



HAL
open science

Learning Analytics-Based Formative Assessment Recommendations for Technology-Enhanced Learning Practices

Rialy Andriamiseza

► **To cite this version:**

Rialy Andriamiseza. Learning Analytics-Based Formative Assessment Recommendations for Technology-Enhanced Learning Practices. Education. Université Paul Sabatier - Toulouse III, 2022. English. NNT: 2022TOU30186 . tel-04006037

HAL Id: tel-04006037

<https://theses.hal.science/tel-04006037v1>

Submitted on 27 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE

**En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE
Délivré par l'Université Toulouse 3 - Paul Sabatier**

**Présentée et soutenue par
Rialy ANDRIAMISEZA**

Le 25 novembre 2022

**Recommandations fondées sur les Learning Analytics pour les
pratiques d'évaluation formative assistée par la technologie.**

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et
Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Julien BROISIN et Franck SILVESTRE

Jury

M. Sébastien GEORGE, Rapporteur
Mme Elise LAVOUÉ, Rapporteuse
M. Philippe DESSUS, Examineur
M. Julien BROISIN, Directeur de thèse
M. Franck SILVESTRE, Co-directeur de thèse
Mme Vanda LUENGO, Présidente

Résumé

L'évaluation formative est un levier pour améliorer l'apprentissage et l'enseignement. La fourniture de feedback aux étudiants et aux enseignants dans le but d'atteindre les objectifs d'apprentissage est au cœur de tout processus d'évaluation formative. Dans les contextes d'enseignement de masse, des systèmes technologiques sont apparus pour soutenir la mise en œuvre de processus d'évaluation formative. Ces systèmes génèrent des données qui peuvent servir de base pour améliorer ces processus et les services qu'ils offrent. Par conséquent, nous adressons les questions de recherche suivantes :

- Quelles informations utiles peut-on obtenir à partir de l'analyse d'un jeu de données collecté via l'utilisation en contexte réel d'un système d'évaluation formative ?
- Comment peut-on exploiter ces informations pour aider les enseignants à orchestrer leurs séquences d'évaluation formative ?

À partir de la littérature et d'un jeu de données collecté via l'usage d'un système d'évaluation formative nommé *Elaastic*, nous mobilisons les learning analytics pour apporter des connaissances sur les pratiques d'évaluation formative. Ces connaissances nous ont permis de concevoir (i) des recommandations pour les concepteurs de systèmes d'évaluation formative (ii) des recommandations pour les enseignants qui orchestrent des séquences d'évaluation formative (iii) un modèle d'orchestration conçu pour aider les enseignants à prendre des décisions tout au long de la séquence.

Par la suite, Nous testons ce modèle d'orchestration en l'implantant dans *Elaastic* par le moyen de recommandations explicables et en collectant des données de son utilisation. L'analyse de ces données montre (1) que les enseignants ne suivent pas les recommandations et (2) que si ces recommandations avaient été suivies, les séquences auraient été significativement plus bénéfiques pour les apprenants. Les travaux futurs proposent des améliorations et extensions possibles de ce modèle d'orchestration.

Abstract

Formative assessment is a useful teaching method for improving learning and teaching. Providing teachers and learners with feedback designed to help them reach the learning objectives is at the core of every formative assessment processes. To conduct large scale formative assessment, technology-enhanced formative assessment systems emerged to support the usage of formative assessment processes. These systems generate data that can serve as a basis for improving these processes and services they provide. Consequently, we tackle the following research questions:

- Which useful information can be inferred from the analysis of data gathered from a tool implementing formative assessment processes used in authentic contexts?
- How can such information contribute to improve formative assessment processes orchestration?

Based on literature and using a dataset gathered from the use of a formative assessment tool named Elaastic, we use learning analytics to provide evidence-based knowledge about formative assessment practices. This knowledge led us to design (i) recommendations for system designers of formative assessment tools (ii) recommendations for teachers orchestration of formative assessment sequences (iii) an orchestration model to assist teachers decision-making during the sequence.

Afterwards, we put this orchestration model to the test by implementing it within Elaastic through explainable recommendations and collecting data of its usage. The analysis of these data provides evidences that show that (1) teachers do not follow the recommendations and (2) if teachers had followed them, there would be significantly improved benefits for learners. Future works discuss the way our orchestration model could be improved and expanded to other contexts.

Acknowledgement

First and foremost, I would like to convey my sincere gratitude to those who helped me make the decision to become a PhD student. It all begins with Celia Picard with whom I discussed a lot. She explained to me what being a PhD student is about and how it works. Getting to know the stakes and outcomes of a thesis was at the core of my concerns so thank you for your time and your wisdom. The next step was to find a PhD offer and this is where my former teacher Jean-Baptiste Raclet helped me. He sent me a PhD offer that seemed to be tailored for me, which I will never be thankful enough for.

Once I began my PhD thesis, I met a lot of likeable people such as Anis Bey with whom I played some basketball. I thank all my other colleagues as well for the good moments we had at the office: Mar Pérez-Sanagustín, Azzedine Benabou, Cathy Pons, Esteban Villalobos, Esther Felix, Denis Olivier. Above anyone else, the people I spent most time with are the two I shared an office with, namely Louis Sablayrolles and Mika Pons. I would like to thank both of them in particular for the great company and for having to deal with my not so funny puns. Sharing nice moments with others during my thesis was very important to me since I spent more than a year working from home due to Covid.

When it comes to work itself, I want to thank Jean-François Parmentier for his precious help regarding my works, especially when it comes to statistical analysis. I also thank a lot of other people who helped me conduct my research such as John Tranier, Gilles Schaeck, Erik Martin-Dorel and Frederic Migeon. However, no one helped me more than my PhD directors Julien Broisin and Franck Silvestre. I would like to thank both of them for their availability and advices. Furthermore I want to thank them for giving me enough freedom in my research so that I would never feel pressured to work on something I do not like. They found the perfect balance between authority and empathy, which helped me work in an emotionally secure environment. I look up to both of them and hope I will become like them in the future, both on the human and professional aspects.

Beyond my research, I would like to thank Christelle Chaudet who helped me a lot during my teaching activities. Given the subject of my thesis, such activities were crucial and helped me a lot during my research. One could not hope for a better person to enter the world of education as a teacher. She supported me on a human and professional level a lot, which I am very thankful for.

Having the courage and confidence to enter the intricate world of research would not have been possible if it was not for my parents and sister who gave

me a unwavering support. More generally, I would like to thank all the other people in my life who helped me be a better and happier person. Spending good times with friends and family is something I will never be thankful enough for and it helped me a lot to relieve the pressure and doubts that come with being a PhD student. But more than anyone and anything, the person I would like to thank is my lover and soon to be wife Julia Ehamelo. No matter how I would come home, feeling doubtful, delighted or chatty about my work, she was always patient enough and found the perfect balance between supporting me and pressuring me to perform better.

Contents

Résumé	II
Abstract	III
Acknowledgement	IV
Introduction	1
1 Introduction	3
1.1 Scientific Context	3
1.1.1 Research Unit	3
1.1.2 Research Team	5
1.1.3 The B4MATIVE! project	6
1.1.4 Formative assessment and learning analytics	6
1.2 Research questions and contributions	7
1.2.1 Research questions	7
1.2.2 Contributions	8
1.2.3 Publications	11
1.3 Organization of the manuscript	12
I State of the Art	13
2 Formative Assessment and Feedback	15
2.1 Formative Assessment	15
2.2 Feedback	16
3 TEFA Processes	21
3.1 Questioning the Class	21
3.2 Peer Confrontation Processes	24
3.3 The Two-votes-based Processes	27
3.3.1 Overall Description	27
3.3.2 Benefits of Two-Votes-Based Processes	30

3.3.3	Limits of Two-Votes-Based Processes	34
II	Research method	37
4	Research method	39
4.1	From Limits to Research Questions	39
4.2	Design-Based Research	40
4.3	The Cycle of Learning Analytics	41
III	First Iteration: Empirical study	45
5	Elaastic, Our Formative Assessment Tool	47
5.1	Overview of Elaastic	48
5.2	Learner perspective	48
5.2.1	Phase 1: Well-argued Response	48
5.2.2	Phase 2: Comparing Viewpoints	49
5.2.3	Phase 3: Results	49
5.3	Teacher perspective	50
5.3.1	Orchestrating the Sequence	50
5.3.2	Retrieving the Question and Answers of the Sequence	51
5.4	Internal mechanics	53
6	Data Analysis	55
6.1	Data Collection	56
6.1.1	Description of the dataset	56
6.1.2	Description of the sample	57
6.2	Findings	58
6.2.1	Basis for our Analysis: Measuring the Benefits of a Sequence	58
6.2.2	Proportion of Correct Answers during the First Vote	61
6.2.3	Learners Confidence Degree during the First Vote	63
6.2.4	Peer Grading Phase and Sequence Benefits	67
6.2.5	Peer Grading Phase and Learner's Confidence Degree during the First Vote	68
6.2.6	Self-Grading as a Substitute for Peer Grading	70
6.2.7	Peer Grading per Learner as Configured by Teachers	72
6.3	Discussion	73
6.3.1	Regarding the Research Questions	73

6.3.2	Implications for Research and Practice: Improving Vickrey's model	74
6.3.3	Limitations	75
6.4	Conclusion of the Empirical Study	76
IV Second Iteration: Evaluation of the New Orchestration Model		77
7	Implementation	79
7.1	Enabling The New Orchestration Model	79
7.1.1	Implemented Orchestration Features	80
7.1.2	Providing Teachers with Relevant Indicators	82
7.2	Implementation of Explainable Recommendations	85
7.2.1	Content of the Explanations	85
7.2.2	Presentation of the Explanations	89
7.3	Action Tracking	90
7.4	Additional documentation	91
8	Evaluation	93
8.1	Expected Values of the Orchestration Model	94
8.2	Data Collection	96
8.2.1	Description of the Dataset	96
8.2.2	Description of the Sample	97
8.3	Data Analysis	98
8.3.1	Teachers and Recommendations	98
8.3.2	Comparing s_{2022} with expected values	100
8.3.3	Comparing s_model_{2022} with expected values	101
8.4	Verifying the Findings from the First Iteration	103
8.4.1	Proportion of Correct Answers during the First Vote	103
8.4.2	Learners Confidence Degree and Proportion of Correct Answers during the First Vote	104
8.4.3	Peer Grading Phase and Sequence Benefits.	105
8.4.4	Peer Gradings Phase and Learner's Confidence Degree during the First Vote	105
8.5	Discussion	108
Conclusion		109
9	Conclusion	111

9.1	Summary	111
9.2	Future Works	112
9.2.1	Short Term: New Version of Elaastic	112
9.2.2	Mid-term: Expansion of the new process	113
9.2.3	Long term: Additional exploratory studies	113
Bibliography		114
Appendix		135
A	Asynchronous Settings	135
A.1	Execution Contexts	135
A.2	The Cold Start Problem	136
A.3	Alternative to ρ_{conf} and ρ_{peer}	136
B	Relevancy of Rationales	139
B.1	Cosine Similarity	139
B.2	Early Results	139

List of Tables

6.1	Summary of the 104 sequences included in the dataset we analysed.	59
6.2	Results of the two-sample test with various grouping methods of our sequences (* with parametric t-test).	73
7.1	Explanations at the end of the first vote.	87
7.2	Explanations for the discussion phase.	88
7.3	Format of the actions saved.	91
8.1	Summary of our 118 sequences per topic.	98
8.2	Contingency table of recommendations regarding phase 2, compared to teachers' final decision.	100
B.1	Average cosine of rationales ordered from lowest to highest. . .	140

List of Figures

1.1	New Formative Assessment Process	11
2.1	Usage of cardboards for formative assessment purposes [100].	18
3.1	Poll Everywhere: Feedback displaying the distribution of votes	22
3.2	Overview of learners short written responses to a physics related question [24].	23
3.3	Poll Everywhere: Feedback displaying the individual votes.	23
3.4	Peer Studio: Learner’s screen during the review process. The screenshot shows (1) the rubric, (2) the student draft, (3) an example of excellent work to compare student work against (4) automatically generated tips for commenting [85].	24
3.5	Juxtapeer: An example of a review form [28].	25
3.6	Juxtapeer: Visual representation of the review process [28].	26
3.7	ComPAIR: Feedback displaying the individual performances of learners [117].	26
3.8	The 5 steps of a two-votes-based process.	27
3.9	MyDalite: Viewpoints confrontation and second vote [29].	30
3.10	MyDalite: Learners’ response switching [29].	31
3.11	Percentage of correct answers before and after the confrontation step for various studies.	32
3.12	Vickrey’s model of Peer Instruction.	35
4.1	The <i>Learning Analytics</i> cycle [41]	41
4.2	Organisation of our Works Based on the <i>Learning Analytics Cycle</i>	42
5.1	Elaastic’s implementation of a two-votes-based process	48
5.2	Phase 1 of Elaastic: Well-argued Response.	49
5.3	Phase 2 of Elaastic: Comparing Viewpoints.	50
5.4	Phase 3 of Elaastic: Results.	51
5.5	Elaastic: Teacher’s perspective to manage a sequence	52

5.6	Pairing Algorithm: Matching example with 2 evaluations per learner.	54
6.1	Summary of the dataset.	56
6.2	Summary of the filtering criteria applied to our initial sample. The blue bar represent the 104 sequences of the final sample for our data analysis.	58
6.3	s_{2019} compared to isovalues of the Cohen's d effect size: no effect ($d = 0$), small ($d = 0.2$), medium ($d = 0.5$), large ($d = 0.8$), and very large ($d = 1.2$).	60
6.4	The effect size d of the sequences from s_{2019} depending on the distance between $p1$ and 50% (with 95% confidence intervals).	62
6.5	Mean effect size d of the sequences from s_{2019} depending on whether $p1$ is inside the interval or not, alongside the results of the Wilcoxon rank sum test.	63
6.6	The effect size d depending on ρ_{conf} . Each point represents a sequence from s_{2019}	64
6.7	Kernel density estimates and mean of the effect size (d) of our sequences depending on the confidence consistency (ρ_{conf}) and the proportion of correct answers at the first vote ($p1$).	65
6.8	Effect size d depending on the consistency of peer gradings ρ_{peer} . Each point is a sequence from s_{2019}	68
6.9	Peer grading consistency ρ_{peer} depending on the confidence consistency ρ_{conf} . Each point is a sequence from s_{2019}	69
6.10	Effect size d depending on the percentage of learners who graded themselves. Each point is a sequence from s_{2019}	70
6.11	Stacked bar chart of the grade attributed during our sequences depending on the type of grading.	71
6.12	Orchestration model of two-votes-based processes based on [155]. Each white number represents the matching recommendation for orchestration.	74
7.1	Button to skip phase 2 and directly enter into phase 3 after the end of phase 1.	80
7.2	New set of rationales to display the recommended rationales.	81
7.3	Displaying the writer's confidence degrees on each rationales	81
7.4	Design to illustrate $p1$	82
7.5	Design to illustrate d	82
7.6	Initial design to illustrate ρ_{conf}	83
7.7	Initial design to illustrate ρ_{peer}	83
7.8	Final design to illustrate ρ_{conf}	84

7.9	Final design to illustrate ρ_{peer} .	84
7.10	Histograms added to the results frame of the teacher interface.	85
7.11	Recommendation: 1st level of detail for the explanations	89
7.12	Recommendation: 2nd level of detail for the explanations.	89
7.13	Recommendation: 3rd level of detail for the explanations.	90
7.14	Simplified xAPI statement schema.	91
7.15	Help button on the left menu.	92
7.16	Online documentation.	92
8.1	Outcome of sequences for s_{2019} and s_model_{2019} .	95
8.2	Kernel densities of s_{2019} and s_model_{2019} .	96
8.3	Summary of the dataset of Elaastic 5.0	97
8.4	Summary of the filtering criteria applied to our dataset. The blue bar represents the 118 sequences of the final sample for our data analysis.	98
8.5	Our new sample compared to isovalues of the Cohen's d effect size: no effect ($d = 0$), small ($d = 0.2$), medium ($d = 0.5$), large ($d = 0.8$), and very large ($d = 1.2$). Each point is a sequence from s_{2022} .	99
8.6	Outcome of sequences from s_{2022} compared to expected values.	101
8.7	Kernel densities of the effect size d of s_{2022} compared to expected values.	102
8.8	Outcome of sequences from s_model_{2022} compared to expected values.	103
8.9	Kernel densities of the effect size d of s_model_{2022} compared to expected values.	104
8.10	The effect size d of the sequences from s_{2022} depending on the distance between $p1$ and 50% with 95% confidence intervals. The last bar has too few sequences to compute a meaningful confidence interval.	105
8.11	Kernel density estimates and mean of the effect size (d) of our sequences depending on the confidence consistency (ρ_{conf}) and the proportion of correct answers at the first vote ($p1$).	106
8.12	Effect size d depending on the consistency of peer gradings ρ_{peer} . Each point is a sequence from s_{2022} .	106
8.13	Peer grading consistency ρ_{peer} depending on the confidence consistency ρ_{conf} for s_{2022} . Each point is a sequence from s_{2022} .	107
9.1	New design of recommendations for teachers.	113
A.1	Elaastic: Configuration popup before teachers start a sequence	135

A.2 l_{conf} depending on l_{peer} from s_{2019} 137

Introduction

Chapter 1

Introduction

Contents

1.1 Scientific Context	3
1.1.1 Research Unit	3
1.1.2 Research Team	5
1.1.3 The B4MATIVE! project	6
1.1.4 Formative assessment and learning analytics	6
1.2 Research questions and contributions	7
1.2.1 Research questions	7
1.2.2 Contributions	8
1.2.3 Publications	11
1.3 Organization of the manuscript	12

1.1 Scientific Context

1.1.1 Research Unit

This manuscript presents the synthesis of three years of work done in the context of a thesis that started on September the 1st of 2019 under the supervision of Julien BROISIN and Franck SILVESTRE.

To conduct my research, I worked at the IRIT¹, one of the largest Joint Research Units (UMR 5505) at the national level, which is one of the pillars of research in Occitania with its 600 members, permanent and non-permanent,

¹Institut de Recherche en Informatique de Toulouse

and a hundred external collaborators. Because of its multi-tutelage (CNRS², Toulouse Universities), its scientific impact and its interactions with other fields, the laboratory is one of the structuring forces of the landscape of computer science and its applications in the digital world, both at the regional and national levels. The IRIT has been able, through its advanced work and its dynamics, to define its identity and to acquire an undeniable visibility, while positioning itself at the heart of the evolutions of the local structures: University of Toulouse, as well as the various devices resulting from the investments of future (LabEx CIMI, IRT Saint-Exupéry, SAT TTT...). Research conducted at the IRIT is structured around five main scientific topics:

- Design and construction of systems (reliable, safe, adaptive, distributed, communicating, dynamic...)
- Digital modeling of the real world
- Concepts for cognition and interaction
- Study of autonomous systems that adapt to their environment
- Transformation of raw data into intelligible information

Six strategic application areas materialize such research:

- Health, Autonomy, Wellness
- Smart City
- Aeronautics, Space, Transportation
- Digital social media and information distribution
- e-Education
- Cybersecurity, Security of goods and people

Finally, these research are materialized by one strategic action which is "Computing, Data, AI". The spectrum of this research is wide. It allows the unit to be a key player in theoretical and applied research in Data Science and Computation, which find their application in everyday life, changing the practices in terms of data dynamics, access to knowledge and decision support, at the heart of the digital world. From an organizational point of view, this unit is structured in 7 research departments that group the 24 teams

²Centre Nationale de Recherche Scientifique

of the laboratory. The research department that I joined is named ICI³. Within it, I became part of the research team TALENT⁴ which focuses on technology-enhanced learning.

