at the first and second vote respectively. Sequences where $p1 = 0$ were removed, since the confrontation can not operate under these conditions (there is no rationales for correct answers to convince incorrect peers). Sequences where $p2 = 1$ or $p1 = 1$ were removed as well, as they point out questions that were too easy to measure an effect size. After cleaning our data, we obtained 104 sequences conducted by 21 teachers where 616 learners provided 1981 answers and performed 4072 peer ratings. For our analysis, even though our sample does not follow a normal distribution of the variables, we consider it as large enough to conduct analysis with parametric tests [16].

4.1 Benefits of Sequences Increase when the Proportion of Correct Answers is Close to 50% during the First Vote

In 2001, Crouch and Mazur defined [35% - 70%] as the desired interval of $p1$ for optimal benefits of formative assessment sequences [9]. Later works suggested [30% - 80%] as the threshold values [20]. Finally, in 2010, Watkins and Mazur [23] noticed that their implementation of Peer Instruction is of high benefits for students when between 30–70% of their first answers are correct. Based on these statements, we make the hypothesis that benefits of a sequence are linked to the distance between $p1$ and 50%.

In order to verify this hypothesis, we measured the effect size between the first and second votes. To this end, we used the estimation of Cohen's effect size d proposed by Parmentier [29]: $d = 0.6ln\left(\frac{p2}{1-p2}\frac{1-p1}{p1}\right)$. Based on this estimation, we define sequences as *beneficial* when $d > 0$ (since it implies that $p1 < p2$). Figure 4a shows the mean effect size depending on the distance between $p1$ and 50%. As an example, the first bar represents 37 sequences where the distance of $p1$ to 50% is between 0% and 10%. In other words, when $p1$ is comprised between 40% (50% – 10%) and 60% (50% + 10%), the mean effect size is close to 0.4. The chart suggests that the effect size of a sequence decreases when the distance between $p1$ and 50% increases.

The Pearson correlation between $|p1 - 0.5|$ and d is -0.31 with p-value = .001 and a 95% confidence interval equal to [-0.48:-0.13], which supports our hypothesis.

The distance between $p1$ and 50% is a useful indicator to predict benefits of a two-votes-based sequence. In other words, benefits of peer interactions are maximized when correct and incorrect answers are equally represented. We argue that too few correct answers may indicate that learners lack understanding or knowledge to engage in productive discussions, whereas too many correct answers may indicate that the question is too easy and does not require discussions.

Recommendations for system designers: Formative assessment systems implementing a two-votes-based process should provide teachers with the proportion of correct answers at the first vote. They should also feature flexibility regarding the way to conduct the sequence, especially according to the proportion of correct answers at the first vote and its distance to 50%.

As Lasry stated [21], the threshold values of the ideal percentages of correct answers are indicative. In our context, the interval that best suits our result is [20%-80%]. Indeed, Figure 4a suggests that when $p1$'s distance from 50% is greater than 30%, the effect size is significantly lower.

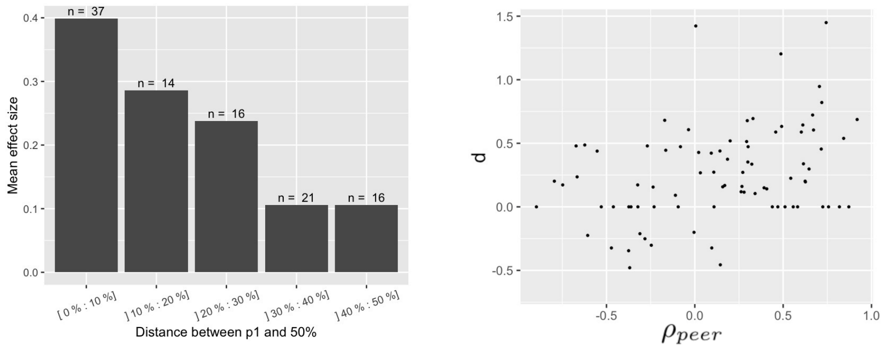
Recommendations for orchestration: If there are too few correct answers at the first phase ($p1 < 20\%$), teachers should either provide detailed explanations and restart the sequence, or provide learners with hints before engaging learners in a confrontation phase. If there are a lot of correct answers ($p1 > 80\%$), teachers can interrupt the sequence and provide learners with a brief explanation. ❶

4.2 Benefits of Sequences Increase when Peer Ratings are Consistent

Double & al. argue that reflecting on peers answers is expected to lead to a higher percentage of correct answers [12]. Since correct learners are expected to convince incorrect learners, we make the hypothesis that the consistency of the peer rating phase is linked to the sequence benefits.

