Data-informed Decision-making in TEFA Processes: An Empirical Study of a Process Derived from Peer-Instruction

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

Data-informed Decision-Making in TEFA Processes: An Empirical Study of a Process Derived from Peer-Instruction

Rialy Andriamiseza
University of Toulouse, IRIT
Toulouse, France
rialy.andriamiseza@irit.fr

Franck Silvestre
University of Toulouse, IRIT
Toulouse, France
franck.silvestre@irit.fr

Jean-François Parmentier
INP-ENSEEIHT
Toulouse, France
jf.parmentier@gmail.com

Julien Broisin
University of Toulouse, IRIT
Toulouse, France
julien.broisin@irit.fr

ABSTRACT
When formative assessment involves a large number of learners, Technology-Enhanced Formative Assessments are one of the most popular solutions. However, current TEFA processes lack data-informed decision-making. By analyzing a dataset gathered from a formative assessment tool, we provide evidence about how to improve decision-making in processes that ask learners to answer the same question before and after a confrontation with peers. Our results suggest that learners' understanding increases when the proportion of correct answers before the confrontation is close to 50%, or when learners consistently rate peers' rationales. Furthermore, peer ratings are more consistent when learners' confidence degrees are consistent. These results led us to design a decision-making model whose benefits will be studied in future works.

Author Keywords
technology-enhanced formative assessment, learning analytics, peer instruction, decision-making

CCS Concepts
•**CCS** → Applied computing → Education → E-learning;

INTRODUCTION
To address formative assessment challenges and the growing number of students in higher education, Technology-Enhanced Formative Assessment (TEFA) and its interactive response systems emerged. Such systems implement different processes allowing teachers to conduct formative assessment sequences. Among them, a group of processes, namely the "two-votes-based processes", requires learners to provide an answer before and after a confrontation with peers. However, these various processes lack evidence to help teachers make decisions depending on learners' behavior. In this paper, we address the following research question: Which meaningful information for teachers can be inferred from the analysis of data gathered from a tool implementing a two-votes-based process, to improve decision-making in face-to-face formative assessment sequences? We tackle this question by (i) identifying hypotheses based on literature, and (ii) applying various data mining techniques to evaluate these hypotheses and infer relevant information about decision-making in formative assessment sequences.

L@S '21, June 22–25, 2021, Virtual Event, Germany.

ACM ISBN 978-1-4503-8215-1/21/06.
http://dx.doi.org/10.1145/3430895.3460153

RELATED WORKS

Formative assessment
Formative assessment aims to improve learning by providing teachers and students with feedback designed to help them adapt their behavior. In 1998, Black and William defined formative assessment as: "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" [1]. This definition highlights the importance of collecting data in order to make decisions during teaching. However, Ellis emphasized the difficulty of capturing learning interactions in a face-to-face context [8], especially when the number of learners increases. Collecting and analysing interaction data make learning analytics a relevant field for formative assessment. More precisely, the involvement of technology is needed so as to effectively capture learning interactions and thus help teachers conduct a formative assessment sequence.

Technology-Enhanced Formative Assessment
TEFA is one of the emerging solutions for delivering formative assessment with immediate feedback [17]. Since questioning an audience enters in the frame of formative assessment [1], Classroom Response Systems (CRS) are the most commonly used systems supporting TEFA in face-to-face context. A generic formative assessment process of CRS is implemented by web-based platforms such as Poll Everywhere [4] which

simply allows teachers to ask a question, and learners to vote for the correct answer(s). Histograms are then immediately displayed as feedback in order to help teachers and learners engage in a debriefing phase. A richer formative assessment process implemented by ComPAIR [15] lets teachers ask open-ended question, while learners provide textual answers. Afterwards, learners engage in a peer review loop. Elaastic [16] and myDalite [2] offer even richer processes with even more interactions. Both systems stand on Mazurs' Peer Instruction [5] and implement a two-votes-based process illustrated in Figure 1. More precisely, they ask user to answer an exclusive choice question and to provide a textual justification (also called rationale). However, when it comes to step 3 and 4, Elaastic engages learners in a peer rating phase before they submit their second answers.

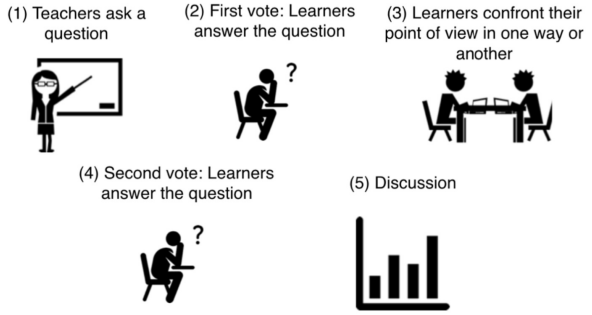

Figure 1. The 5 phases of the two-votes-based process.