1.1.2 Research Team

The 3 main axis of research conducted by TALENT are the followings:

- **Engineering of Technology-Enhanced Learning Environment to support active learning.** These works focus on the design and implementation of digital environments to enhance learning, but also on their evaluation in authentic learning situations. The objective is to bring knowledge to support active pedagogical methods designed to maximize learners' engagement in the learning process. In particular, this axis aims to promote deep learning through the use of visualizations and reflection tools, methods such as peer instruction and confrontation, or exploratory approaches based on the development of practical activities.
- **Learning analytics for personalized and self-regulated learning and teaching.** The main objective of this axis is to study how learning data can be exploited to support teaching and learning processes. Innovative methods and mechanisms are explored to design technological solutions with educational benefits for teachers and learners, where methods and techniques from big data research fields are used to provide effective tools and systems. The objective is to (1) collect user experience in digital educational environments; (2) apply data mining techniques (such as process/pattern mining for example) to identify learning behaviors that promote learner success; and (3) implement adaptive tools to guide teachers and learners through tasks based on knowledge inferred from behavioral data mining.
- **Digital skills and development of future education.** This axis studies the skills that teachers, students and institutions should promote to adapt to the future challenges of education. The type of research questions posed in this axis are intended to promote the acquisition and development of 21st century skills, as well as the way educational institutions should transform to achieve this.

³Intelligence Collective, Interaction

⁴Teaching And Learning ENhanced by Technology

1.1.3 The B4MATIVE! project

When it comes to the thesis itself, it is part of the project B4MATIVE! [121] which is the result of a collaboration between the TALENT team and the Académie de Nancy-Metz funded by the French DNE⁵. This project aims at designing, deploying and evaluating an innovative digital environment for delivering effective formative assessments. To this end, the Académie offers an experiment field for research, whereas our team provides teachers with a digital tool dedicated to formative assessment. More precisely, in this thesis, we focus on supporting teachers by providing them with tools that are helpful to conduct effective formative assessment sequences.

1.1.4 Formative assessment and learning analytics

Being a teacher and conducting formative assessment sequences are more and more difficult. The main reason for that is the ever-growing amount of learners in face-to-face classrooms. TEL⁶ represents an opportunity to help teachers in these tasks. More precisely, the use of technology within formative assessment practices generates a big amount of data that are an opportunity to capture a lot of learning interactions and provide immediate feedback to both learners and teachers. However, prior works on technology-based formative assessment do not include advanced analysis of the variety of data collected during these learning activities. Such advanced analysis falls into the scope of *Learning Analytics*. Learning Analytics is a fairly young field of research that originated from other data-driven fields [52] such as business intelligence, web analytics, academic analytics, educational data mining, and action analytics [55]. The big data era allows various industries and fields to grow significantly. This applies to education as well [44] which increasingly incorporated more devices and technological systems in the learning process. As defined by the Society for Learning Analytics Research (SoLAR) [93],

”Learning Analytics are the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.”

Formative Assessment and Learning Analytics are often paired together as compatible fields [145], one of the main reason being that they both serve the goal of continuous improvement of learning. More precisely, the cycle of Learning Analytics and Heritage’s statement about formative assessment

⁵Direction du Numérique pour l’Education

⁶Technology-Enhanced Learning

being a systematic process to continuously gather evidences about learning [67], make the association relevant and meaningful. Learning Analytics hold the potential to (i) explain unexpected learning behaviors (ii) identify successful learning patterns (iii) detect misconceptions and misplaced effort (iv) introduce appropriate interventions and (v) increase users' awareness of their own actions and progress [93] which is highly compatible with strategies of formative assessment centered around feedback.

Even though Learning Analytics might greatly improve teaching and learning, data-based decision-making as opposed to intuition-based decision making is at the core of users' and researchers' concerns. People have mixed opinions about fairness and usefulness of automated decision-making at a societal level, with general attitudes influenced by individual characteristics [11]. Trusting a device is essential when users are pressured by external constraints such as time limits and results requirements [71]. Although this is the case in education, few studies addressed explainability in such contexts [3] where finding the balance between trusting a device and trusting a human intuition is a challenging task [133]. This has been shown by a previous study in an elementary school which suggests that teachers tend to prioritize their own intuition [153]. In summary, the usage of technology and data for education is promising yet challenging. The aim of our works is to include more technology in teaching to assist practitioners while still accounting for the crucial role of the human factor.

1.2 Research questions and contributions

1.2.1 Research questions

This manuscript starts by arguing that technology is a solution to the challenges that teachers face when conducting large scale formative assessment. Consequently, it goes through numerous Technology-Enhanced Formative Assessment processes that can be found in literature and focuses on a family of process that we name *two-votes-based processes*. Basically, a two-votes-based sequence ask students to give their answer to a question before and after they confront their viewpoint with each other. More details on these processes are given in Section 3.3. More precisely, related works about two-votes-based processes are mentioned. Even though they emphasize the benefits of such processes for learners, they also mention their limits such as the likelihood of undesired outcomes and the lack of available data about their usage. This leads to a lack of studies that propose advanced feedback for teachers to assist them when orchestrating two-votes-based sequences. As a

consequence, we address the following research questions:

- RQ1 - Which useful information can be inferred from the analysis of data gathered from a tool implementing formative assessment processes used in authentic contexts?
- RQ2 - How can such information contribute to improve formative assessment processes orchestration?

1.2.2 Contributions

To answer these questions, we analyse data issued from instances of a formative assessment system named *Elaastic*. The process that it proposes belongs to the two-votes-based processes family. More precisely, we conduct Design-Based Research by going through two iterations of the Learning Analytics cycle described in Section 4.3. Regarding RQ1, two contributions are inferred from this thesis. The first one includes findings about formative assessment:

- Within two-votes-based formative assessment sequences, benefits of sequences increase when close to 50% of learners' first votes are correct;
- Sequences are beneficial to learners when peers can accurately grade their peers rationales, i.e., when they give high grades to rationales for correct answers and low grades to rationales for incorrect answers.
- Benefits of sequences do not significantly increase when learners who provided correct answers are more confident than learners who did not;
- Learners can accurately grade their peers rationales when they are accurately confident, i.e., when learners who provided a correct answer reported a high confidence degree whereas those who provided an incorrect answer reported a low confidence degree.
- Self-grading is inaccurate in peer grading context;
- The amount of evaluations each learner performs within peer grading activities makes no significant differences in terms of sequences benefits.

We provide evidences that support these findings for secondary and higher education contexts.

The second contribution for RQ1 is a set of Recommendations for System Designers (RSD) of formative assessment tools.

RSD-1: 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 especially according to such consistency.

RSD-2: Formative assessment systems implementing a peer grading activity should provide teachers with the consistency of peer grading and feature flexibility regarding the selection of the rationales in the focus of the discussion, especially according to the consistency of peer grading.

RSD-3: Formative assessment systems implementing a two-votes-based process should feature flexibility regarding the selection of the rationales in the focus of the discussion according to the consistency of confidence degree.

RSD-4: Peer grading activities in formative assessment systems should not include self-grading.

RSD-5: Formative assessment systems should feature flexibility regarding the number of peers involved in confrontation of viewpoints.

Two other contributions provide answers to RQ2. First, based on computed indicators (namely $p1$, d , ρ_{conf} and ρ_{peer}) that we designed and describe in Chapter 6, we provide Recommendations for Teachers Orchestration (RTO) of two-votes-based sequences:

RTO-1: After the second vote, explanations provided by teachers should be more detailed if the proportion of correct answers decreased or stagnated between the first and second vote. ❶

RTO-2: At the end of the first vote:

- If there are more than 80% of correct answers, teachers should skip the confrontation phase. ❷
- When there are less than 30% of correct answers:
 - if learners who provided correct answers are more confident than the others, teachers should provide learners with hints before starting the confrontation phase.

- Else, teachers should provide detailed explanations and restart a new sequence to evaluate the same concept.

③

RTO-3: If rationales for incorrect answers are better graded than the others, teachers should focus on such rationales during the discussion. Else, teachers should focus on correct rationales during the discussion.

④

RTO-4: When the confrontation step was skipped as well as the second vote step:

- If learners who provided a correct answer are more confident than the others, teachers should focus on rationales for incorrect answers during the discussion.
- Else, teachers should focus on rationales for correct answers during the discussion. ⑤

RTO-5: Teachers can decide the number of peers involved in viewpoints confrontation.

Second, based on these findings and on Vickrey’s model [155], a new and deterministic data-informed formative assessment process has been inferred as shown in Figure 1.1.

Afterwards, we implement such a model within Elaastic by providing teachers with feedback in the form of explainable recommendations and collect new data to measure its impact on learners. The results suggest that such feedback do not improve benefits for learners. A deeper analysis provided evidences that show that this is due to teachers not following the recommendations. By ignoring teachers final decision but only accounting for our model’s decisions, we conducted an analysis that suggest that our model improve the outcome and benefits of formative assessment sequences. Therefore, future works will focus on improving the presentation of our feedback and recommendations.

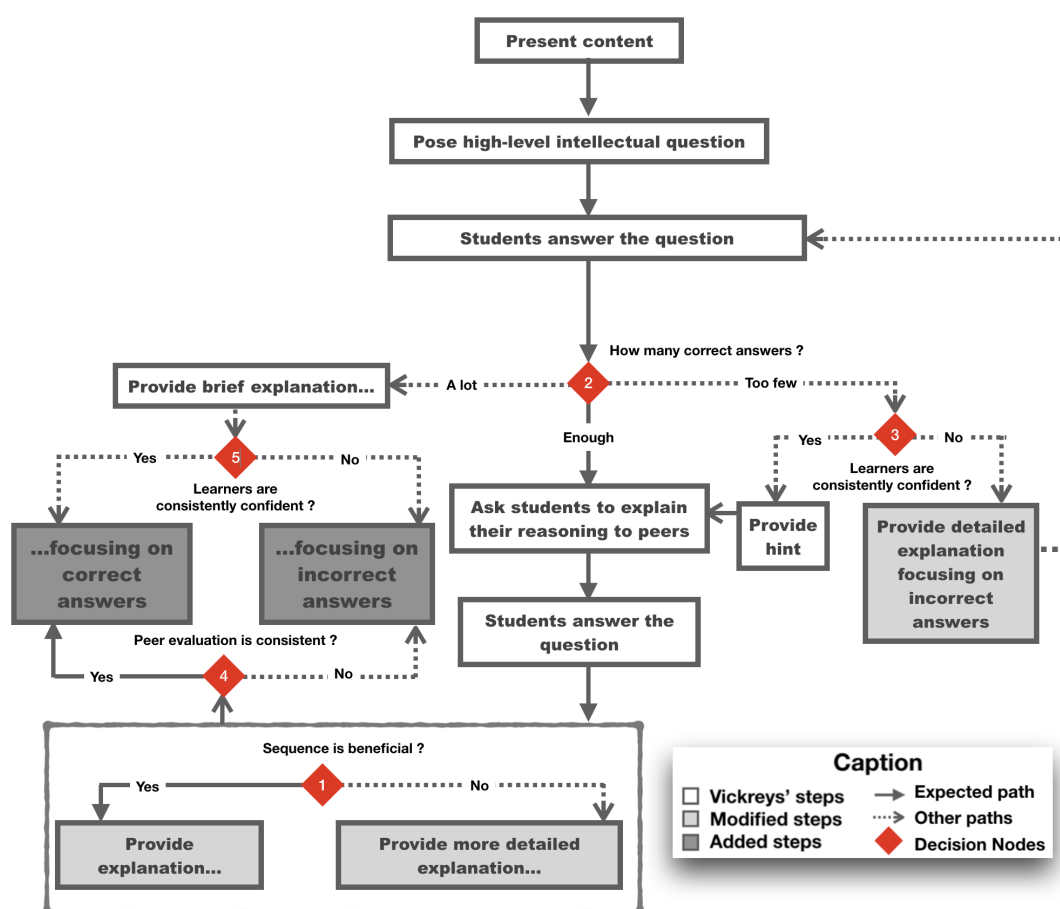


Figure 1.1: New Formative Assessment Process

1.2.3 Publications

Five papers were published in the span of these three years. The conference ranks mentioned below are the ones given by the ATIEF⁷.

- Andriamiseza, R. (2022, May). Implanter un processus d'évaluation formative : le défi de l'explicabilité. Accepted for the national B ranked RJC conference of 2022. **This paper was nominated for best paper award.**
- Andriamiseza, R., Silvestre, F., Parmentier, J. F., & Broisin, J. (2021, September). Recommendations for Orchestration of Formative Assessment Sequences: a Data-driven Approach. Accepted for the interna-

⁷Association des Technologies de l'Information pour l'Education et la Formation - <http://www.atief.fr/>

tional A+ ranked EC-TEL conference of 2021 [8]. **This paper won the award for best paper of the conference.**

- Andriamiseza, R., Silvestre, F., Parmentier, J. F., & Broisin, J. (2021, June). Data-informed Decision-making in TEFA Processes: An Empirical Study of a Process Derived from Peer-Instruction [7]. Accepted for the international B ranked conference L@S 2021.
- Andriamiseza, R., Silvestre, F., Parmentier, J. F., & Broisin, J. (2021, June). Vers la conception de feedback pour enseignants dans un contexte d'évaluation formative à grande échelle: une approche analytique [9]. Accepted for the national A ranked EIAH conference of 2021.
- Andriamiseza, R. (2020, June). Évaluer la réutilisabilité d'une question : une utilisation des learning analytics dans un contexte d'évaluation formative [6]. Accepted for the national B ranked RJC conference of 2021.

At the time of writing the manuscript, one long paper has been submitted to the A+ international journal IEEE Transactions on Learning Technologies. A revised version has been submitted and is being peer reviewed.

1.3 Organization of the manuscript

The manuscript is organized as follows. Part I explores the literature in order to justify our objectives and research questions. Part II describes our research method. Part III describes the empirical study that we conducted to design solutions that are expected to answer our research questions. Part IV describes the study designed to put our solutions to the test.

Part I
State of the Art

Chapter 2

Formative Assessment and Feedback

Contents

2.1	Formative Assessment	15
2.2	Feedback	16

2.1 Formative Assessment

Assessing is part of the natural process of learning. It consists in asking a learner to complete one or more tasks in order to collect the results of her performance. Assessing a learner allows to certify her understanding and mastery of a concept (e.g. midterm exam, paper, final project, etc.). Such an assessment *of* learning is called summative assessment [81]. However, another kind of assessment aims at helping teachers and learners adapt their behavior regarding their method of teaching and learning. This assessment *for* learning is called formative assessment [159]. Various studies support formative assessment as a positively impactful teaching instructional process. As an example, in his project named "Visible Learning" [66], Hattie listed influences that have a positive and negative impact on learning outcomes. Among them, formative assessment and feedback were above average. Similarly, other studies stated that formative assessment is an effective teaching and learning method to include in the instructional process [23, 61, 89, 109]. Some qualitative studies about instructors' opinions and practices shed light on the ways formative assessment contributes to ensure immediate feedback, increase motivation and focus on learning materials. More precisely, instructors stated that formative assessment allows them to better detect students

deficiencies and therefore accurately adjust their teaching [80, 161]. Regarding learners' opinion, students acknowledged that feedback collected from formative assessment remains important for them as it helps to fill their learning gaps [45]. More generally, they agreed that formative assessment helps them to identify their weaknesses.

The theoretical definition of formative assessment is the following one [20, p. 2]:

"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."

In addition to this definition, earlier works by Black and Wiliam [21] proposed a unifying basis for diverse practices that are said to be formative. More precisely, they proposed key activities such as "classroom questioning" and "formative use of summative tests" as well as strategies such as "engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding" and "providing feedback that moves learners forward". This definition and theoretical framework emphasize how crucial the role of feedback is when it comes to formative assessment.

2.2 Feedback

Even though feedback has various meaning depending on the field of research, the one relevant to formative assessment is defined as follows: "Feedback is (...) defined in terms of information about how successfully something has been or is being done." [128, p. 120]. When it is provided in the context of a formative assessment activity, it acts as a support for both teachers and learners.

In 2007, Hattie [65] provided a framework that identifies questions whose answers determine how useful a feedback for learners is. These questions are (i) Where am I going? (ii) How am I going? (iii) Where to next? And, in addition to that, four levels (which embody four aspects of the performance of the student regarding the learning activity) of feedback are defined:

- Feedback about the task is a feedback that focuses on a task or a product. It can be a simple statement about whether the result is right or wrong [94], or about what is missing in the product. This feedback is powerful when the task information is then used to improve strategy processing or enhance self-regulation.

- Feedback about the processing of the task digs deeper. It is aimed at the processes underlying tasks or relating and extending tasks. The examples mentioned in the study [65] are feedback that relate to students' strategies for error detection. In this case, the feedback is not about the task itself but more about the way learners performed the task. These feedback are perceived as powerful in terms of deep processing and mastery of tasks.
- Feedback about self-regulation is about the way students monitor, drive, and regulate actions towards the learning goal. As an illustration, they can be about the way learners seek help. According to Hattie and Temperley, these feedback are also powerful.
- Feedback about the self as a person goes beyond the student as a performer by focusing on the student as a person. Such feedback can be a simple commentary such as "You are a good student". The study argues that this kind of feedback are the least effective ones.

This study [65] focuses on feedback for learners that are intended to help them to adapt their learning through self-regulation. They propose extensive theoretical definition and frameworks. However, as the definition of formative assessment stated, feedback obtained through learning activities in the context of formative assessment can also be used as a way to adapt teaching. When this is the case, the feedback is meant to inform teachers.

When studying the positive influences of various factors on learners achievement (such as teachers practices or context of learning), Hattie [66, p. 4] made the following statement:

"When I completed the first synthesis of 134 meta analyses of all possible influences on achievement [64], it soon became clear that feedback was among the most positive influences on achievement... The mistake I was making was seeing feedback as something teachers provided to students. I discovered that feedback is most powerful when it is from the student to the teacher. What they know, what they understand, where they make errors, when they have misconceptions, when they are not engaged – then teaching and learning can be synchronized and powerful. Feedback to teachers makes learning visible."