In order to measure the consistency of peer ratings in a sequence, we used ρ_{peer} which is the correlation between the level of agreement given by peers to a rationale, and the correctness of the matching answers (self-rating included). Since these two variables are latent [15], the polychoric correlation is the adequate tool [28]. More precisely, ρ_{peer} will tend to be close to 1 if the rationales matching with correct answers are positively evaluated by peers, whereas those matching with incorrect answers are negatively evaluated. Conversely, ρ_{peer} will tend to be close to -1 if the rationales matching with incorrect answers are better evaluated than those matching with correct answers. Figure 4b shows a plot diagram of the effect size d depending on ρ_{peer}.

The Pearson correlation between ρ_{peer} and d is 0.34 with a p-value < .002 and a 95% confidence interval equal to [0.14:0.52], which supports our hypothesis. Let us note that ρ_{peer} is not significantly correlated to the distance between $p1$ and 50% (p-value = 0.25). Consequently, this subsection and subsection 4.1 identified two independent predictors of the benefits of a sequence, namely ρ_{peer} and $|p1 - 50\%|$.



(a) The effect size d depending on the distance between $p1$ and 50%.

(b) The effect size d depending on the consistency of peer ratings ρ_{peer}.

Fig. 4. d given $|p1 - 50\%|$ and ρ_{peer}.

When $\rho_{peer} < 0$, it means that incorrect answers are more popular than correct answers which should be addressed by teachers.

Recommendations for system designers: Formative assessment systems implementing a peer rating process should provide teachers with the consistency of peer rating and feature flexibility regarding the selection of the rationales in the focus of the discussion (phase 5), especially according to the consistency of peer rating.

Recommendations for orchestration: If peer rating is inconsistent ($\rho_{peer} < 0$), teachers should focus on incorrect rationales during the discussion. Else ($\rho_{peer} \geq 0$), teachers should focus on correct rationales during the discussion. 2

4.3 Peer Ratings are Consistent when Learners Confidence Degrees are Consistent

Back to the first vote, Curtis used the confidence of learners about their answers as a way to identify misinformed learners [10]. More precisely, he defined misinformed learners as confident but incorrect learners. Starting from this research, we propose an indicator to measure the consistency of learners confidence degree given the correctness of their answers. Since correct learners are expected to be more confident than incorrect learners, we believe that misinformed learners are not able to consistently rate peers rationales. As a consequence, we make the hypothesis that consistency of peer ratings is linked to the consistency of learners confidence degree.

Similarly to ρ_{peer}, confidence consistency ρ_{conf} can be computed by using the polychoric correlation between learners confidence degree and correctness of their first answers. If correct learners are confident whereas incorrect ones aren't, ρ_{conf} will tend to be close to 1 . Conversely, if incorrect learners are confident whereas correct ones aren't, ρ_{conf} will tend to be close to -1 . Figure 5a is a plot diagram of ρ_{peer} according to ρ_{conf}. The Pearson correlation between ρ_{conf} and ρ_{peer} is 0.38 with a p-value $< 4e-4$, and a 95% confidence interval equal to [0.18:0.55], which supports our hypothesis.

Recommendations for system designers: Formative assessment systems implementing a two-votes-based process should provide teachers with the consistency of learners confidence degree. They should also feature flexibility regarding the way to conduct the sequence, but also regarding the selection of the rationales in the focus of the discussion (phase 5) according to this consistency.

ρ_{conf} is an adequate measure of learners understanding of the concept targeted by the question. Beyond learners correctness, their confidence degree allows teachers to obtain more precise feedback, including the proportion of misinformed learners (incorrect but confident) and lucky learners (correct but not confident). Similarly to ρ_{peer}, when $\rho_{conf} < 0$, it means that incorrect answers are more popular than correct answers. This may indicate that some misconceptions need to be addressed by teachers.