Even though quantitative studies [12, 14, 18] and the ICAP framework [3] emphasized the benefits of such interactivity-rich processes , these 5 phases might not always be the best choice to orchestrate the sequence. Alternative options should be considered by teachers depending on learners' behavior and understanding.. To the best of our knowledge, Mazur's recommendation to skip phases 3 and 4 when less than 30% or more than 70% of learners' first answers are correct [12] is the only recommendation that can be found in literature (with few variations [19, 11, 5]). On the basis of a dataset gathered from an authentic usage of Elaastic in higher education from 2015 to 2019, the remaining of the paper explores which information can be used to help teachers driving a FA sequence.

CONTENT OF THE DATASET

With Elaastic, a sequence is characterised by a learning context (i.e. face-to-face, distant or hybrid), learners' answers for the first and second votes, as well as the number of participants. For each answer, the following data are collected: the learner's 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 level of agreement of peers (self-reported on a 5-items Likert scale) to the rationale, and the confidence degree of the learner who provided the answer (self-reported on a 4-items Likert scale). Questions are described by their statement and their type (e.g. open ended, multiple- or exclusive-choice). Finally, for each evaluation resulting from the confrontation phase, the following data are collected: the rated rationale, the identifier of the rater, and the rate she assigned.

Data Analysis

The whole dataset has been filtered in order to keep only relevant data for our study. First, we only considered choice questions so as to evaluate correctness of answers. 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). Then we removed sequences with less than 10 participants because we wanted to focus on larger 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 there is no rationales for correct answers to convince incorrect peers) as well as sequences where $p1 = 1$ or $p2 = 1$ (as they point out questions that were too easy to measure an impact). We finally obtained 104 sequences conducted by 21 teachers where 616 learners provided 1981 answers and performed 4072 peer ratings. Even though this sample doesn't follow a normal distribution, it is large enough to conduct analysis with parametric tests [10].

Benefits of formative assessment sequences increase when the proportion of correct answers is close to 50%

Based on Mazur's statement about the proportion of correct answers [12], we make the hypothesis that sequence benefits increase as the distance between $p1$ and 50% decreases. 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 [14] and calculated as follows: $d = 0.6ln\left(\frac{p2}{1-p2}\frac{1-p1}{p1}\right)$. Based on this estimation, we define sequences as *beneficial* when $d > 0$ (which implies that $p1 < p2$) because it means that students understanding of the topic has been enhanced [18]. As suggested by Figure 2, the mean effect size decreases when the distance between $p1$ and 50% increases. The calculated

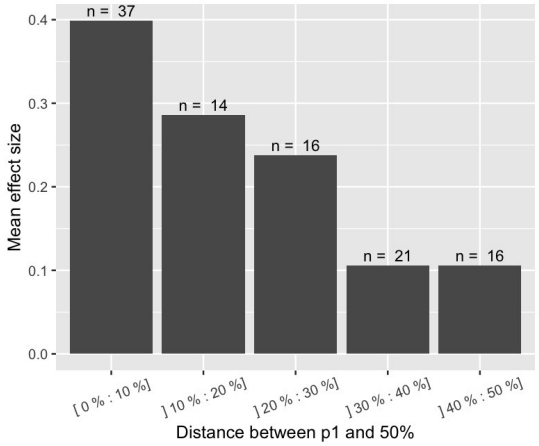

Figure 2. The effect size d depending on $p1$'s distance to 50%.

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

Based on our results, the recommended interval for $p1$ should be [20%-80%] because when $p1$'s distance from 50% is

greater than 30%, the effect size is significantly lower (< 0.2).

1 Recommendations: If there are too few correct answers ($p1 < 20\%$) teachers should provide detailed explanations and restart the sequence, or provide hints before peer interactions. If there are enough correct answers ($20\% \leq p1 \leq 80\%$), learners can engage in a peer interaction phase. Else ($p1 > 80\%$), teachers can provide brief explanations and end the sequence.

Furthermore, the effect size d serves as a good indicator to determine how well learners understood the topic and how detailed teachers' explanation should be.