However, research has hardly focused on feedback for teachers compared to feedback for learners [146]. One of the reasons could be that capturing all learning interactions in a face-to-face context [53] and accounting for numerous variables such as each learner's skills, knowledge and understanding [114] is a hard task. Consequently, teachers struggle with for-

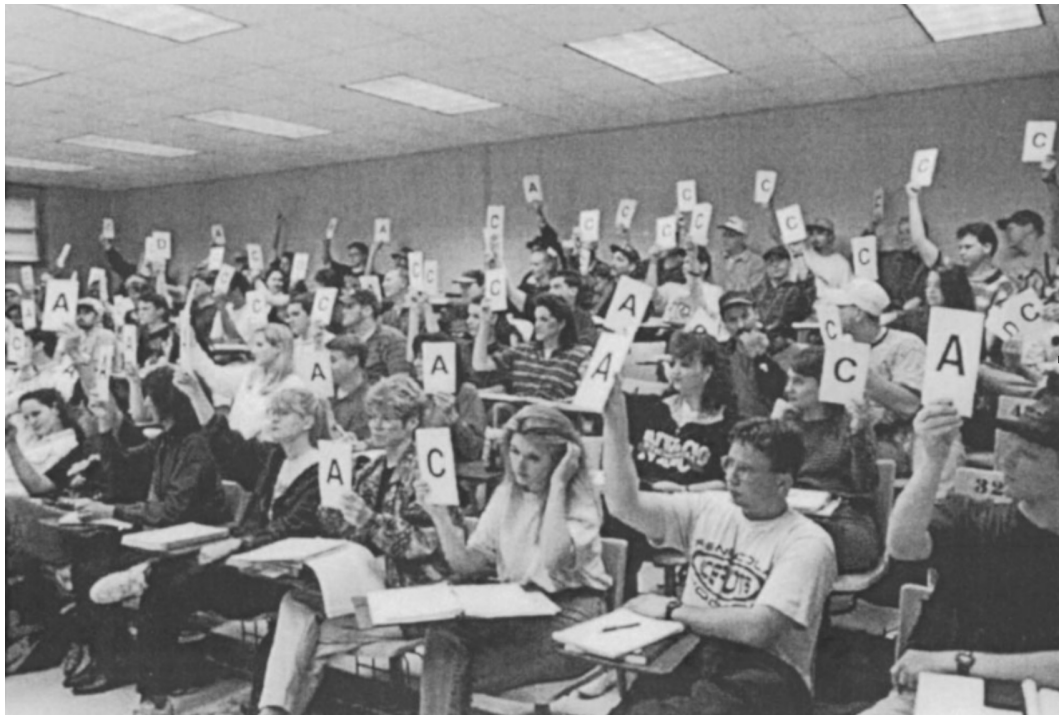


Figure 2.1: Usage of cardboards for formative assessment purposes [100].

formative assessment [134] and often use it in an informal and approximate way [5]. This leads to report on the mixed-effects of formative assessment in classroom practices [13, 58]. This statement's importance is considerable since the number of learners in educational settings is constantly increasing [1, 26, 101, 118, 119] and teachers often lack time [2, 42].

For instance, in face-to-face settings, Meltzer and Mannivan reported on the usage of pieces of papers or cardboards to allow students to answer questions asked by teachers [100]. Thanks to this feedback (see Figure 2.1), teachers can collect learners' answers and adapt their teaching. However, this formative assessment process gets harder to conduct when the number of learners increases because it implies more answers to simultaneously collect and process.

In summary, teachers would benefit from feedback designed to help them adapt in response to emergent class patterns [148]. Such an adaptation places the teacher at the center of the learning process as an orchestrator of a sequence of activities, and makes her responsible for making timely and context-relevant adjustments [82]. As a consequence, we want to focus our research on feedback for teachers designed to facilitate orchestration of formative assessment sequences in large scale settings. To do so, technology-based

environments are a relevant solution, particularly in activities that require the tracking of every student in the class [48, 103, 111, 125].

Chapter 3

Technology-Enhanced Formative Assessment Processes

Contents

3.1	Questioning the Class	21
3.2	Peer Confrontation Processes	24
3.3	The Two-votes-based Processes	27
3.3.1	Overall Description	27
3.3.2	Benefits of Two-Votes-Based Processes	30
3.3.3	Limits of Two-Votes-Based Processes	34

The usage of technology to deliver formative assessment is called Technology-Enhanced Formative Assessment (TEFA). It is one of the emerging solutions for delivering formative assessment with immediate feedback for both learners and teachers [142].

3.1 Questioning the Class

TEFA facilitates the implementation of formative assessment processes such as collecting learners' answers to a question. Class questioning is one of the formative assessment strategies mentioned by Black and Wiliam (see subsection 2.1). In addition to that, questioning a class is perceived as a powerful formative assessment approach [107, 157] because it is an efficient way to obtain immediate feedback for teachers and learners. Instead of using cardboards, Classroom Response Systems (CRS) [15] can be used as a more

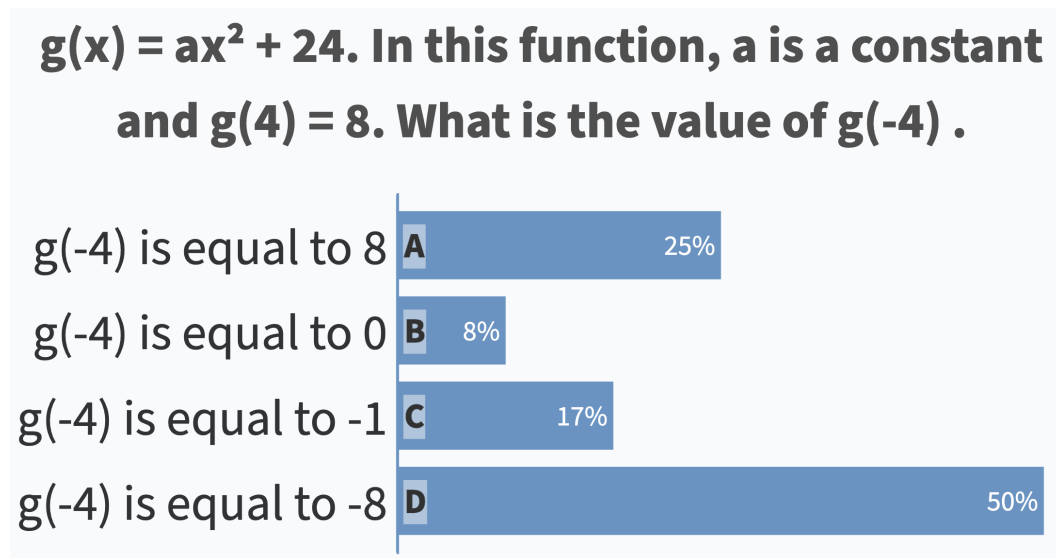


Figure 3.1: Poll Everywhere: Feedback displaying the distribution of votes

efficient alternative to question a class and obtain feedback regardless of the size of the classroom. Indeed, collecting learners' answers is a generic process used in TEFA that was first implemented by clickers [27]. Typically, a formative assessment sequence based on this process works as follows:

1. Teachers ask a choice or open-ended question (so that the answers can be automatically processed).
2. Learners submit an answer.
3. Based on learners' answer, a feedback is immediately given to teachers and learners in order to help them engage in a discussion with learners.

Some studies support such a computer-based assessment process and state that quality questioning makes both teaching and learning more effective [10, 19, 76].

The feedback given to teachers and learners during step 3 depends on the type of question. As an example, if the question is a choice question, a web-based questioning platform called Poll Everywhere [36] proposes to provide teachers with the distribution of votes as shown in Figure 3.1. When it comes to questions that require short written responses (such as open-ended questions), a web-based tool called Concept Warehouse [24] provides teachers with a word cloud as well as a representative explanation (see Figure 3.2). Several other platforms such as Kahoot [73], Socrative [12] and Plickers [83] support the same process. However, beyond the overview of learners' vote,

Word Cloud of Written Explanations	Representative Explanation (emphasis added)
<p>air temperature pressure volume decrease increase perfectly insulated heat remain system</p>	<p>“if the pressure decreases and the volume increases then there will be no change in the temperature”</p>

Figure 3.2: Overview of learners short written responses to a physics related question [24].

Response	Via	Screen name	Received at
g(-4) is equal to -8	pollev.com/rialyandriam928	Learner 11	March 18, 2022, 05:38 AM
g(-4) is equal to -8	pollev.com/rialyandriam928	Learner10	March 18, 2022, 05:35 AM
g(-4) is equal to -8	pollev.com/rialyandriam928	Learner9	March 18, 2022, 05:35 AM
g(-4) is equal to -8	pollev.com/rialyandriam928	Learner8	March 18, 2022, 05:34 AM
g(-4) is equal to -8	pollev.com/rialyandriam928	Learner7	March 18, 2022, 05:34 AM

Figure 3.3: Poll Everywhere: Feedback displaying the individual votes.

some of them propose a feedback providing teachers with the answer of each individual learner. Such feedback is implemented by Poll Everywhere as well. As shown in Figure 3.3, each answer is presented in a table. For each one of them, teachers can see the learner who voted and its timestamp. Finally, a former version of Tsaap-Note allowed student to participate in a notetaking activity and offered teachers the opportunity to test learners through a “notes as questions” feature [136]. More precisely, it integrates a questioning feature based on closed question in a notetaking tool, allowing learners to take notes during step 3. One of the main observation from the usage of Tsaap-Note was that students tend to take notes more often when they answered the question incorrectly.

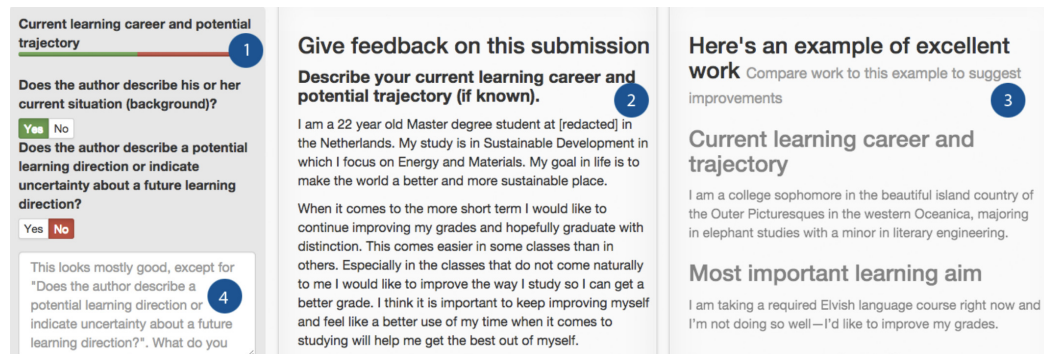


Figure 3.4: Peer Studio: Learner’s screen during the review process. The screenshot shows (1) the rubric, (2) the student draft, (3) an example of excellent work to compare student work against (4) automatically generated tips for commenting [85].

3.2 Peer Confrontation Processes

Activating learners as instructional resources is a strategy that also enters the frame of formative assessment as identified by Black & Wiliam (see Section 2). Therefore, many TEFA systems propose learners to review each other’s answer after they submit their own. The resulting process is the following:

1. Teachers ask a question.
2. Learners submit their work.
3. Learners review works by peers.
4. Both learners and teachers engage in discussion based on the results.

This process allows learners to undertake the critical process of determining and articulating what makes a “better” answer through writing individual feedback. It is implemented by PeerStudio [85], Juxtapeer [28] and ComPAIR [117]. Peer Studio is designed to allow students to create and receive feedback on any number of drafts for every open-ended assignment. When learners submit their draft, they are asked to review their peers’ submissions in order to unlock their own feedback. More precisely, they must give rubric-based feedback on two peers’ drafts as shown in Figure 3.4.

Juxtapeer goes further by asking learners to grade more than two of their peers’ work. On each review, learners complete an evaluation which consists in providing grades based on specific criteria as shown in Figure 3.5, comparative preferences, and open-ended comments. The review process can either

Figure 3.5: Juxtapeer: An example of a review form [28].

be by comparison or serial, as shown on Figure 3.6. Both cases ask learner to review only one submission at a time (i.e., the red one in the "compare" case). The algorithm that determines which submission will be used for comparison is a modified version of the Crowd-BT active learning algorithm [30]. This algorithm is based on the Bradley-Terry model [22] which suggests that learners are more likely to make a comparison 'correctly' (i.e. agree with the consensus) if the difference (regarding the criterion of comparison) between items is large. Conversely, it can compute the reliability of a rater based on his ability to make a correct comparison depending on how different the two submissions are. Over repeated comparisons, Crowd-BT improves estimates of quality and reliability. By weighting more reliable raters higher, and asking raters to compare pairs with smaller perceived differences, Crowd-BT improves ranking accuracy with a given number of comparisons. The final item on each rubric asks learners which submission they prefer overall; Crowd-BT uses this pairwise preference to update its quality estimate of both submissions. Once learners ended their review phase, they can receive feedback regarding their work. Rather than simply providing learners with a list of all the reviews their work received, Juxtapeer displays each feedback in the context they were provided. In other words, each feedback is presented alongside the other work that served as a comparison. This allows learners to see more of their peers work, and contextualize the reviews they received. Finally, with ComPAIR, teachers can view the feedback provided by each learner (see Figure 3.7).

Effectiveness of the peer review process is supported by previous studies.



Figure 3.6: Juxtapeer: Visual representation of the review process [28].

3 comparisons found for "Student 12345"

Student 12345 ↓

Student 12345 submitted on Jan 21 @ 4:12 pm:

Comparison for "Which answer has the better critical idea?"

The first critical statement focuses on describing and observing Prospero's ambitions. The second statement broadens Prospero's ambition to humans in general and makes a point to the lengths that we go to fulfill our desires. The first could have added a point to argue for. The second offers more analysis.

[Show compared answers](#)

Comparison for "Which answer is more effectively articulated? Explain the reason for your preference."

The first critical statement seemed a bit awkward to me. "William Shakespeare implies the representation of human nature of ambition" is perhaps too wordy. I would change the sentence to something more simple such as "represents human ambition". The second critical statement is read smoothly and clear.

[Show compared answers](#)

Feedback for Odd-Numbered Answer

I agree that Shakespeare uses Prospero's magical power and manipulation to represent human ambition. Prospero's goals are mentioned in detail in this critical premise when words could have been used to make another point such as the lengths we go to meet our goals.

Feedback for Even-Numbered Answer

I think this critical premise is articulated well. Mentioning how Shakespeare paints a larger picture of human ambition and the lengths we go to fulfill our desires is a nice generalization to take from Prospero's ambition.

Figure 3.7: ComPAIR: Feedback displaying the individual performances of learners [117].

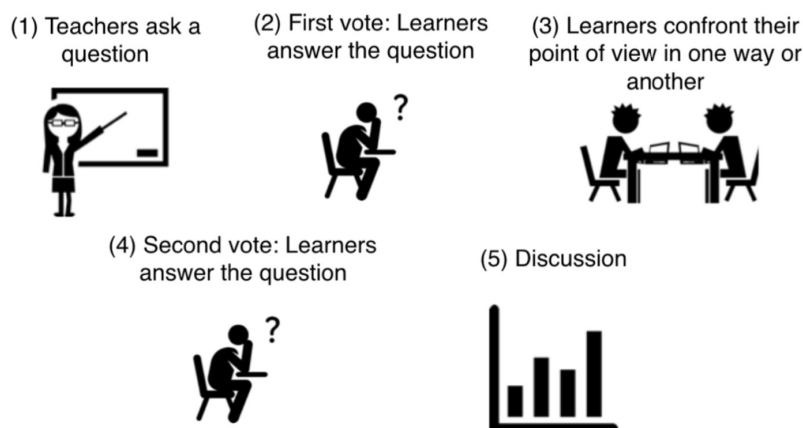


Figure 3.8: The 5 steps of a two-votes-based process.

First, student performance over a course of an academic program can be significantly affected and positively influenced thanks to feedback provided by peers [106]. This is because using peer feedback has numerous benefits: it is a scalable practice [149], it has long-term benefits [74] and it impacts motivation and self-regulatory skills such as giving feedback [158].

3.3 The Two-votes-based Processes

3.3.1 Overall Description

Questioning the class again after a peer review activity allows teachers to measure the effect of such activity on learners' understanding of the topic targeted by the question. We named the resulting set of processes the *Two-votes-based processes* [8].

The two-votes-based process is a generalisation of a group of formative assessment processes that can be illustrated in Figure 3.8. It consists in 5 steps. In this figure, we can see that a two-votes-based sequence asks learners to vote for the correct answer during the second and fourth steps, which respectively take place before and after a confrontation of viewpoints.

One of the earliest implementation of this process is Mazur's Peer Instruction [40] which was mainly used with clickers and worked as follows:

1. Teachers ask a question (multiple choice question, exclusive choice question or open-ended question).
2. Learners submit an answer.

3. Learners discuss with their neighbour about their answer.
4. Learners submit an answer again, they can either change their answer or keep their initial one.
5. Both learners and teachers engage in a discussion, based on the results.

When it comes to Peer Instruction, Mazur used exclusively conceptual questions called "ConceptTests" such as the Force Concept Inventory which is a Physics related Concept Test [70]. ConceptTests are choice questions that have been studied and validated by numerous studies.

Another implementation of the two-votes-based process was obtained through an evolution of Tsaap-Note. More precisely, based on Tsaap-Note's former process described in Section 3.1, two limits were identified regarding learners' note [138]. First, every single notes taken by learners on a given question are automatically included into the feedback of the quiz; the notes are not evaluated nor filtered. Secondly, only few learners participate in the writing of explanations that are required to build relevant feedback. Consequently, a new feature to manage interactive questions was implemented. This new feature leads to a new process that consists in questioning the class again after the peer review activity. Consequently, the implementation of the two-votes-based process that the new version of Tsaap-Notes proposes consists in allowing learners to confront their viewpoint through a peer rating activity:

1. Teachers ask a choice question.
2. Learners select an answer, provide their confidence degree and a rationale.
3. Learners are presented with an alternative response to theirs.
4. Learners are given a chance to change their answer.
5. Each learner is asked to grade three of their peers' rationales using a Likert scale [77] graduated from 1 to 5. The graded rationales are exclusively the ones that are related to the correct answer. Then, both learners and teachers engage in a discussion based on the results.

In order to (i) save time (ii) allow learners who provided the incorrect answer to receive graded feedback and (iii) allow teachers to identify popular misconceptions, Tsaap-Notes' process evolved again. It was renamed Elaastic and proposed a variation of its original process:

1. Teachers ask a question (multiple choice question, exclusive choice question or open-ended question).
2. Learners submit an answer by providing a written rationale and their confidence degree. They must also provide their selected choice(s) in case of choice question.
3. Each learner grades a set of her peer's rationales on a Likert Scale ranging from 1 to 5.
4. Each learner submits an answer again, she can change her selected choice and/or her confidence degree or keep her initial choice and confidence degree.
5. Both learners and teachers engage in discussion based on the results.

At anytime of the sequence, Elaastic can display first and second votes of learners and provide teachers with each learner's written explanation and the mean rate attributed by peers. More details on Elaastic are given in Chapter 5.

Another technological implementation of the two-votes-based process that can be found in literature is myDalite [29]. The five steps that it proposes are the following ones:

1. Teachers ask a question (multiple choice question, exclusive choice question or open-ended question).
2. Learners submit an answer.
3. Each learner is presented with a set of her peers rationales.
4. Each learner can select one of these rationales as her second submitted answer. She can either select her own rationale and therefore keep their initial answer or select a peer rationale and therefore change her initial answer.
5. Both learners and teachers engage in a discussion, based on the results.

As shown in these five steps, myDalite groups the viewpoint confrontation and second submission of an answer in a single activity (see Figure 3.9). 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 (see Figure 3.10).

You answered **A** and gave this rationale:

The bottom rays will not pass through the lens, but the top rays will. Since the final real image is inverted, then only the bottom part of the image will be present (representing the top part of the object).

Consider the problem again, noting the rationales below that have been provided by other students. They may, or may not, cause you to reconsider your answer. Read them and select your final answer.