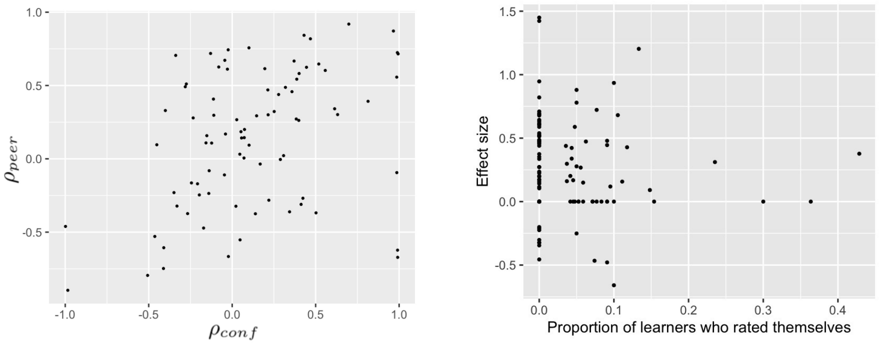
Recommendations for orchestration: When there are too many correct answers in the first vote, teachers should focus the discussion on incorrect rationales if learners are inconsistently confident ($p1 > 80\%$ and $\rho_{conf} < 0$), and on correct rationales if learners are consistently confident ($p1 > 80\%$ and $\rho_{conf} < 0$). When there are too few correct answers, teachers should provide detailed explanations and restart the sequence if learners are inconsistently confident ($p1 < 20\%$ and $\rho_{conf} < 0$). If learners are consistently confident ($p1 < 20\%$ and $\rho_{conf} > 0$), teachers should provide learners with hints before starting the confrontation phase. ❸

4.4 Self-Rating is Inconsistent in Peer Rating Contexts

Regarding peer interactions-related factors, some studies about self-rating [24,12] provide support for its use as a formative practice to improve academic performances. Consequently, we make the hypothesis that there is a relationship between the number of self-rated students and the benefits of a sequence.

Our results suggest that self-rating tends to nullify the effect size (see Figure 5b). We explored the data and found out that learners who rated themselves during the confrontation of viewpoints tend to give their rationale the highest grade whether they where correct or not. We compared grades given when learners rated themselves with grades given when learners rated peers (see Figure 6). The difference in means was significant (95% CI = [-1.68:-1.014] and p-value $< 10e-11$). Furthermore, self-rating was less consistent ($\rho_{self_r} = 0.139$) than peer rating ($\rho_{peer_r} = 0.219$).

This result rejects our hypothesis and suggests that self-rating does not benefit learners within peer rating contexts. An informal discussion with 9 learners has been conducted and allowed us to make three hypotheses. First, learners stated that they logically agree with themselves. This implies that they do not revise their own answer based on peers rationales as expected. Second, learners know that rationales with the highest



(a) Peer rating consistency ρ_{peer} depending on the confidence consistency ρ_{conf}.

(b) Effect size d depending on the percentage of learners who rated themselves.

Fig. 5. ρ_{peer} given ρ_{conf} and d given the proportion of self-grades

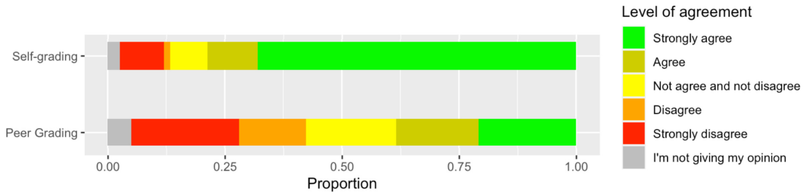


Fig. 6. Stacked bar chart of the grade attributed depending on the type of rating.

grades are more likely to be noticed. Therefore, learners game the system in order to receive oral feedback from teachers during phase 5. Third, learners perceive this activity as competitive and, therefore, want to obtain the highest mean grade.

Recommendation for system designers: Peer rating activities in formative assessment systems should not include self-rating.

4.5 The Amount of Peer Ratings Makes no Significant Difference in Terms of Sequences Benefits

Group discussion in formative assessment is a challenging task. Depending on the context (e.g. the physical location of learners or the nature of the course), different ways to confront learners' viewpoints can be found in literature. Some implementation paired learners with their neighbour in classes [34], whereas others involved teachers in the collective discussion [35]. Therefore, we want to explore the impact of the number of learners involved in group discussions. With Elaastic, the number of learners involved in group discussion is represented by the number of peers rationales rated by each learner. We believe that the effect size depends on such a number.