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

Benefits of formative assessment sequences increase when learners consistently rate peers' rationales

Double and al. argue that reflecting on peers answers is expected to lead to a higher percentage of correct answers [7]. Since correct learners are expected to convince incorrect learners, we make the hypothesis that sequence benefits increase alongside peer rating consistency. Consistency of peer ratings ρ_{peer} can be computed by using the correlation between the level of agreement of a learner to a peer's rationale with the correctness of his answer. Since both these variables are latent [9], the polychoric correlation is the adequate one [13]. In other words, ρ_{peer} indicates if the rationales matching with correct answers are positively evaluated by peers, and if those matching with incorrect answers are negatively evaluated. Figure 3 shows a plot diagram of the effect size d depending on ρ_{peer}. The calculated Pearson correlation between ρ_{peer}

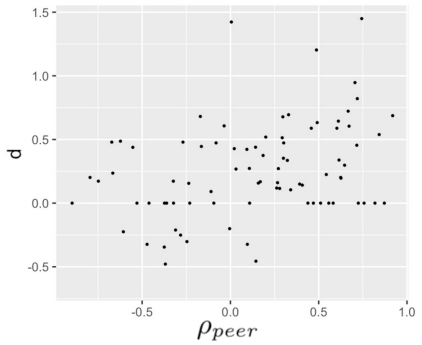

Figure 3. The effect size d depending on consistency of peer ratings ρ_{peer}.

and d is 0.34 with a p-value lower than .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, we identified two independent predictors of the benefits of a sequence. When $\rho_{peer} < 0$, it means that incorrect answers were better rated than correct ones which must be addressed by teachers.

3 Recommendations: If peer ratings are inconsistent ($\rho_{peer} < 0$), teachers should focus on incorrect rationales during the oral feedback. Else ($\rho_{peer} \geq 0$), teachers should focus on correct rationales.

Peer ratings are more consistent when learners' confidence degrees are consistent

Back to the first vote, Curtis used confidence of learners to identify misinformed learners [6]. As a consequence, we make the hypothesis that the consistency of peer ratings increases alongside the consistency of learners' confidence degrees.Confidence consistency ρ_{conf} can be computed by using the polychoric correlation [13] between learners confidence degree and correctness of their first answers. Figure 4 is a plot diagram of ρ_{conf} according to ρ_{peer}. The calculated Pearson correlation between ρ_{conf} and ρ_{peer} is 0.38 with a p-value lower than $4e-4$, and a 95% confidence interval equal to [0.18:0.55], which supports our hypothesis. This result suggests that the consistency of learner confidence degree is correlated to the consistency of peer's rates. When

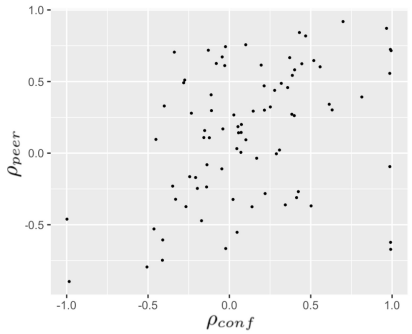

Figure 4. The consistency of peer rating ρ_{peer} depending on the confidence consistency ρ_{conf}.

$\rho_{conf} < 0$, it means that incorrect answers are more popular than correct answers. Depending on the result of the first vote ($p1$), this information can determine either the next steps to engage in, or the answers that need to be addressed by teachers.

4 Recommendations: When there are too many correct answers ($p1 > 80\%$), if learners are inconsistently confident ($\rho_{conf} < 0$), teachers should focus their brief explanations on incorrect rationales. Else ($\rho_{conf} \geq 0$), teachers should focus their brief explanations on correct rationales. When there a too few correct answers ($p1 < 20\%$), if learners are inconsistently confident ($\rho_{conf} < 0$), teachers should provide detailed explanations and restart the sequence. Else ($\rho_{conf} \geq 0$), teachers should provide hints before engaging learners in a peer interaction phase.

DISCUSSION AND FUTURE WORKS

Thanks to our findings, we can extend Vickrey's model of two-votes-based processes [19] and propose an orchestration model as shown in Figure 5. Future works will implement and evaluate this experimental model within Elaastic.