A. Considering that the image formed by a converging lens is inverted, we know that rays that pass through the lower half converge and end up in the top half of the image, and that rays that pass through the top half of the lens converge and end up in the lower half of the image. Therefore, if the lower half of the lens is covered, rays that usually go through it and converge to the top half of the image will be missing; the top half of the real image will be missing.

Since the image is inverted, the bottom part that is covered would have been placed at the top. And since it is covered, that part will be missing in the final image.

I stick with my own rationale.

D. Since half the lens is covered, the image produced will have less light being projected, resulting in a dimmer image. The full image still forms because the light rays will still converge in the top half of the lens.

Clearly not all rays will hit the screen, but enough rays emerging from all of the object WILL hit the screen. The final real image will be complete, but will be less bright (hence dimmer) because not all of the light intensity goes through the lens.

Figure 3.9: MyDalite: Viewpoints confrontation and second vote [29].

3.3.2 Benefits of Two-Votes-Based Processes

With two-votes-based processes, the number of students who provided the correct answer at the second vote is expected to be higher than at the first one. When this is the case, we qualify such sequence as *beneficial* because it means that students' general understanding of the topic has been enhanced. This statement is supported by Smith's study [140] who conducted Peer Instruction sequences and followed each one of them with a second question that is intended to target the same concept as the previously asked one. The results showed that the improvement observed in the second vote of the first question was similar to the result from the second question, which suggests that students can actually benefit from a two-votes-based sequence.

Previous quantitative studies reported on the usage of two-votes-based

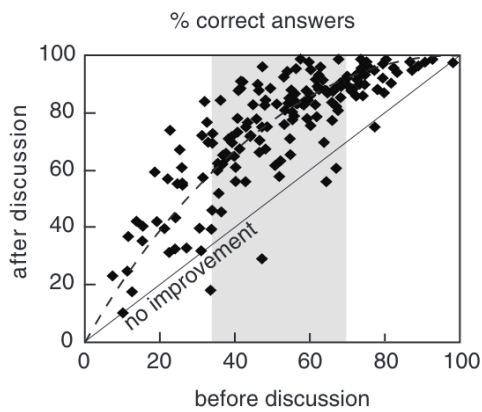
Question	N	RR	RW	WR	WW
Reading - RC Circuits	19	19	0	0	0
Rc circuit - Shape of Discharging graphs in Real Lab	20	10	0	3	7
RC circuit - Shape of discharging graphs	19	10	1	4	4
RC circuit - Shape of charging graphs	19	9	0	5	5
RC circuit - increase charging time	19	16	0	2	1
RC circuit - discharging graph time constant	19	12	0	0	7

Figure 3.10: MyDalite: Learners' response switching [29].

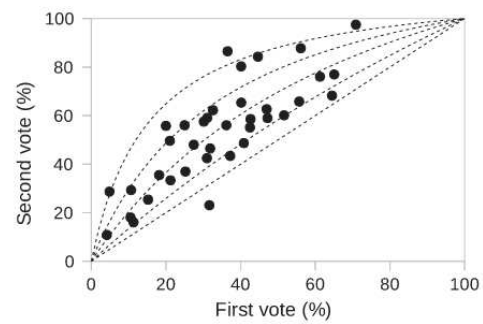
processes and obtained a fairly high percentage of beneficial two-votes-based sequences (see Figure 3.11). As illustrations, Mazur reported on the usage of Peer Instruction where he obtained more than 140 beneficial sequences and very few non beneficial sequences, as shown in Figure 3.11a. With 37 two-votes-based sequences (see Figure 3.11b), Parmentier obtained 36 beneficial sequences. Among 86 sequences, Tullis ended up with 55 beneficial sequences (see Figure 3.11c). Finally, in a previous study that we conducted and that is detailed in the Chapter 6, we obtained 65 beneficial sequences out of 104 (see Figure 3.11d). Contrary to the 3 previously mentioned studies, this one did not include exclusively ConceptTests. Instead, teachers wrote their own questions.

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 [29]. According to Crouch and Mazur [40], this process cognitively engages students at different levels.

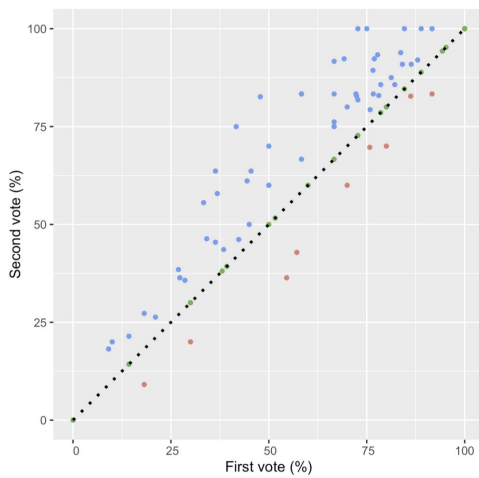
Regarding engagement of learners, the ICAP framework [33] is a relevant taxonomy to classify the type of engagement of learners during a formative assessment sequence supported by the Two-votes-based process. This framework categorises a learner's engagement into four modes ordered from the most to the least desired one. These four levels are Interactive, Constructive, Active and Passive. Passive is the lowest engagement level, it is defined as an activity where learners receive information without doing anything else related to learning (e.g. listening to a lecture). The next level of engagement is active, it goes beyond the passive level by having learners perform motoric activities when receiving information (e.g. highlighting a text or replaying a specific segment of a video). If those motoric activities generate a product or some knowledge beyond the information received by the learners, the level of engagement is constructive (e.g. writing an answer to a question). Finally, the interactive engagement consists in conducting dialogues where (a) every



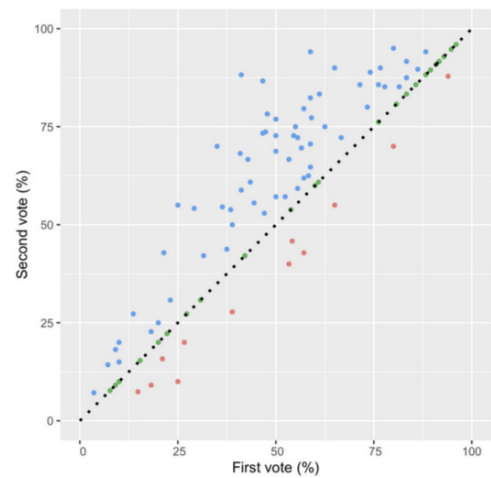
(a) Results of Mazur's study [97] with ConceptTests.



(b) Results of Parmentier's study [115] with ConceptTests.



(c) Results of Tullis' study [150] with ConceptTests.



(d) Results of Andriamiseza's study [8] with free questions.

Figure 3.11: Percentage of correct answers before and after the confrontation step for various studies.

partners' involved must primarily utter constructive statements, and (b) a sufficient degree of turn taking must occur. Based on this framework, we can classify the level of engagement of the 5 steps of the two-votes-based process.

1. **Teachers ask a question:** During this step, learners must read the question. The level of engagement involved here is the *passive* one. However, one might argue that such a step is not a learner activity by itself and should therefore not measure learners' engagement. It should be paired with step 2 so as to determine the level of engagement of the resulting activity.
2. **Learners submit an answer:** Since the answer to a question is an additional output produced beyond the question, we believe such an activity to have a *constructive* level of engagement (regardless whether this step is to be paired with step 1 or not).
3. **Learners confront their viewpoints:** The confrontation of viewpoint is based on each learner's rationale. Therefore, we can argue that each peer contributes to the discussion. As stated earlier, this confrontation can take several forms depending on the implementation. As an example, MyDalite and Elaastic only present learners with peer's rationale which is arguably not a discussion. More precisely, learners do not take turns talking, as stated in the definition of the interactive level of engagement. However, such a requirement is expected to make it easier for students to incorporate their partners' understanding of the domain and to make adjustments to their own mental model [33]. Such an expectation is met even though it is not done through a discussion as formally defined by the ICAP framework. In other words, the confrontation of viewpoint that occurs in the two-votes-based process does not meet the definition of the interactive level of engagement but meets its expectations. Consequently, we can classify such a step as a step with an engagement level between the constructive and interactive ones.
4. **Learners submit an answer again:** Similarly to step 2, this step has a constructive level of engagement.
5. **Both learners and teachers engage in a discussion based on the results:** Depending on how teachers lead the discussion, this activity can range from passive (simple oral correction from teachers) to interactive (teachers and students engage in a dialogue to better understand the answer to the question).

Silvestre's previous works show that the two-votes-based process fulfills most of the activities and strategies of formative assessment [137]

3.3.3 Limits of Two-Votes-Based Processes

Even though the usage of two-votes-based processes mostly led to beneficial sequences for learners, there are still some sequences that failed to be beneficial (see Figure 3.11). In other words, there are situations where engaging learners in the confrontation and second vote steps does not lead to beneficial outcomes. Therefore, teachers should be provided with feedback during the sequence so as to help them decide which step they should engage learners in next. However, few studies proposed feedback designed to help teachers make decision during a two-votes-based sequence. Most instructional feedback occurs via student evaluations, which, by definition, focuses on learners' behavior and therefore lack specific feedback for improvement of teachers' behavior [63]. More generally, due to the lack of data related to two-votes-based processes [18], little work has explored how to use these interactions to provide advanced feedback for teachers orchestration by inferring new knowledge about formative assessment. Consequently, we want to focus on feedback for teaching and, more specifically, orchestration and decision-making by teachers during formative assessment sequences.

In 2015, Vickrey [155] proposed a decision making model designed to support teachers' orchestration of Peer instruction, which is an implementation of the two-votes-based process (see Figure 3.12). However, to the best of our knowledge, this model has never been implemented in any TEFA system, and therefore has never been evaluated at large scale. Furthermore, it is a non-deterministic decision model. More precisely, when there are too few correct answers, it recommends teachers to provide hints to learners, or provide them with detailed explanations before restarting a sequence. There are no further indications regarding the most suited decision of the two.

In summary, even though some studies support two-votes-based process as beneficial formative assessment practices, there is still room for improvement, especially regarding feedback for teachers orchestration.

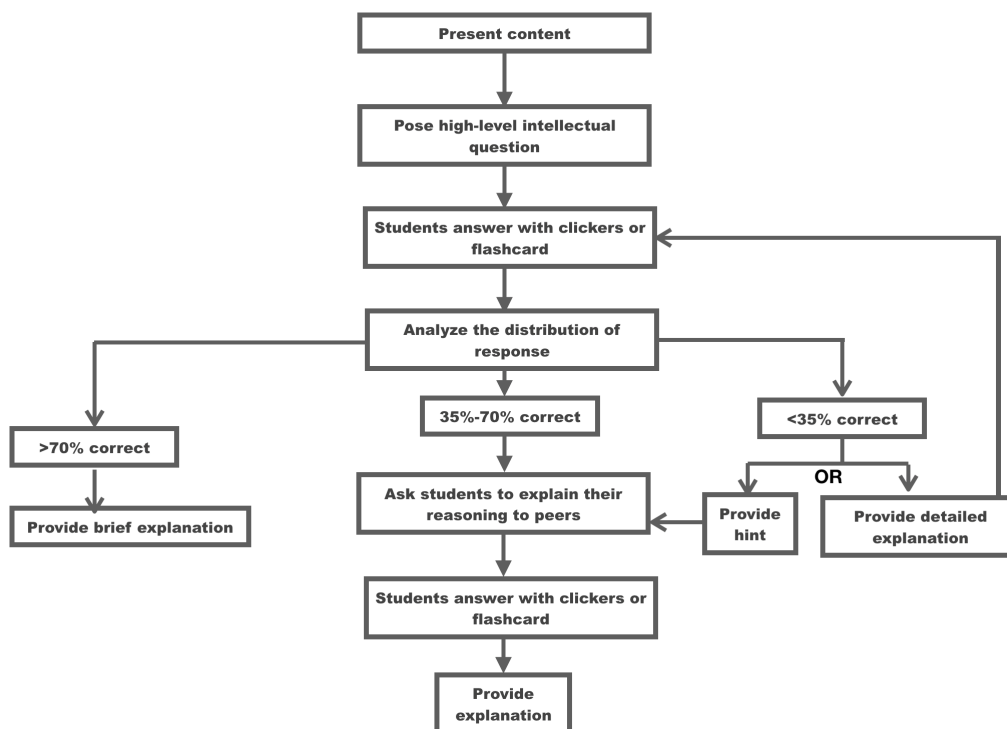


Figure 3.12: Vickrey's model of Peer Instruction.

Part II
Research method

Chapter 4

Research method

Contents

4.1	From Limits to Research Questions	39
4.2	Design-Based Research	40
4.3	The Cycle of Learning Analytics	41

4.1 From Limits to Research Questions

In the previous chapter, we focused on formative assessment and emphasized the challenge of its usages. More particularly, we emphasized the difficulty for teachers to implement it at scale due to the lack of feedback provided to them. We introduced TEFA as a way to perform effective formative assessment at scale and showed that there is a lack of advanced feedback for teachers orchestration. As a solution, we want to explore data that were generated from the process implemented by a formative assessment tool. Hence, in this thesis, we identify one main objective to improve formative assessment practices: to improve TEFA systems by identifying and integrating relevant feedback for teachers. To do so, we focus on two-votes-based processes because they are beneficial to learners and offer a wide variety of interactions. The current limits of formative assessment processes and the potential of the two-votes-based processes lead us to the following questions:

- **RQ1 - Which useful information can be inferred from the analysis of data gathered from a tool used in authentic contexts?**
- **RQ2 - How can such information contribute to improve formative assessment processes orchestration?**

To answer these questions, we have access to the data collected from the authentic usage of Elaastic, the web-based platform introduced in the previous chapter.

4.2 Design-Based Research

Research conducted in such context falls into the category of Design-Based Research (DBR) [123]. It is a research methodology that is highly associated with education and TEL due to its compatibility with real and authentic contexts [17, 91]. More precisely, one of the advantages of DBR is its impact on practice [120], as well as its collaborative aspect between researchers and practitioners [4]. A basic process of Design-Based Research consists in designing solutions to problems and to test how well such solutions work. This can be done with an iterative approach where each iteration is designed to (i) adapt the solution (ii) test it again (iii) gather more data. Each iteration consists of the following phases: analysis, design, development and implementation [156, p. 6]. It starts with an issue and a product, and ends with new knowledges and hypotheses that will be verified in the next iteration. Consequently, data analysis often take the form of iterative comparisons [75].

We propose to conduct Design-Based Research which consists in following these principles:

- Conduct our works in an authentic context.
- Adopt an iterative method.
- Involve teachers in the loop.
- Conclude with practices that have a short-term impact on teaching.

In concrete terms, the starting point of our first iteration is the research questions mentioned in our previous section, whereas our product are Elaastic and the data that stem from its usage. Based on this data, we infer a new orchestration model that is intended to assist teachers in their decision-making. Finally, based on this model, we implement a new version of Elaastic and evaluate its impact on formative assessment through another iteration.

Since our method starts with an analysis of learning data, we propose to focus on the cycle of *Learning Analytics* which is at the core of the next subsection.

4.3 The Cycle of Learning Analytics

The cycle of Learning Analytics [41] shown in Figure 4.1 is a cycle that is intended to provide a framework to conduct successful Learning Analytics work. Its cyclic nature also makes it compatible with the DBR methods.

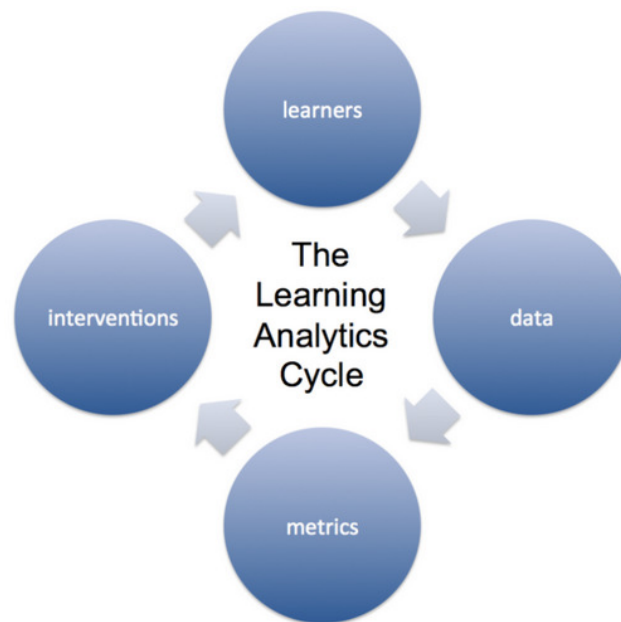


Figure 4.1: The *Learning Analytics* cycle [41]

The four components of this cycle are the followings:

- **Learners** is self-explanatory. It encompasses all kinds of learners in any settings. Informal learners participating in a MOOC are as valid as student of a course. The technology involved in the data collection also enters in the frame of this step because it plays a crucial role in defining the context of the data collected from learners.
- **Data** can be data about learners (such as demographic information) or generated by learners (such as traces).
- **Metrics** are obtained from the analysis of the data. These are some information that provide some insight about the learning process. Presenting these information to the various stakeholders of the learning process is at the core of a subdomain of Learning Analytics called *learning dashboards* [102, 154].

- **Interventions** are based on metrics. As an example, it can be the implementation of a new dashboard within a TEFA platform that enables learners to improve their self-regulation. In our case, the example that best suits our needs is any intervention that will help teachers improve her teaching, so as benefits to learners are increased.

The research conducted that we describe in this manuscript follow this cycle. Figure 4.2 describes the method we used to conduct our research

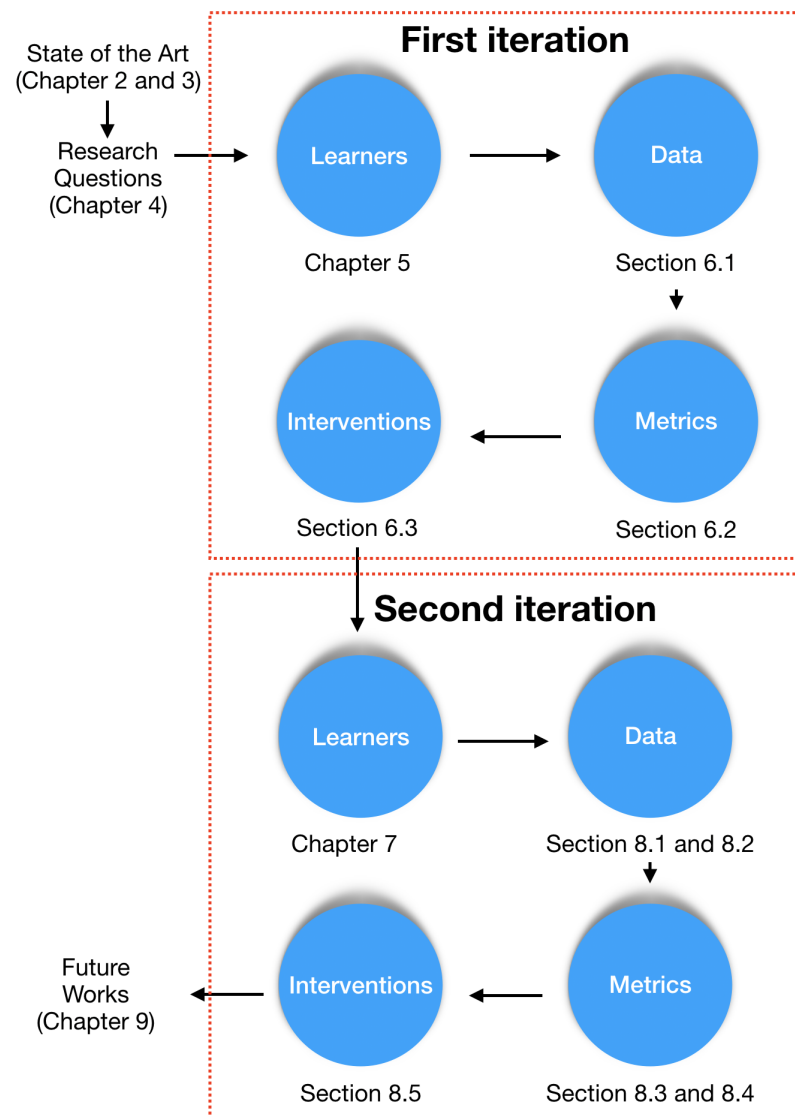


Figure 4.2: Organisation of our Works Based on the *Learning Analytics Cycle* based on this cycle as well as the matching chapters and sections of each

step. The first iteration will be based on data collected from Elaastic's usage from 2015 to 2019. With the intervention that is designed at the end of it, we will go through a second iteration to put such intervention to the test. More precisely, we implement our intervention within Elaastic and collect data by recruiting teachers. The data collected from this usage will be the ones that we will use at the start of our second iteration. Finally, we will base our future works on the results and intervention proposed at the end of this second iteration.