Since there were not enough sequences with 1 and 4 rates given, we ran a t-test with various grouping methods (see Table 1). According to our result, the number of learners involved in peer interactions has no significant impact, which rejects our hypothesis.

Table 1. Results of the two sample t-test with various grouping methods.

Group 1			Group 2			two sample t-test	
nb rates given	mean	sd	nb rates given	mean	sd	95% CI	p-value
1, 2	0.18	0.39	3	0.26	0.42	[-0.3 : 0.13]	0.42
1, 2	0.18	0.39	4, 5	0.29	0.34	[-0.31 : 0.08]	0.25
3	0.26	0.42	4, 5	0.29	0.34	[-0.2 : 0.14]	0.73
1, 2	0.18	0.39	3, 4, 5	0.28	0.38	[-0.09 : 0.29]	0.29
1, 2, 4, 5	0.25	0.36	3	0.26	0.42	[-0.17 : 0.15]	0.88
1, 2, 3	0.23	0.41	4, 5	0.29	0.34	[-0.2 : 0.09]	0.44

Recommendation for system designers: Formative assessment systems should feature flexibility regarding the number of peers involved in group confrontation.

Recommendation for orchestration: Teachers can decide the number of peers involved in group confrontation.

5 Resulting Orchestration Model

Figure 7 summarises our recommendations for orchestration of formative assessment sequences. The presented model is derived from Vickrey's model designed to support orchestration of Peer Instruction [36]. When sequences are not beneficial, deep and detailed explanations are needed from teachers during the oral feedback. Consequently, we added the following recommendation to our model:

Recommendation for orchestration: After the second vote, teachers explanation should be more detailed if the proportion of correct answers did not increase ($d \leq 0$). 4

6 Limitations

Our main limitations come from the dataset itself. The 104 sequences that we analysed addressed mainly STEM topics from higher education classes. A more refined study of sequences from various topics and educational levels could lead to broader findings.

In the context of multiple choice answers, if a learner obtains a score of 33/100 during the first vote and 66/100 during the second vote, both her answers are considered as wrong, and the information stating that she improved is lost. Even though multiple

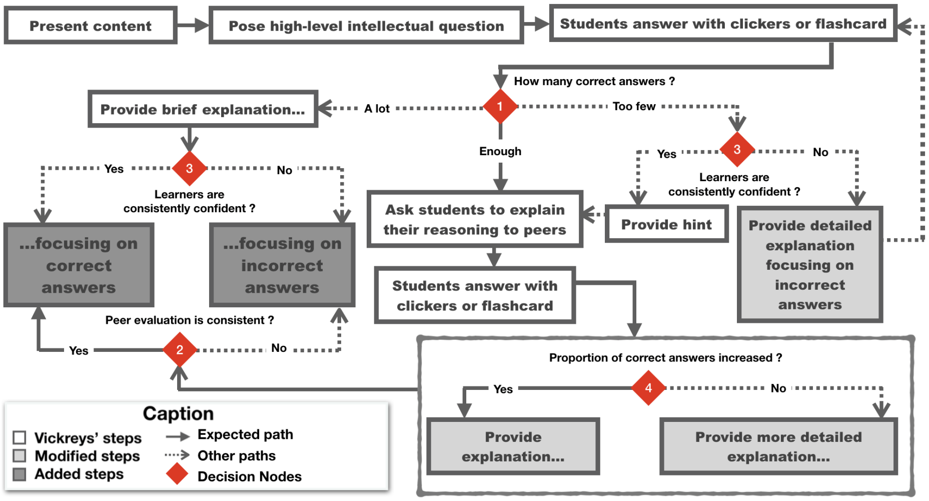

Present content
Pose high-level intellectual question
Students answer with clickers or flashcard
How many correct answers ?
1
A lot
Too few
Enough
Provide brief explanation...
3
Yes
No
Learners are consistently confident ?
...focusing on correct answers
...focusing on incorrect answers
Peer evaluation is consistent ?
2
Yes
No
Ask students to explain their reasoning to peers
Students answer with clickers or flashcard
3
Yes
No
Learners are consistently confident ?
Provide hint
Provide detailed explanation focusing on incorrect answers
Proportion of correct answers increased ?
4
Yes
No
Provide explanation...
Provide more detailed explanation...
Caption
□ Vickreys' steps
▨ Modified steps
■ Added steps
→ Expected path
⇢ Other paths
◆ Decision Nodes

Fig. 7. Orchestration model of two-votes-based processes based on [36]. Each white number represents the matching recommendation for orchestration.

choice questions are only a small portion of our sample (~10%), a deeper study addressing this distinction would be a more adequate way to refine our results.