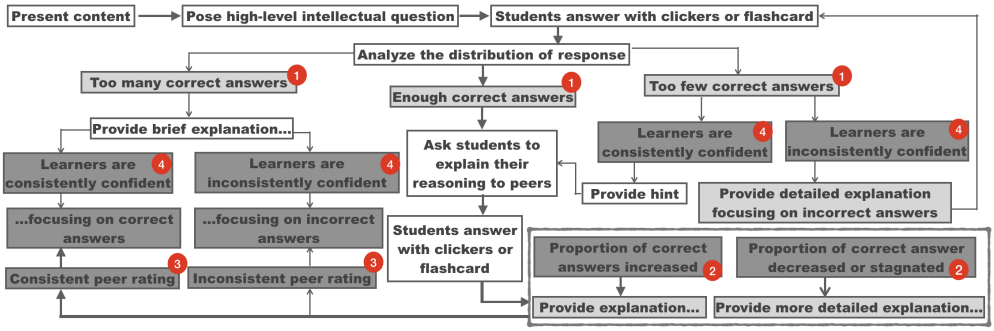

Figure 5. A decision-making model of two-votes-based process. White steps are from the original model [19]. Light grey steps are steps that we modified. Dark grey steps are steps that we added. The numbers refer to the corresponding recommendations and bold arrows represent the nominal case.

REFERENCES

[1] Paul Black and Dylan Wiliam. 1998. Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice* 5, 1 (March 1998), 7–74.

[2] Elizabeth S Charles, Nathaniel Lasry, Sameer Bhatnagar, Rhys Adams, Kevin Lenton, Yann Brouillette, Michael Dugdale, Chris Whittaker, and Phoebe Jackson. 2019. Harnessing peer instruction in-and out-of class with myDALITE. In *Education and Training in Optics and Photonics*. Optical Society of America, SPIE, Quebec City, Quebec, Canada, 11143_89.

[3] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (Oct 2014), 219–243.

[4] Sarah Clark. 2017. *Enhancing Active Learning: Assessment of Poll Everywhere in the Classroom*. Technical Report. University of Manitoba.

[5] Catherine H Crouch and Eric Mazur. 2001. Peer instruction: Ten years of experience and results. *American journal of physics* 69, 9 (2001), 970–977.

[6] Donald A Curtis, Samuel L Lind, Christy K Boscardin, and Mark Dellinges. 2013. Does student confidence on multiple-choice question assessments provide useful information? *Medical education* 47, 6 (2013), 578–584.

[7] Kit S Double, Joshua A McGrane, and Therese N Hopfenbeck. 2020. The impact of peer assessment on academic performance: A meta-analysis of control group studies. (2020).

[8] Cath Ellis. 2013. Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics. *British Journal of Educational Technology* 44, 4 (2013), 662–664.

[9] B Everett. 2013. *An introduction to latent variable models*. Springer Science & Business Media.

[10] Asghar Ghasemi and Saleh Zahediasl. 2012. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism* 10, 2 (2012), 486.

[11] Nathaniel Lasry, Eric Mazur, and Jessica Watkins. 2008. Peer instruction: From Harvard to the two-year college. *American journal of Physics* 76, 11 (2008), 1066–1069.

[12] Eric Mazur and Jessica Watkins. 2010. Just-in-time teaching and peer instruction. In *Just-in-time Teaching: Across the Disciplines, Across the Academy*. Stylus Publishing, LLC, 22883 Quicksilver Drive, Sterling, Virginia 20166-2102, 39–62.

[13] Ulf Olsson. 1979. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* 44, 4 (1979), 443–460.

[14] Jean-François Parmentier. 2018. How to quantify the efficiency of a pedagogical intervention with a single question. *Physical Review Physics Education Research* 14, 2 (2018), 020116.

[15] Tiffany Potter, Letitia Englund, James Charbonneau, Mark Thomson MacLean, Jonathan Newell, Ido Roll, and others. 2017. ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry* 5, 2 (2017), 89–113.

[16] Franck Silvestre, Philippe Vidal, and Julien Broisin. 2015. Reflexive learning, socio-cognitive conflict and peer-assessment to improve the quality of feedbacks in online tests. In *Design for Teaching and Learning in a Networked World*. Springer, Toledo, Spain, 339–351.

[17] J Michael Spector, Dirk Ifenthaler, Demetrios Sampson, Joy Lan Yang, Evode Mukama, Amali Warusavitarana, Kulari Lokuge Dona, Koos Eichhorn, Andrew Fluck, Ronghuai Huang, and others. 2016. Technology enhanced formative assessment for 21st century learning. *International Forum of Educational Technology and Society* 19, 3 (2016), 58–71.

[18] Jonathan G. Tullis and Robert L. Goldstone. 2020. Why does peer instruction benefit student learning? *Cognitive Research: Principles and Implications* 5, 1 (Dec 2020), 15.

[19] Trisha Vickrey, Kaitlyn Rosploch, Reihaneh Rahmanian, Matthew Pilarz, and Marilyne Stains. 2015. Research-Based Implementation of Peer Instruction: A Literature Review. *CBE—Life Sciences Education* 14, 1 (Mar 2015), es3.