Part III

First Iteration: Empirical study

Chapter 5

Elaastic, Our Formative Assessment Tool

Contents

5.1	Overview of Elaastic	48
5.2	Learner perspective	48
5.2.1	Phase 1: Well-argued Response	48
5.2.2	Phase 2: Comparing Viewpoints	49
5.2.3	Phase 3: Results	49
5.3	Teacher perspective	50
5.3.1	Orchestrating the Sequence	50
5.3.2	Retrieving the Question and Answers of the Sequence	51
5.4	Internal mechanics	53

This chapter introduces our tool by providing an overview of Elaastic and explaining how it works from the learner perspective, before exposing the main features available to teachers. Afterwards, we provide further details regarding Elaastic’s pairing algorithm for the confrontation phase. If the reader wants to know more about Elaastic, its documentation is available online¹.

¹<https://elaastic.github.io/elaastic-questions-server/en/overview/>

5.1 Overview of Elaastic

Elaastic is an open source [37] web platform [122] implementing an instance of two-votes-based process, and used since 2015 in different higher education curricula across various disciplines such as computer science, physics or project management. Thanks to the B4MATIVE! project (see Section 1.1.3), it is now also used in secondary education in different subjects such as mathematics, history, geography, biology and music.

Elaastic is a class questioning tool that allows teachers to create exclusive and multiple choice, as well as open-ended questions. Its process consists of 3 *phases* that can be matched with one or more steps of the two-votes-based process. As shown in Figure 5.1, steps 1 and 2 occur in Elaastic's

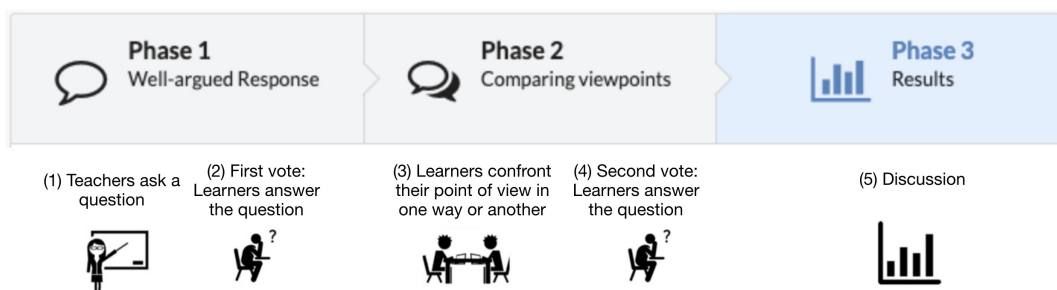


Figure 5.1: Elaastic's implementation of a two-votes-based process

phase 1, steps 3 and 4 occur in phase 2, and step 5 occurs in phase 3. When teachers want to ask a question to learners, they start a *sequence* which is to be interpreted as an instance of the process. A sequence can be run in synchronous or asynchronous context. However, our works focus on synchronous sequences. More details about the way asynchronous sequences are delivered can be found in Appendix A.

5.2 Learner perspective

5.2.1 Phase 1: Well-argued Response

During phase 1, learners must 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 5.2). This scale has 4 items because a neutral value would be difficult to interpret [108] regarding the confidence degree.

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 $x = 4$ as well.
So if $g(4) = 8$ then $g(-4) = 8$

Your evaluation:

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

Strongly disagree

Strongly agree

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.

Your answer: 1 2 3 4

Submit

Figure 5.3: Phase 2 of Elaastic: Comparing Viewpoints.

frame "My results" in Figure 5.4). If teachers want to, an overview of learners scores, rationales and mean rates can be displayed on learners' perspective for a debriefing (as shown in the frame "Results" in Figure 5.4).

5.3 Teacher perspective

Teachers orchestrate a sequence by starting, stopping or restarting a phase. The teacher perspective of a sequence is shown in Figure 5.5 which contains three frames described below.

5.3.1 Orchestrating the Sequence

The control frame is designed to help teachers orchestrate the sequence. The three phases are shown in a timeline that changes depending on the current state of the sequence. The numbers on top of the frame inform the teachers about the number of learners who completed each activity along the phases. From left to right, they represent the number of learners who provided a first answer, the number of learners who completed their evaluation(s), and

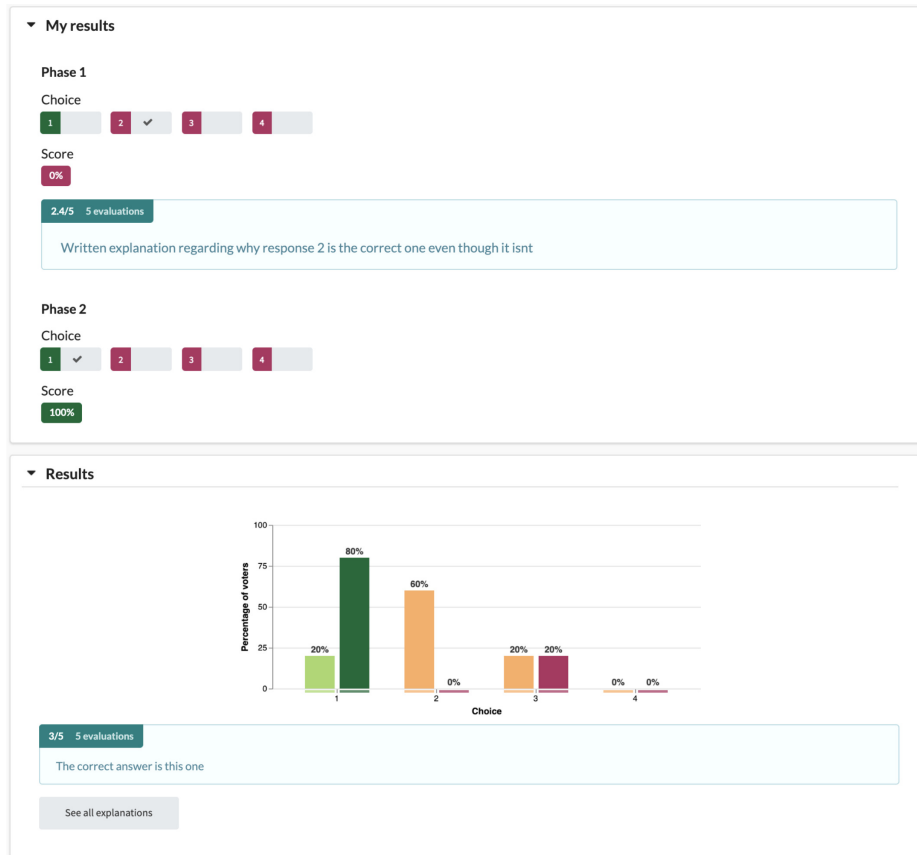


Figure 5.4: Phase 3 of Elaastic: Results.

the number of learners who provided a second answer. Below this timeline is a section to drive the sequence. It allows teachers to choose which phase learners should engage in next. The buttons that are available depend on the current state of the sequence. In this figure the sequence has ended and teachers published the results to learners. Consequently, they can end the sequence or cancel the publication of the results. More generally, when teachers have ended a phase, they can reopen it or start the next phase. They can also end the sequence whenever they want.

5.3.2 Retrieving the Question and Answers of the Sequence

At any moment of the sequence, the question summary frame displays the content of the question that is currently being played.

The results frame is a dashboard designed to summarise the results of

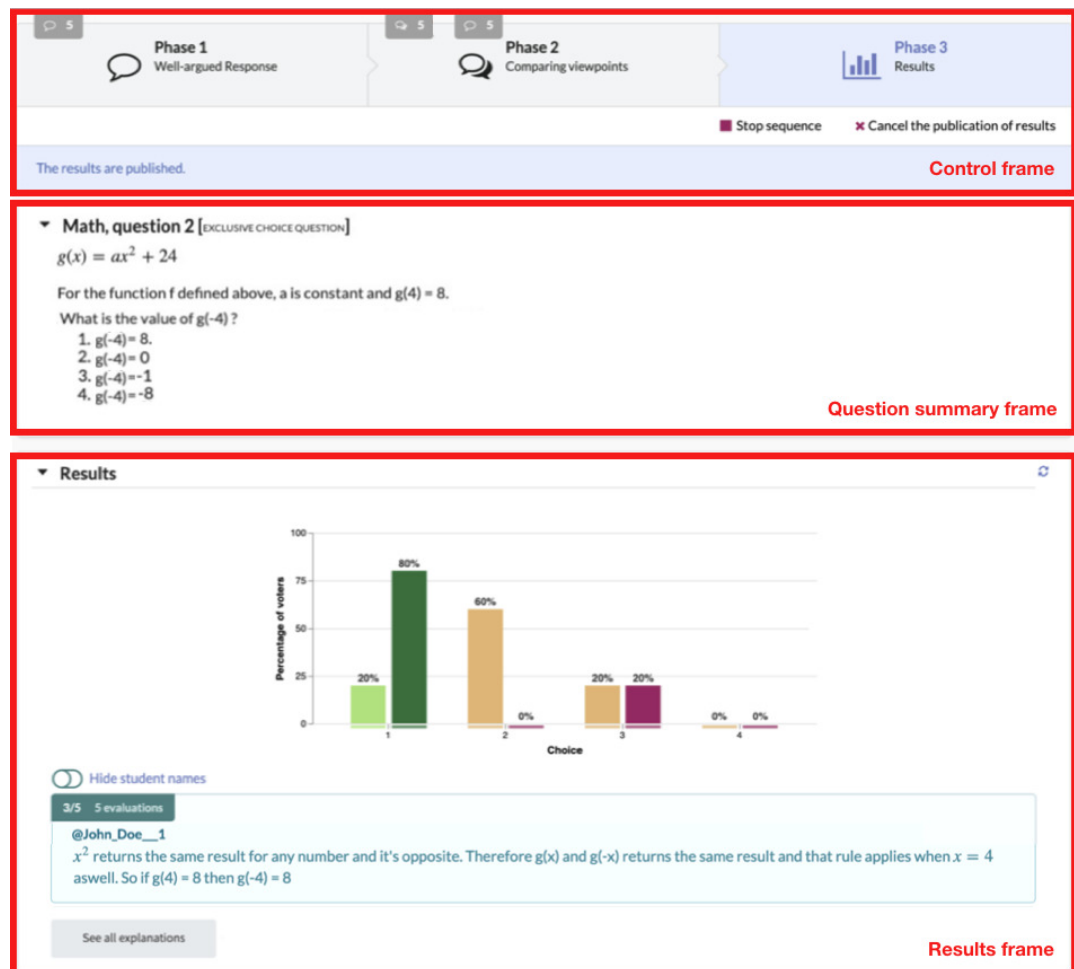


Figure 5.5: Elaastic: Teacher's perspective to manage a sequence

learners' answers along the phases. Each possible answer is shown on the x-axis. The correct ones are represented by green bars whereas the incorrect ones are represented by orange and red bars. Each possible answer has two columns. The left one represents the proportion of learners who chose the matching answer at the first vote, whereas the right one represents the proportion of learners who chose the answer at the second vote. Thanks to this diagram, teachers can quickly see the difference between the distribution of the first and second vote. Below this diagram is a selection of a few written explanations. The correct ones only are displayed from the highest graded one to the lowest. Teachers can see all explanations by clicking on the button labeled "See all explanations". Each rationale's content is displayed, as well as their mean grade given by peers.

5.4 Internal mechanics

As mentioned before, when teachers start a sequence, they can configure the number of rationales (up to 5) evaluated by each learner during the peer rating phase. Regarding the process of selecting and pushing the rationales to learners, it relies on an algorithm designed to fulfill, when possible, five requirements that are given below from the highest to the lowest priority:

1. A rationale with less than 10 characters is considered as irrelevant and is therefore discarded from the list of rationales to be evaluated by peers.
2. A learner can not evaluate the same rationale more than once.
3. For each set of rationales given to a learner, the number of rationales related to an incorrect answer and the number of rationales related to a correct one must be as close as possible. The aim is to avoid representation bias [116].
4. To promote sociocognitive conflicts [16], each learner must alternatively evaluate rationales related to correct answers and incorrect ones. The first one she sees must be related to an answer whose correctness is different than her own.
5. Rationales should receive the maximum number of evaluations.

In the example shown by Figure 5.6, each of the 6 learners provided a rationale. Two of them provided a correct answer (learners 1 and 5), whereas 4 of them provided an incorrect one (learners 2, 3, 4 and 6). The algorithm paired each learner with their peers' rationales according to the criteria mentioned above. Teachers configured the sequence so that each learner evaluates two rationales.

In this chapter, we presented the tool at the basis of our works. The high level of interactivity of this tool generates a high number of data that are captured. Furthermore, the two-votes-based process that it proposes makes it relevant to our research. In the following chapter, we present the data mining techniques we applied on the Elaastic's dataset to provide insights about information that impact the outcome of a formative sequence.

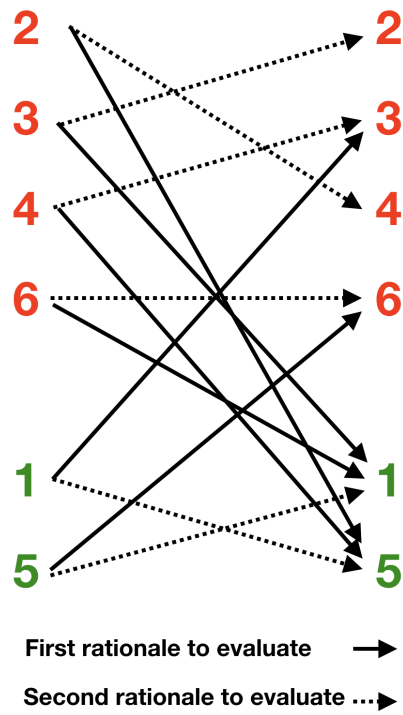


Figure 5.6: Pairing Algorithm: Matching example with 2 evaluations per learner.

Chapter 6

Data Analysis

Contents

6.1	Data Collection	56
6.1.1	Description of the dataset	56
6.1.2	Description of the sample	57
6.2	Findings	58
6.2.1	Basis for our Analysis: Measuring the Benefits of a Sequence	58
6.2.2	Proportion of Correct Answers during the First Vote	61
6.2.3	Learners Confidence Degree during the First Vote	63
6.2.4	Peer Grading Phase and Sequence Benefits	67
6.2.5	Peer Grading Phase and Learner’s Confidence De- gree during the First Vote	68
6.2.6	Self-Grading as a Substitute for Peer Grading	70
6.2.7	Peer Grading per Learner as Configured by Teachers	72
6.3	Discussion	73
6.3.1	Regarding the Research Questions	73
6.3.2	Implications for Research and Practice: Improving Vickrey’s model	74
6.3.3	Limitations	75
6.4	Conclusion of the Empirical Study	76

This section details the contributions of the empirical study of our first Learning Analytics cycle iteration as described in our research method (see Section 4). Section 6.1 presents the data that we collected from the use of

Elaastic in authentic settings. Section 6.2 details our analysis and presents two contributions. First, it identifies meaningful information impacting effectiveness of formative assessment sequences on the basis of significant correlation. Second, it proposes recommendations to assist designers of formative assessment systems (RQ1), as well as recommendations to assist teachers when orchestrating two-votes-based sequences (RQ2). Section 6.3 is a discussion about the results and introduces a new orchestration model as a contribution (RQ2). Section 6.4 concludes the study.

6.1 Data Collection

6.1.1 Description of the dataset

We gathered data from the use of Elaastic in higher education from 2015 to 2019. In this timeframe, we collected 623 sequences conducted by 53 teachers where 1,769 learners provided 8,757 answers and performed 9,256 peer ratings.

Figure 6.1 is a simplified UML class diagram that provides a summary of the structure of the dataset. A sequence is characterised by an execution

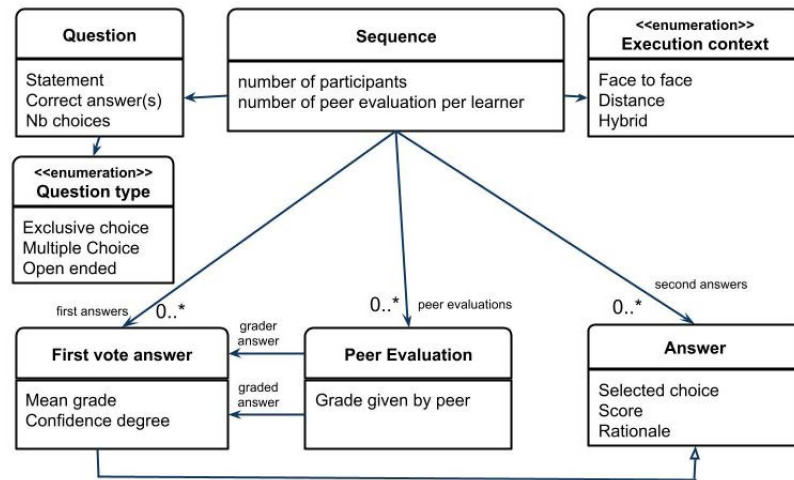


Figure 6.1: Summary of the dataset.

context, the number of peer evaluation each learner is asked to perform (as configured by teachers) as well as the number of participants. A sequence also contains a question, a set of answers (for the first and second answers) and a set of evaluations. For each answer, the following data are collected: the content of the rationale, the score and the selected choice(s) when applicable.

If the answer is a first answer, 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 and, in case of choice questions, by the number of different choices proposed to learners as well as a list of the correct choices. Finally, for each evaluation resulting from the peer grading activity, the following data are collected: the graded answer, the answer of the grader, and the grade given by the peer.