Moreover, as stated earlier, Elaastic does not capture all learning interactions in a face-to-face context, thus making us unable to identify every decisive aspects of a formative assessment sequence such as its context (i.e. the subjects and themes of the questions) as well as oral and informal interactions between learners and teachers.

Finally, we consider rationales associated to correct answers as correct rationales. However, learners can answer correctly and provide incorrect rationales. As an example, if learners give a low rate to an incorrect rationale corresponding to a correct answer, ρ_{peer} will decrease even though this rationale was rightfully given a low rate. Such a possibility is not addressed by our works regarding the quality of peer interactions.

7 Conclusion and Future Works

This paper focused on formative assessment and emphasized the challenge of its application in face-to-face contexts. We introduced TEFA as the solution that emerged to perform face-to-face formative assessment and also introduced rich formative assessment processes generating a lot of meaningful interactions. Based on literature and on a dataset gathered from the usage of a two-votes-based process in an authentic learning context, we proposed to study these interactions to (i) highlight new understandings of formative assessment; (ii) provide system designers with evidences intended to help them to design a formative assessment system; (iii) identify meaningful indicators to assist teachers when orchestrating a face-to-face formative assessment sequence.

Future works will implement our orchestration model within Elaastic while taking in account the explainability issues regarding our indicators. After the first vote, teachers will receive textual description to help them make decisions regarding the next phase to engage. After the second vote, teachers will be provided with recommended learners' rationale to address during the discussion phase. Then, we will measure this evolution's impact on teaching and learning thanks to a qualitative and quantitative analysis.

References

1. Andersson, C., Palm, T.: The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. Learning and Instruction **49**, 92–102 (Jun 2017)
2. Beatty, I.D., Gerace, W.J.: Technology-enhanced formative assessment: A research-based pedagogy for teaching science with classroom response technology. Journal of Science Education and Technology **18**(2), 146–162 (2009)
3. Bhatanagar, S., Zouaq, A., Desmarais, M.C., Charles, E.: A dataset of learnersourced explanations from an online peer instruction environment. International Educational Data Mining Society **13**, 350–355 (2020)
4. Black, P., Wiliam, D.: Assessment and Classroom Learning. Assessment in Education: Principles, Policy & Practice **5**(1), 7–74 (Mar 1998)
5. Black, P., Wiliam, D.: Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education) **21**(1), 5 (2009)
6. Charles, E.S., Lasry, N., Bhatnagar, S., Adams, R., Lenton, K., Brouillette, Y., Dugdale, M., Whittaker, C., Jackson, P.: Harnessing peer instruction in-and out-of class with mydalite. In: Education and Training in Optics and Photonics. p. 11143_89. Optical Society of America, SPIE, Quebec City, Quebec, Canada (2019)
7. Chi, M.T.H., Wylie, R.: The icap framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist **49**(4), 219–243 (Oct 2014)
8. Clark, S.: Enhancing active learning: Assessment of poll everywhere in the classroom. Tech. rep., University of Manitoba (2017)
9. Crouch, C.H., Mazur, E.: Peer instruction: Ten years of experience and results. American journal of physics **69**(9), 970–977 (2001)
10. Curtis, D.A., Lind, S.L., Boscardin, C.K., Dellinges, M.: Does student confidence on multiple-choice question assessments provide useful information? Medical education **47**(6), 578–584 (2013)
11. Davis, M.: Technology fed growth in formative assessment. Education Week p. 11 (2015)
12. Double, K.S., McGrane, J.A., Hopfenbeck, T.N.: The impact of peer assessment on academic performance: A meta-analysis of control group studies (2020)
13. Elaastic: `https://elaastic.irit.fr`, last consulted: 2021-06-25.
14. Ellis, C.: Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics. British Journal of Educational Technology **44**(4), 662–664 (2013)
15. Everett, B.: An introduction to latent variable models. Springer Science & Business Media (2013)
16. Ghasemi, A., Zahediasl, S.: Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism **10**(2), 486 (2012)
17. Hattie, J.: Visible learning for teachers: Maximizing impact on learning. Routledge, 711 Third Avenue, New York, NY 10017 (2012)
18. Ismail, M.A.A., Mohammad, J.A.M.: Kahoot: A promising tool for formative assessment in medical education. Education in Medicine Journal **9**(2), 19–26 (2017)