6.1.2 Description of the sample

The whole dataset has been filtered in order to meet the requirements of our upcoming data analysis that would answer our research questions. More precisely, we removed aborted sequences, survey sequences and test sequences. Then, 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). Afterwards, we kept face-to-face sequences only, since the asynchronous nature of distant and hybrid execution contexts in Elaastic does not require full orchestration from teachers (see Section 5.1). The next step consisted in removing sequences where there were less than 10 participants because we wanted to focus on large scale settings. Finally, based on the remaining sequences, we removed those where there were no correct answers, since the confrontation can not operate under these conditions (there are no rationales for correct answers to convince peers who provided incorrect answers). Sequences where all answers are correct at the first and/or second vote were removed as well, as they point out questions that were too easy to measure an effect size.

Figure 6.2 displays the amount of remaining sequences after each filter we applied to the dataset. We obtained 104 sequences conducted by 21 teachers where 616 learners provided 1,981 answers and performed 4,072 peer gradings. The questions asked in these sequences address mainly STEM¹ topics from higher education courses. Table 6.1 summarises the topics of the sequences included in the sample (i.e, the filtered dataset).

Starting from this sample that we name s_{2019} , we verify hypotheses based on literature and infer recommendations for system designers and/or teachers.

¹Science, Technology, Engineering and Mathematics

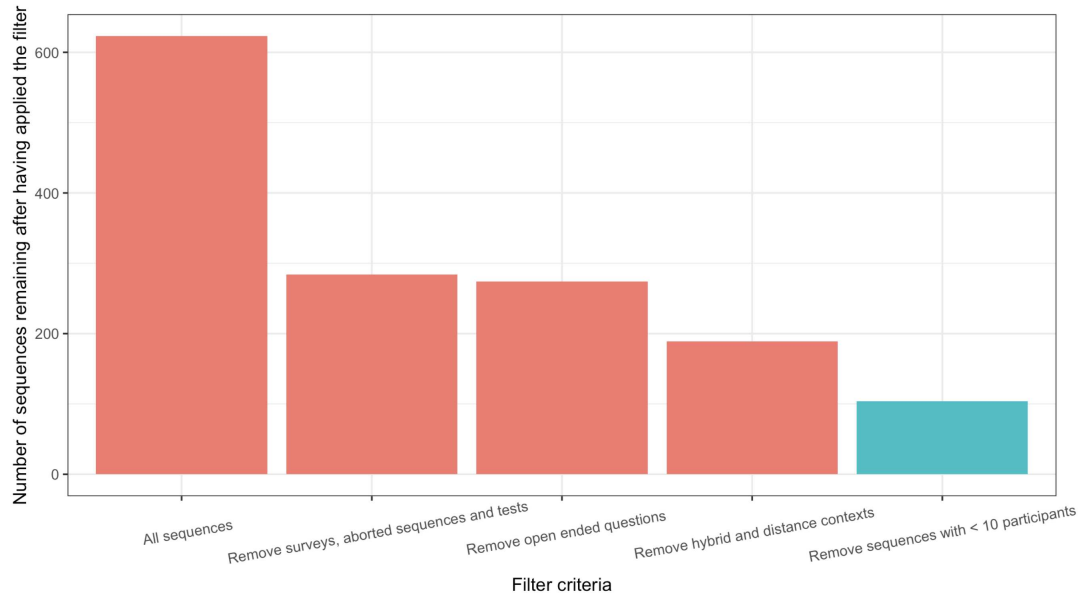


Figure 6.2: Summary of the filtering criteria applied to our initial sample. The blue bar represent the 104 sequences of the final sample for our data analysis.

6.2 Findings

6.2.1 Basis for our Analysis: Measuring the Benefits of a Sequence

First of all, we need to determine how we can measure the outcome of a sequence in terms of benefits. We considered the variables p_1 and p_2 which are the proportion of learners who answered correctly at the first and second vote respectively. As stated in subsection 3.3.2, sequences where $p_2 > p_1$ are qualified as *beneficial* because it means that students' general understanding of the topic has been enhanced [140]. Based on this property, we also qualify sequences where the proportion of correct answers does not change ($p_2 = p_1$) as *non beneficial* sequences. Finally, sequences where the proportion of correct answers decreases at the second vote ($p_2 < p_1$) are called *harmful* sequences. However, this classification is limited. It does not measure how beneficial a sequence is. As an example, we argue that a sequence where $p_1 = 10\%$ and $p_2 = 11\%$ is not as beneficial as a sequence where $p_1 = 10\%$ and $p_2 = 90\%$ even though, with this classification, they will both be simply labeled as beneficial. To measure the benefits of a two-votes-based sequence, Parmentier studied various indicators [115].

Topic	Number of sequences
Computer Science	72
Project Management	10
Law	5
Sociology	4
Physics	4
Psychology	3
Mathematics	2
Professionalization	2
History	1
Medicine	1

Table 6.1: Summary of the 104 sequences included in the dataset we analysed.

As rejected candidates, he studied the risk difference (see Equation 6.1), the risk ratio (see Equation 6.2) and odds ratio (see Equation 6.3). Indeed, these indicators have major drawbacks. First, they are not interval scales [143]. In other words, it does not take into account the value of the initial proportion of correct answers p_1 . For instance, when a sequence ends, if $RD = 40\%$, it is not precise enough to tell whether the proportion of correct answer went from 10% to 50% or from 50% to 90%. Secondly, RD and RR are limited by p_1 . As an example, RD can not be higher than 60% when $p_1 = 40\%$. In summary, these measures cannot be used to compare two interventions starting with two different initial scores, or to quantify how much an intervention is greater than another, even if they both start with the same initial score.

$$RD = p_2 - p_1 \quad (6.1)$$

$$RR = p_2/p_1 \quad (6.2)$$

$$OR = \frac{p_2}{1-p_2} / \frac{p_1}{1-p_1} = \frac{p_2}{1-p_2} * \frac{1-p_1}{p_1} \quad (6.3)$$

Based on the odds ratio and on item response theory [78], Parmentier proposed the estimation of Cohen's effect size shown in Equation 6.4.

$$d = 0.6 \ln \left(\frac{p_2}{1-p_2} \frac{1-p_1}{p_1} \right) \quad (6.4)$$

One notable property of this estimation is that the comparison between p_1 and p_2 can be inferred thanks to the sign of d . When the effect size is positive ($d > 0$), it means that the sequence is *beneficial* ($p_1 < p_2$). Similarly, when the effect size is null ($d = 0$), the sequence is *non beneficial* ($p_1 = p_2$).

Finally, when the effect size is negative (when $d < 0$), the sequence is *harmful* ($p1 > p2$). When applying s_{2019} to the diagram of the effect size as shown in Figure 6.3, we can use the rules of thumb [129] to interpret the benefits of each sequence in terms of very small to huge. We can see that Elaastic’s implementation of the two-votes-based process mainly leads to benefits for learners. Amongst 104 sequences, there are 65 beneficial sequences, 27 non-beneficial and 12 harmful. This observation reinforces the need to predict

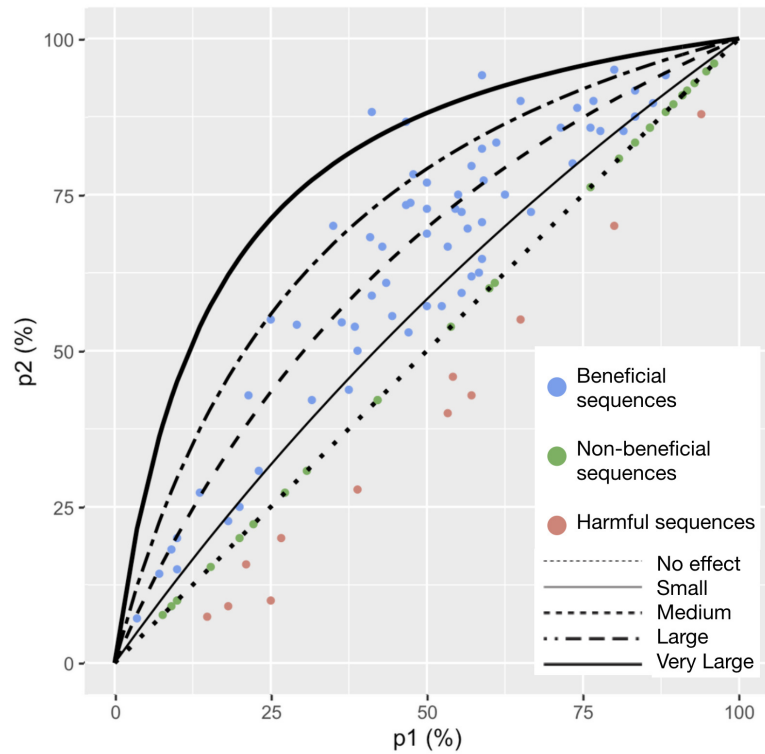


Figure 6.3: s_{2019} compared to isovalues of the Cohen’s d effect size: no effect ($d = 0$), small ($d = 0.2$), medium ($d = 0.5$), large ($d = 0.8$), and very large ($d = 1.2$).

the benefits of a sequence, and therefore to provide teachers with feedback during the sequence as stated in Section 3.3.3.

We can already use the benefits of a sequence to infer recommendations for teachers. When sequences are not beneficial, teachers should provide learners with deep and detailed explanations about the concepts involved in the question during the oral feedback. Consequently, we can infer the following recommendation:

Recommendation for Teachers Orchestration: After the second vote, explanations provided by teachers should be more detailed if the proportion of correct answers did not decrease or stagnated between the first and second votes ($d \leq 0$). ①

The next steps of our analysis focus on identifying variables of a sequence that are correlated with its effect size. For our data analysis, unless mentioned otherwise, we used the Spearman correlation instead of the Pearson one because (i) most of the variables involved in our correlations are bounded; (ii) we want to identify correlations that are not necessarily linear; (iii) the variables do not follow a normal distribution; (iv) the variables are ordinal, interval or ratio; (v) the variables are paired.

6.2.2 Proportion of Correct Answers during the First Vote

Let us begin with phase 1. In 2001, Crouch and Mazur stated that the improvement of student responses is largest when the initial percentage of correct answers is around 50%, and defined [35% : 70%] as the desired interval of $p1$ for optimal benefits of formative assessment sequences [40]. They argued 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. Based on these statements, we make the hypothesis that benefits of a sequence are linked to the distance between $p1$ and 50%.

Figure 6.4 shows the mean effect size of our sequences depending on the distance between $p1$ and 50%. As an example, the first bar represents 37 sequences where the distance of $p1$ from 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 (i.e. a medium effect size). The chart suggests that the effect size of a sequence decreases when the distance between $p1$ and 50% increases. We computed the Spearman correlation between $|p1 - 0.5|$ and obtained the following results: d is equal to -0.32 with a p-value = $8e - 4$ and a 95% confidence interval equal to $[-0.49 : -0.14]$, 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 sequences are more likely to be high when correct and incorrect answers are equally represented.

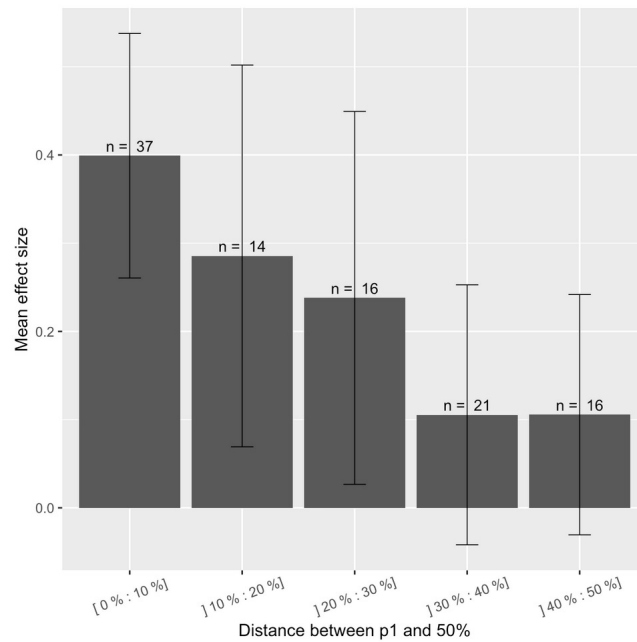


Figure 6.4: The effect size d of the sequences from s_{2019} depending on the distance between $p1$ and 50% (with 95% confidence intervals).

Recommendation for System Designers: Formative assessment systems implementing a two-votes-based process should provide teachers with the proportion of correct and incorrect 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%.

Even though Crouch and Mazur chose [35% : 70%] as their ideal interval, later works suggested [30% : 80%] as the threshold values [86]. And in 2010, Watkins and Mazur [97] noticed that their implementation of Peer Instruction is of high benefits for students when between [30% : 70%] of their first answers are correct. Finally, in our previous works [8], we chose the interval [20% : 80%] because of the significant decrease of mean effect size as shown in Figure 6.4. For these 4 proposed intervals from the literature, we ran a Wilcoxon rank sum test to compare the effect size of the sequence whose $p1$ is inside or outside the interval. Figure 6.5 shows the results of the test and the mean effect size d of the sequences from s_{2019} depending on whether $p1$ is inside the interval or not. The results are fairly similar for all the intervals of the literature. This is reinforced by Lasry [87] which stated that the threshold values of the ideal percentages of correct answers

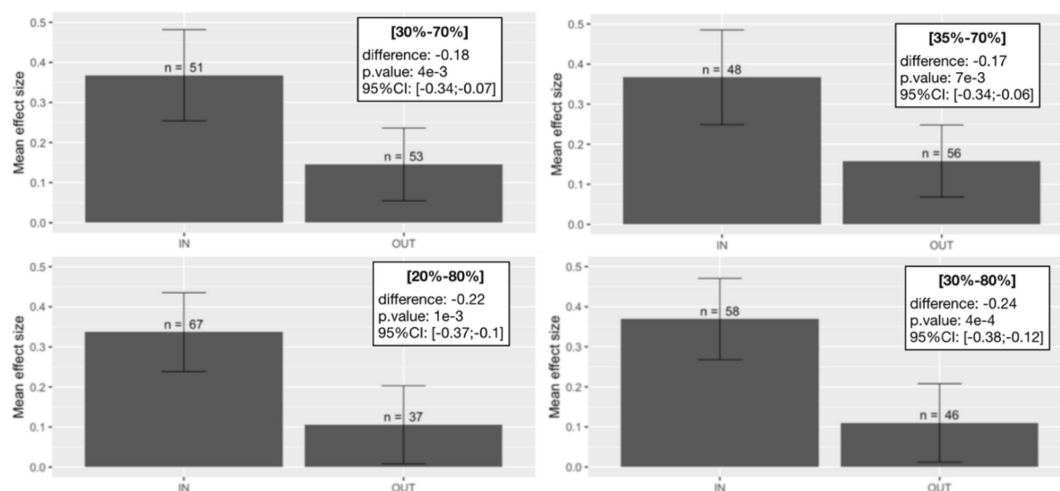


Figure 6.5: Mean effect size d of the sequences from s_{2019} depending on whether $p1$ is inside the interval or not, alongside the results of the Wilcoxon rank sum test.

are indicative and depend on the context.

So far, this result brought knowledge about two-votes-based process in general. However, we could not determine an ideal interval for $p1$ at the end of the first vote. The next section will provide additional evidences to solve this issue.

6.2.3 Learners Confidence Degree during the First Vote

The other information that is available at the end of the phase 1 is each learner confidence degree. Curtis used the confidence of learners about their answers as a way to profile learners beyond the simple correctness of their answers [43]. He defined "misinformed learners" as learners who are confident but provided an incorrect answer. In addition to that, he defined "uninformed learners" as learners who are not confident and provided incorrect answer. We can see with this classification that confidence of learners can help to make the difference between a learner that has a misconception and a learner who simply did not understand the course. Such difference implies that teachers need to behave differently depending on learners confidence degree. As Brooks stated [25], students' self-confidence in their knowledge can significantly affect how they interact and perform in the classroom. This leads us to believe that benefits of sequences could significantly increase when learners who provided correct answers are more confident than learners who did not.

Starting from this hypothesis, we propose an indicator to measure the consistency of learners confidence degree given the correctness of their answers. This indicator is meant to give an overview on how uninformed or misinformed learners are across a sequence. In order to measure the consistency of learners' confidence degree, we designed ρ_{conf} which can be computed by using the correlation between learners confidence degree and learners' understanding of the concept targeted by the question. Such understanding is measured with a binary variable, namely, the correctness of their first answers. Since these two variables are latent [54], the polychoric correlation is the adequate tool [112]. If learners who provided correct answers are confident whereas those who provided incorrect answers are not confident, ρ_{conf} will tend to be close to 1. Conversely, if learners who provided incorrect answers are confident whereas those who provided correct answers are not confident, ρ_{conf} will tend to be close to -1 .

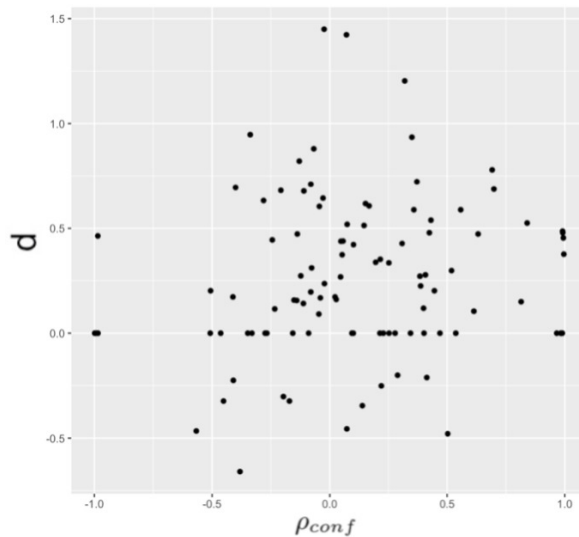


Figure 6.6: The effect size d depending on ρ_{conf} . Each point represents a sequence from s_{2019} .

Figure 6.6 is a plot diagram of d according to ρ_{conf} . The Spearman correlation between ρ_{conf} and d is 0.15 with a p-value = 0.13, and a 95% confidence interval equal to $[-0.05 : 0.33]$. These results are therefore inconclusive on this hypothesis.

We conducted deeper analysis of the benefits of a sequence depending on the combination of both ρ_{conf} and $p1$. Figure 6.7 shows the benefits of sequences depending on the proportion of correct answers at the end of first vote and the consistency of confidence degree (for each interval of the literature). Based on the mean effect sizes observed on this figure, ρ_{conf} seems to

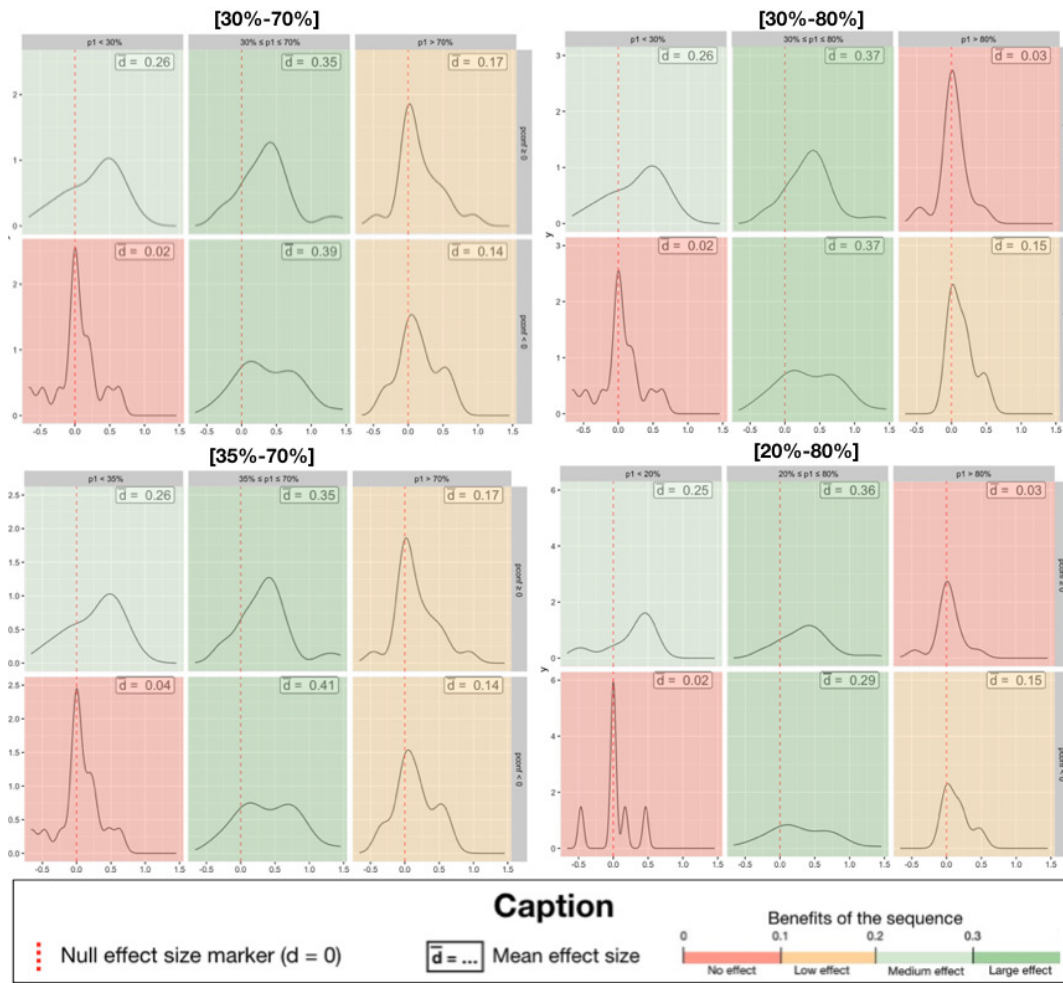


Figure 6.7: Kernel density estimates and mean of the effect size (d) of our sequences depending on the confidence consistency (ρ_{conf}) and the proportion of correct answers at the first vote ($p1$).

play a significant role when $p1$ is low regardless of the interval. Consequently, we propose the following hypothesis: when there are few correct answers at the end of the first vote, the benefits of sequences are higher when learners are consistently confident than when they are not.

To verify this hypothesis, we ran mean comparison tests. For $p1 < 20\%$, the effect size is not normally distributed (as computed by the Shapiro-Wilk test which returned a p-value equal to 0.045). Consequently, we used the Wilcoxon test and obtained a weakly significant difference (p-value = 0.15 and 95% confidence interval = $[-0.16 : 0.52]$). For $p1 < 30\%$, the effect size is normally distributed (as computed by the Shapiro-Wilk test which returned a p-value equal to 0.34) and both groups have homogeneous variances (F test returned a p-value equal to 0.34). We can therefore use the parametric t-test. We also obtained a weakly significant difference (p-value = 0.116 and 95% confidence interval = $[-0.07 : 0.55]$). For $p1 < 35\%$, the assumptions for t-test were also verified (the Shapiro-Wilk test returned a p-value equal to 0.31 and F test returned a p-value equal to 0.46). We obtained a weakly significant difference as well (p-value = 0.125 and 95% confidence interval = $[-0.07 : 0.52]$). These results led us to choose 30% as the low threshold for our recommendations regarding $p1$, which means that the remaining candidate intervals are $[30\% : 70\%]$ and $[30\% : 80\%]$. In order to decide between both these intervals, we examined Figure 6.7 and noticed that d is higher when $p1 > 70\%$ than it is when $p1 > 80\%$. This is due to the fact that sequences where $70\% < p1 < 80\%$ have a higher mean effect size ($\bar{d} = 0.362$) than sequences where $p1 > 80\%$ ($\bar{d} = 0.06$). This difference was proven to be significant thanks to a Wilcoxon test (p-value = 0.003 and 95% confidence interval = $[-0.05 : -0.23]$).

Based on this result and on previous works [155] detailed in Section 4.1, we proposed the following recommendations for orchestration.

Recommendation for Teachers Orchestration: At the end of the first vote:

- If there are a lot of correct answers ($p1 > 80\%$), teachers should skip the confrontation phase. ②
- When there are too few correct answers ($p1 < 30\%$), if learners are consistently confident ($\rho_{conf} \geq 0$), teachers should provide learners with hints before starting the confrontation phase. Else ($\rho_{conf} < 0$), teachers should provide detailed explanations and restart a new sequence to evaluate the same concept. ③

As a conclusion, ρ_{conf} is a relevant measure of learners' understanding

of the concept targeted by the question, beyond learners' correctness. In other words, learners' confidence degree has an impact on the outcome of a sequence.

Recommendation 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 especially according to such consistency.

6.2.4 Peer Grading Phase and Sequence Benefits

Moving onto phase 2, Double & al. argue that reflecting on peers answers is expected to lead to a higher percentage of correct answers [51]. Indeed, some studies support peer rating as a beneficial activity for learners, especially when it is conducted by a device [88, 163]. Peer rating allows learners to give feedback to peers and receive feedback from peers which both contribute to learning [49]. Since learners who provided correct answers are expected to convince those who did not thanks to the peer grading phase, we make the hypothesis that the consistency of the peer grading phase is linked to the sequence benefits.

Similarly to ρ_{conf} , we used ρ_{peer} to measure learners' consistency of peer grading. It is measured through the polychoric correlation between the level of agreement given by a peer to a rationale (see Likert scale on Figure 5.3), and the level of understanding of the learner who wrote the rationale (measured through correctness of the matching answer). 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 6.8 shows a plot diagram of the effect size d depending on ρ_{peer} . The Spearman correlation between ρ_{peer} and d is 0.34 with a p-value $< .002$ and a 95% confidence interval equal to $[0.13 : 0.52]$, which supports our hypothesis. When $\rho_{peer} < 0$, it means that incorrect answers are more popular than correct answers which should be addressed by teachers.

Let us note that ρ_{peer} is not significantly correlated to the distance between $p1$ and 50%. Indeed, the correlation we calculated returned a p-value equal to 0.19. We used the correlation of Kendall since it is a substitute for Spearman correlation that handles ties better. The assumptions for this statistical test are met because both variables are continuous and we wanted

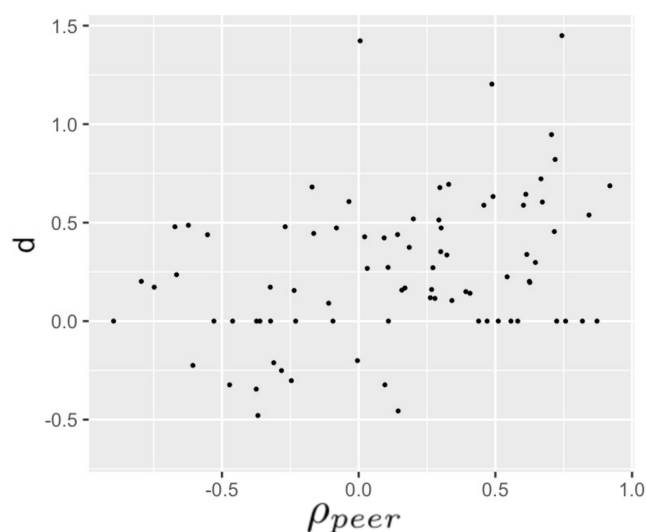


Figure 6.8: Effect size d depending on the consistency of peer gradings ρ_{peer} . Each point is a sequence from s_{2019} .

to identify a monotonous relation between them. Consequently, ρ_{peer} and $|p1 - 50\%|$ (see Section 6.2.2) are two independent predictors of the benefits of a sequence. Thus, this analysis makes it possible to propose the following recommendations.

Recommendation for System Designers: Formative assessment systems implementing a peer grading process should provide teachers with the consistency of peer grading and feature flexibility regarding the selection of the rationales in the focus of the discussion (phase 3), especially according to the consistency of peer grading.

Recommendation for Teachers Orchestration: If peer grading is inconsistent ($\rho_{peer} < 0$), teachers should focus on rationales for incorrect answers during the discussion. Else ($\rho_{peer} \geq 0$), teachers should focus on correct rationales during the discussion. ④

6.2.5 Peer Grading Phase and Learner's Confidence Degree during the First Vote

In Section 6.2.3, we identified no correlations between the consistency of confidence degree and the benefits of a sequence. However, in Section 6.2.4, we identified a correlation between the consistency of peers grading and a

sequence's benefits. These results led us to perform deeper studies about these two consistency. We want to verify the following assumption: misinformed learners are not able to consistently grade peers rationales. As a consequence, we make the hypothesis that consistency of peer gradings is linked to the consistency of learners confidence degree.

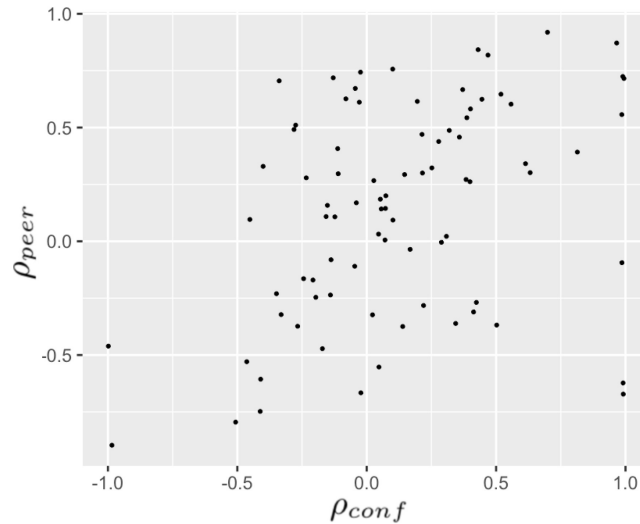


Figure 6.9: Peer grading consistency ρ_{peer} depending on the confidence consistency ρ_{conf} . Each point is a sequence from s_{2019} .

Figure 6.9 is a plot diagram of ρ_{peer} according to ρ_{conf} . The Spearman correlation between ρ_{conf} and ρ_{peer} is 0.37 (p-value = $5e - 4$ and 95% confidence interval = $[0.17 : 0.55]$), which supports our hypothesis.

Based on these results, we argue that ρ_{conf} can serve two purposes. At the end of the first vote, it can be used as a way to predict whether ρ_{peer} is expected to be high or not, and therefore determine whether the confrontation phase should be skipped or not (which supports recommendation for orchestration 2 and 3 of Section 6.2.3). And, during the debriefing phase, it can be used as a substitute for ρ_{peer} when the latter is not available (which typically is the case when the confrontation phase was skipped).

Recommendation for System Designers: Formative assessment systems implementing a two-votes-based process should feature flexibility regarding the selection of the rationales in the focus of the discussion (phase 3) according to the consistency of confidence degree.

Recommendation for Teachers Orchestration: When the confrontation step was skipped as well as the second vote step:

- If learners are inconsistently confident ($\rho_{conf} < 0$), teachers should focus on rationales for incorrect answers during the discussion.
- Else ($\rho_{conf} \geq 0$), teachers should focus on rationales for correct answers during the discussion. ⑤

6.2.6 Self-Grading as a Substitute for Peer Grading

Regarding factors about peer interactions, some studies about self-grading [51, 98] provide support for its use as a formative practice to improve performances. Consequently, we make the hypothesis that there is a relationship between the amount of self-rated students and the benefits of a sequence.

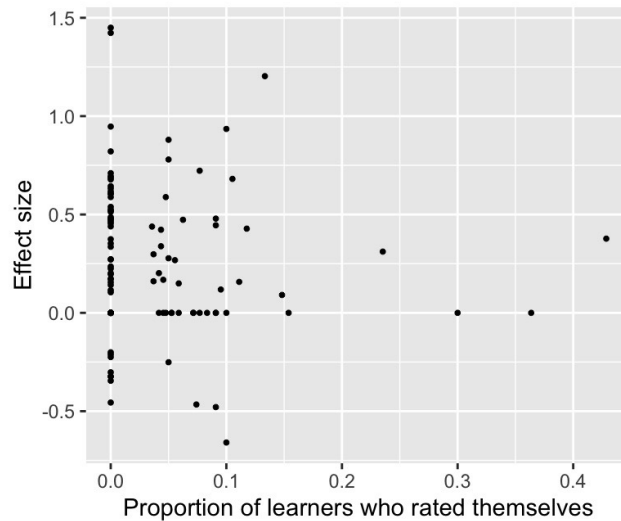


Figure 6.10: Effect size d depending on the percentage of learners who graded themselves. Each point is a sequence from s_{2019} .

Our results suggest that self-grading tends to nullify the effect size (see Figure 6.10). The Spearman correlation between the effect size and the

percentage of learners who performed self-grading returned a p-value equal to 0.2 and a 95% confidence interval = [0.3 : 0.06]. In conclusion, there is no significant relations between both variables. We explored the data and found out that learners who graded themselves during the confrontation of viewpoints tend to give their rationale the highest grade whether their answer was correct or not. Based on s_{2019} , we compared grades given when learners graded themselves with grades given when learners graded peers (see Figure 6.11).

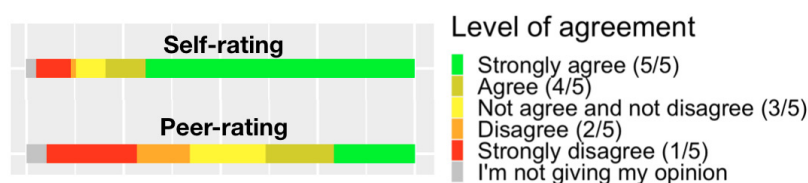


Figure 6.11: Stacked bar chart of the grade attributed during our sequences depending on the type of grading.

Since the grades from s_{2019} do not follow a normal distribution and the observation within self- and peer- grading are independent, we used the non-parametric Wilcoxon-Mann-Whitney test. The difference in means was significant (95% CI = [-2:-1] and p-value = $2.3e-15$). Furthermore, self-grading was less consistent ($\rho_{self_r} = 0.139$) than peer grading ($\rho_{peer_r} = 0.219$).

These results reject our hypothesis and suggest that self-grading does not benefit to learners within peer grading activities where learners have to grade a set of explanations in a row and on a same page. An informal discussion with 9 learners has been conducted and allowed us to make two 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 grades are more likely to be noticed by teachers. Therefore, they game the system in order to receive oral feedback from teachers during phase 3. In other words, learners perceive this activity as competitive to catch teacher's attention.

Recommendation for System Designers: Peer grading activities in formative assessment systems should not authorize self-grading as a substitute for peer grading.

When self-assessment was initially implemented within Elaastic's evaluation phase, it was expected from learners to perform consistent self-assessment. The main hypothesis was that a learner is supposed to give herself a more adequate grade after she evaluated some of her peers' rationales because these

evaluations were expected to improve her knowledge compared to the one she possessed when she provided her initial answer. Deeper analysis need to be conducted in contexts where the competitive settings are reduced. One of the options would be settings where learners see their own rationale a few moments after their peers' rationales, and where self-grading is not taken into account to calculate the mean grade of a given answer.

6.2.7 Peer Grading per Learner as Configured by Teachers

Confrontation of viewpoints 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 [40], whereas others involved teachers in the collective confrontation [110]. Therefore, we want to explore the impact of the number of learners involved in such a confrontation. With Elaastic, the number of learners involved in viewpoints confrontation is represented by the number of peers rationales graded by each learner. The number of rationales each learner grades is chosen by teachers when configuring the sequence (see Section 5.1). We believe that the effect size of a sequence is related to such a number.

Since there were not enough sequences configured with 1 grade on one side, and with 4 grades on the other, we ran a statistical test with various grouping methods. Table 6.2 summarises the results of the mean effect size comparison of two groups of sequences. As an example, the first row shows the results of the two-sample test between sequences where each learner had to grade 1 or 2 rationales (group 1) and sequences where they had to grade 3 rationales (group 2). For each grouping method, we used the parametric t-test when the effect size is normally distributed (determined thanks to the Shapiro-Wilk normality test) and the non-parametric Wilcoxon-Mann-Whitney test otherwise. As shown on the last column of Table 6.2, the number of learners involved in the peer grading activity has no significant impact regardless of the grouping method, which rejects our hypothesis.

Recommendation for System Designers: Formative assessment systems should feature flexibility regarding the number of peers involved in confrontation of viewpoints.

Group 1			Group 2			two-sample test	
nb grades given	mean	sd	nb grades given	mean	sd	95% CI	p-value
1, 2	0.18	0.39	3	0.26	0.42	$[-0.13 : 0.3]$	0.42 (*)
1, 2	0.18	0.39	4, 5	0.29	0.34	$[-7e - 5 : 0.3]$	0.17
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	$[-0.28 : 3e - 5]$	0.17
1, 2, 4, 5	0.25	0.36	3	0.26	0.42	$[-0.14 : 0.16]$	0.75
1, 2, 3	0.23	0.41	4, 5	0.29	0.34	$[-0.2 : 0.07]$	0.41

Table 6.2: Results of the two-sample test with various grouping methods of our sequences (* with parametric t-test).

Recommendation for Teachers Orchestration: Teachers can decide the number of peers involved in viewpoints confrontation.

6.3 Discussion

6.3.1 Regarding the Research Questions

Our first research question was *"Which useful information can be inferred from the analysis of data gathered from a tool used in authentic contexts?"*. The analysis that we performed allowed us to answer it by identifying useful correlations between various aspects of learners' behavior along the two-votes-based process. In brief, our results suggest that benefits of two-votes-based sequences depend on the proportion of correct answers' distance from 50% at the first vote as well as the consistency of peer grading. Confidence consistency of learners also plays a role in determining the consistency of peers' evaluation during the confrontation phase.

Our second research question was *How can such information contribute to support orchestration of formative assessment processes?* We provided answers to this question by suggesting recommendations for orchestration to teachers. Moreover, based on prior works, our recommendations can be used to provide an evidence-based orchestration model described in the next subsection.