19. Kulas, J.T., Stachowski, A.A., Haynes, B.A.: Middle response functioning in likert-responses to personality items. Journal of Business and Psychology **22**(3), 251–259 (2008)
20. Lasry, N.: Clickers or flashcards: Is there really a difference? The Physics Teacher **46**(4), 242–244 (2008)
21. Lasry, N., Mazur, E., Watkins, J.: Peer instruction: From harvard to the two-year college. American journal of Physics **76**(11), 1066–1069 (2008)
22. Martinez, M.E., Lipson, J.I.: Assessment for learning. Educational Leadership **46**(7), 73–75 (1989)
23. Mazur, E., Watkins, J.: Just-in-time teaching and peer instruction. In: Just-in-time Teaching: Across the Disciplines, Across the Academy, pp. 39–62. Stylus Publishing, LLC, 22883 Quicksilver Drive, Sterling, Virginia 20166-2102 (2010)
24. McMillan, J.H., Hearn, J.: Student self-assessment: The key to stronger student motivation and higher achievement. Educational Horizons **87**(1), 40–49 (2008)
25. Meltzer, D.E., Manivannan, K.: Transforming the lecture-hall environment: The fully interactive physics lecture. American Journal of Physics **70**(6), 639–654 (2002)
26. Montebello, M., Pinheiro, P., Cope, B., Kalantzis, M., Amina, T., Searsmith, D., Cao, D.: The impact of the peer review process evolution on learner performance in e-learning environments. In: Proceedings of the Fifth Annual ACM Conference on Learning at Scale. pp. 1–3. ACM, London, UK (2018)
27. Muijs, D.: Doing quantitative research in education with SPSS. Sage Publications (2004)
28. Olsson, U.: Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika **44**(4), 443–460 (1979)
29. Parmentier, J.F.: How to quantify the efficiency of a pedagogical intervention with a single question. Physical Review Physics Education Research **14**(2), 020116 (2018)
30. Parmentier, J.F., Silvestre, F.: La (dé-)synchronisation des transitions dans un processus d'évaluation formative exécuté à distance : impact sur l'engagement des étudiants. In: 9ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH 2019). pp. 97–108. ATIEF, Sorbonne Universite, LIP6Paris, France (2019)
31. Potter, T., Englund, L., Charbonneau, J., MacLean, M.T., Newell, J., Roll, I., et al.: Compair: A new online tool using adaptive comparative judgement to support learning with peer feedback. Teaching & Learning Inquiry **5**(2), 89–113 (2017)
32. Silvestre, F.: Conception et mise en oeuvre d'un système d'évaluation formative pour les cours en face à face dans l'enseignement supérieur. Ph.D. thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier (2015)
33. Spector, J.M., Ifenthaler, D., Sampson, D., Yang, J.L., Mukama, E., Warusavitarana, A., Dona, K.L., Eichhorn, K., Fluck, A., Huang, R., et al.: Technology enhanced formative assessment for 21st century learning. International Forum of Educational Technology and Society **19**(3), 58–71 (2016)
34. Tullis, J.G., Goldstone, R.L.: Why does peer instruction benefit student learning? Cognitive Research: Principles and Implications **5**(1), 15 (Dec 2020)
35. Turpen, C., Finkelstein, N.D.: Not all interactive engagement is the same: Variations in physics professors' implementation of peer instruction. Physical Review Special Topics - Physics Education Research **5**(2), 020101 (Aug 2009)
36. Vickrey, T., Rosploch, K., Rahmanian, R., Pilarz, M., Stains, M.: Research-based implementation of peer instruction: A literature review. CBE—Life Sciences Education **14**(1), es3 (Mar 2015)