6.3.2 Implications for Research and Practice: Improving Vickrey's model

Based on the metrics we identified, our goal is to design interventions that are intended to help teachers make decisions that benefit to learners. In prior works, Vickrey proposed a model designed to support orchestration of Peer Instruction [155]. With such a model as a basis, we introduced decision nodes according to the different findings presented before. Figure 6.12 exposes these nodes, and summarises our deterministic recommendations for orchestration of formative assessment sequences. At the end of the first vote,

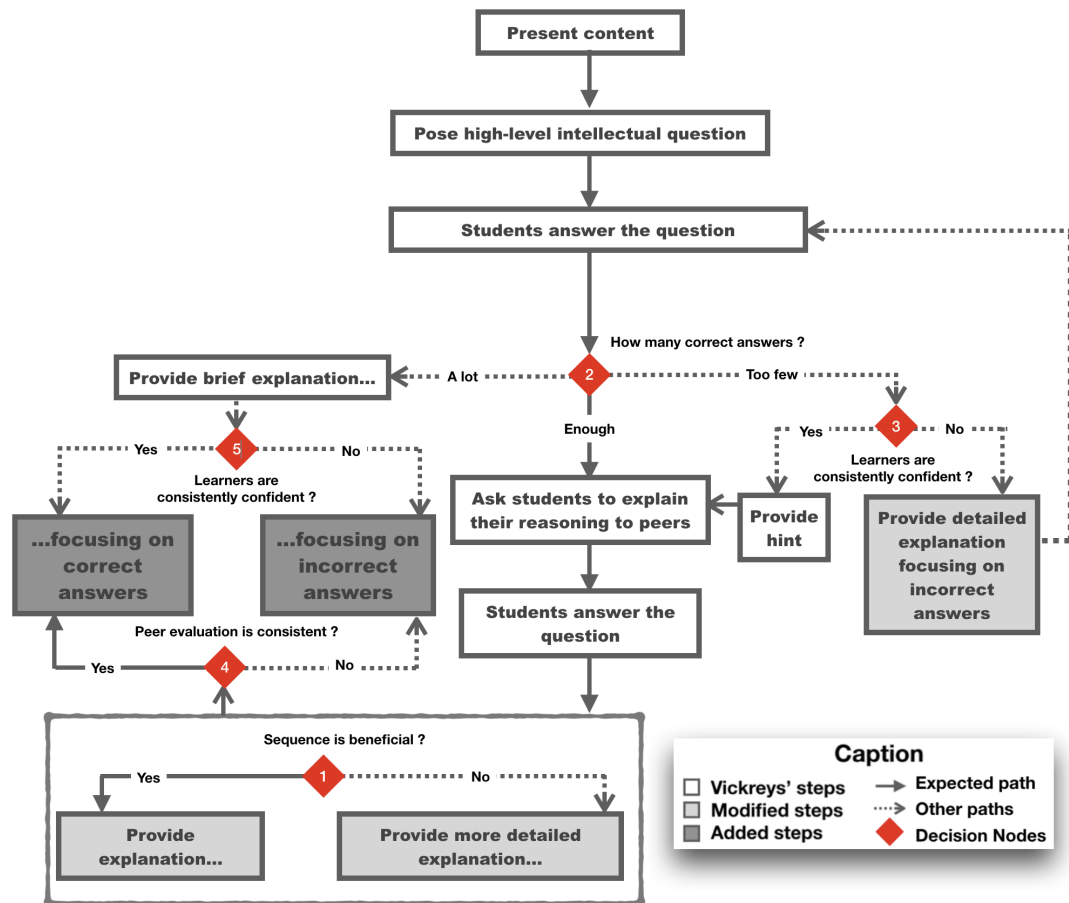


Figure 6.12: Orchestration model of two-votes-based processes based on [155]. Each white number represents the matching recommendation for orchestration.

two indicators are available: the proportion of correct answers as well as the consistency of learners' confidence degree. Therefore, decision nodes labeled

2, 3 and 5 recommend teachers to take different actions depending on their value. Teachers are recommended to skip phase 2 when there a lot of correct answers or when there are few correct answers and learners are inconsistently confident. Skipping phase 2 is expected to prevent sequences from having a negative outcome such as ending up. Another possible recommendation for teachers at the end of the first vote is not to skip the confrontation phase but to provide learners with hint beforehand. This is the case when there are few correct answers but learners are consistently confident. Similarly, at the end of the second vote, two more indicators are available: the benefits of the sequence as well as the consistency of peer evaluation. Therefore, decision nodes labeled 1 and 4 recommend teachers to lead the discussion phase one way or another depending on their value. Such discussion must be detailed if the sequence is not beneficial. It must also focus on correct answers when peer grading is consistent or when there was no confrontation phase and learners confidence degree was consistent.

6.3.3 Limitations

The main limitations of our study come from the dataset itself. The 104 sequences we analysed address mainly STEM-related topics from higher education courses. A broader study including sequences from various topics and educational levels could help to refine our findings. Furthermore, the analysis only identified significant but fairly weak correlations (~ 0.3).

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 of her answers are considered as wrong (because their score is not equal to 100), and the information stating that she improved is lost. Even though multiple choice questions are only a small portion of s_{2019} ($\sim 10\%$), 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 considered rationales associated to correct answers as correct rationales. However, learners can answer correctly and provide incorrect rationales. Consequently, if learners give a low grade to an incorrect rationale corresponding to a correct answer, ρ_{peer} will decrease even though this rationale was rightfully given a low grade. Such a possibility is not addressed by our works regarding the quality of peer interactions.

6.4 Conclusion of the Empirical Study

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 interaction data to answer our research questions. We answered RQ1 by highlighting new understandings of formative assessment, and providing system designers with evidences intended to help them to design a formative assessment system. We answered RQ2 by identifying useful indicators to assist teachers when orchestrating a face-to-face formative assessment sequence, and inferring a deterministic orchestration model designed for synchronous settings based on a previous orchestration model. The next part of this manuscript will focus on implementing and evaluating this new process.

Part IV

Second Iteration: Evaluation of the New Orchestration Model

Chapter 7

Implementation

Contents

7.1	Enabling The New Orchestration Model	79
7.1.1	Implemented Orchestration Features	80
7.1.2	Providing Teachers with Relevant Indicators	82
7.2	Implementation of Explainable Recommendations	85
7.2.1	Content of the Explanations	85
7.2.2	Presentation of the Explanations	89
7.3	Action Tracking	90
7.4	Additional documentation	91

This chapter describes *Elaastic* 5.0, the new version of *Elaastic* that we implemented to evaluate the new orchestration model. More precisely, Section 7.1 describes features that would enable teachers to conduct sequences according to our orchestration model. Section 7.2 shows the implementation of recommendations for orchestration as well as the way they are explained. Then, Section 7.3 details internal features of the system that will allow us to track teachers' interaction with the system. The last section is 7.4. It details the features we implemented to facilitate usages of *Elaastic* and therefore encourage teachers to incorporate it within their formative assessment practices so that we can collect as much data as possible.

7.1 Enabling The New Orchestration Model

Implementing our new orchestration model within *Elaastic* would require teachers to be able to make specific orchestration decisions in accordance with

the new model, such as skipping a phase or selecting rationales according to a specific criterion.

7.1.1 Implemented Orchestration Features

First of all, as shown in Figure 7.1, we added a button in the steps panel to allow teachers to skip phase 2 (which is the confrontation step and the second vote step) at the end of phase 1 and directly start phase 3 (which is the discussion step).

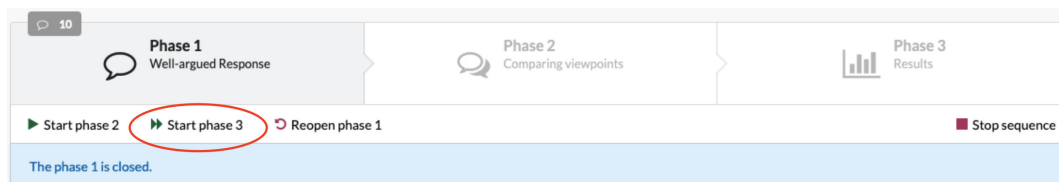


Figure 7.1: Button to skip phase 2 and directly enter into phase 3 after the end of phase 1.

We also provided additional features regarding the selection of rationales by teachers. First of all, the rationales that appear under the diagrams are now the ones identified by the recommendation model. Let us note that, by default, best graded rationales for correct answers used to be the ones shown to users (see Section 5.3).

In addition to that, we added a new set of rationales in the rationale popup window that was originally meant to display all rationales. As shown in Figure 7.2, it gathers all the recommended explanations filtered and ordered according to our decision making model. As an example, when teachers are recommended to focus on rationales related to incorrect answers because $\rho_{conf} < 0$, this set of rationales contains those related to incorrect answers ordered by descending confidence degree. When teachers are recommended to focus on rationales related to the correct answers because $\rho_{peer} > 0$, this set gathers the matching rationales and orders them by descending mean evaluation provided by peers. Since the confidence degree sometimes is a meaningful criterion to determine which rationale should be at the focus of the discussion phase, the interface of the teacher now displays the learner's confidence degree for each rationale, as shown in Figure 7.3. On this figure, we can now see that Joe Walson selected answer 1 and selected "Confident" as his confidence degree.

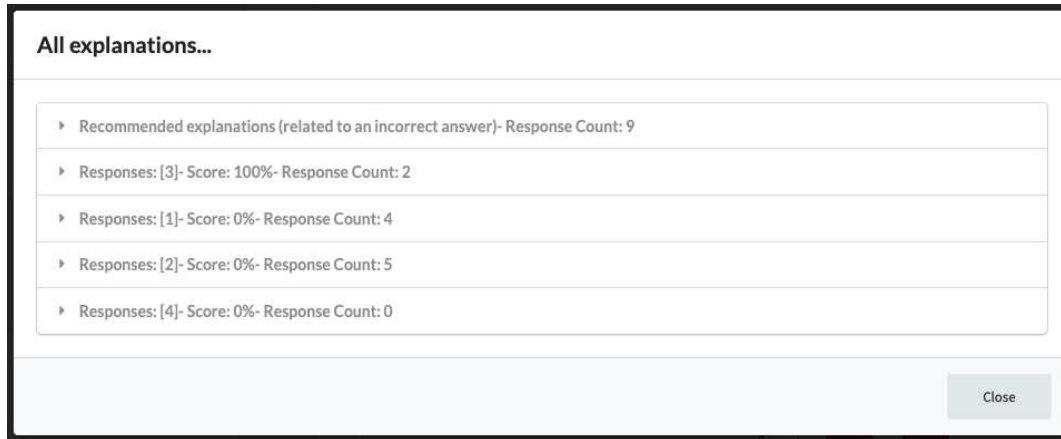


Figure 7.2: New set of rationales to display the recommended rationales.

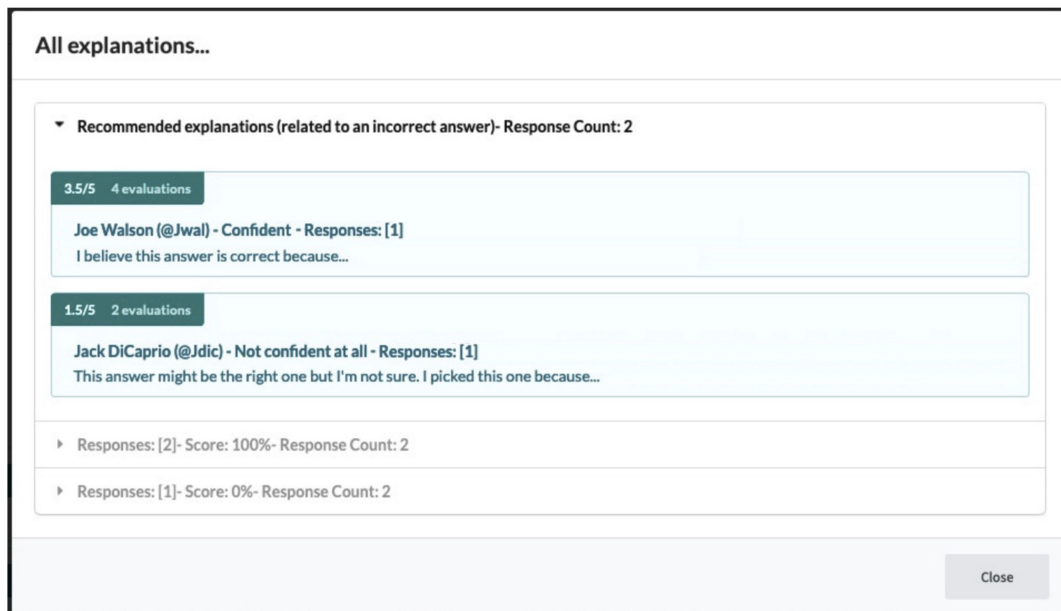


Figure 7.3: Displaying the writer's confidence degrees on each rationales

7.1.2 Providing Teachers with Relevant Indicators

In order to make decisions, teachers must be provided with the value of some indicators that are relevant to the decision they are expected to make. As an example, according to decision node 5 of our orchestration model shown in Figure 6.12, teachers must be able to directly engage in phase 3 (discussion step) without having to go through phase 2 (which is the peer grading step and the second vote step). To do so, they need to be provided with ρ_{conf} and $p1$ that represent the relevant indicators determining the next decision to make. To this end, we designed histograms to expose our indicators and their values. When it comes to $p1$ and d , graphical representations already exist within Elaastic (see Figure 7.4 for $p1$ and Figure 7.5 for d). Based

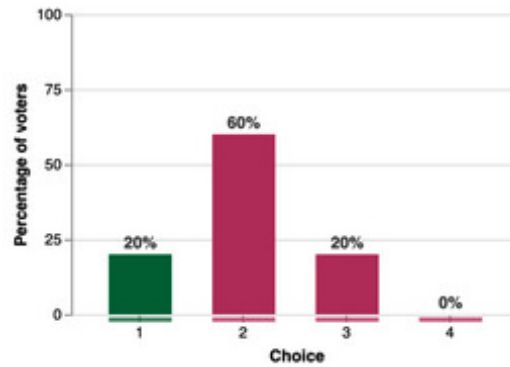


Figure 7.4: Design to illustrate $p1$

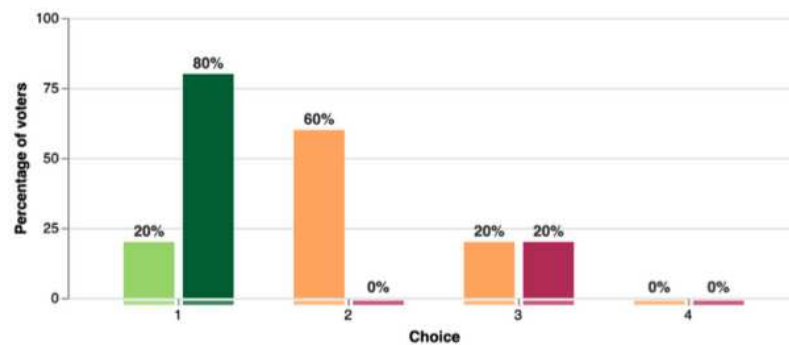


Figure 7.5: Design to illustrate d

on these histograms, teachers can easily see an infer relevant such as *“The second vote has more correct answers than the first one.”*

However, the previous version of Elaastic did not include visualisations for ρ_{peer} and ρ_{conf} . Hence, new graphical representations summarising learners confidence degree and peer evaluations had to be designed. To this end, we originally based our design on Nash’s paper about the best way to plot data based on Likert scales [126]. In these works, the chosen design is a diverging stacked bar chart. Figures 7.6 and 7.7 shows such initial design in order to illustrate respectively ρ_{peer} and ρ_{conf} .

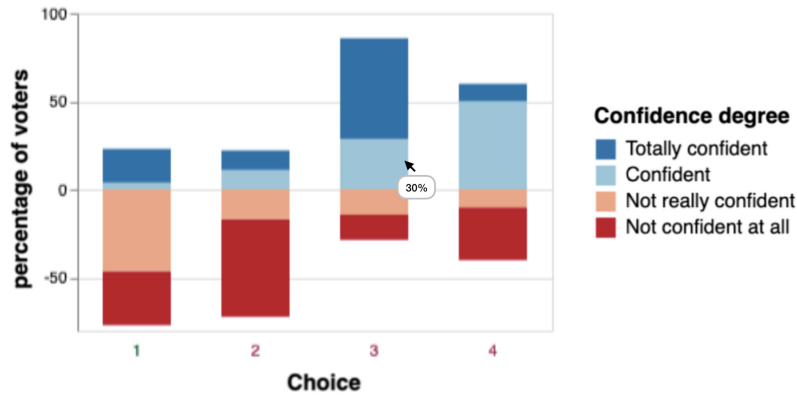


Figure 7.6: Initial design to illustrate ρ_{conf} .

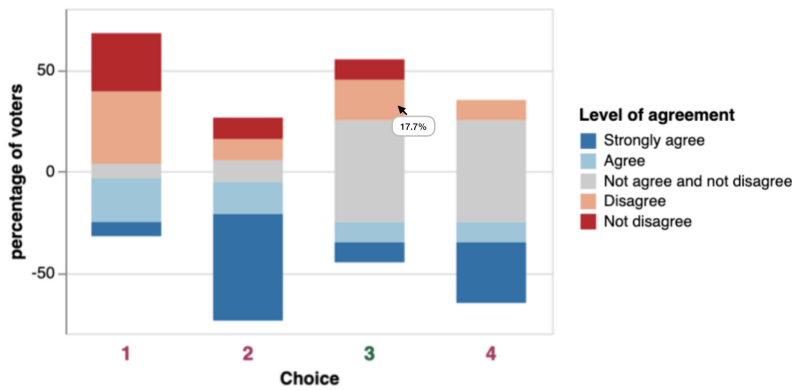


Figure 7.7: Initial design to illustrate ρ_{peer} .

We collected users’ opinion about this design during a workshop with 3 teachers in secondary education and obtained the following feedback:

- *”The diagram is hard to understand because the bottoms of the bars are not visually aligned.”*
- *”It’s counter-intuitive to have negative percentages on the y-axis.”*

- *"The design is not a satisfying explanation regarding the recommendation because I can hardly compare the mean confidence degree of learners who provided the correct answer with the mean confidence degree of learners who didn't. I have to mentally combine the bars of all the answers grouped by correctness."*

The reason why Nash's proposal was not satisfying might come from the difficulty to compare percentages across different groups. As an example, it is hard to compare the percentages that represent learners who provided correct answers and reported that they are *"Totally confident"*, with those who reported the same confidence degree but provided incorrect answers. This might be due to the fact that the bold blue bars representing these information are not next to each other, and can hardly be compared when it comes to their sizes. We thus designed an alternative solution presented in Figures 7.8 and 7.9, which are grouped bar charts. The two groups of correct and incorrect answers face each other, and a tooltip allows to show the exact numeric percentages of each bar.

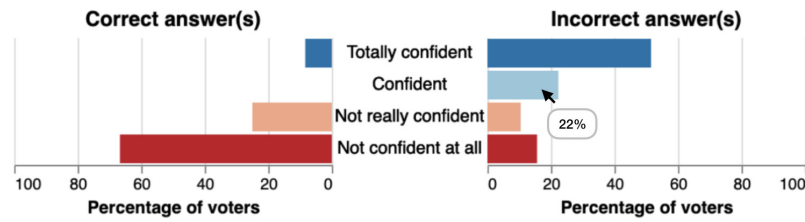


Figure 7.8: Final design to illustrate ρ_{conf} .

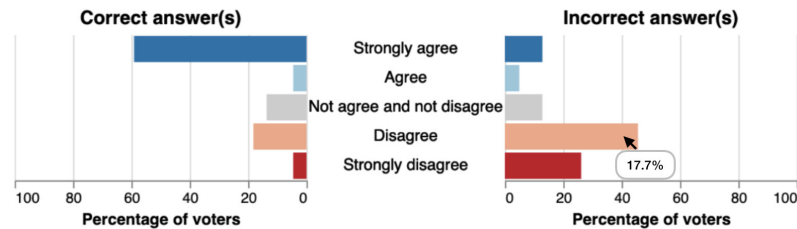


Figure 7.9: Final design to illustrate ρ_{peer} .

We made these histograms available in the main result panel of the teacher interface, as shown in Figure 7.10.

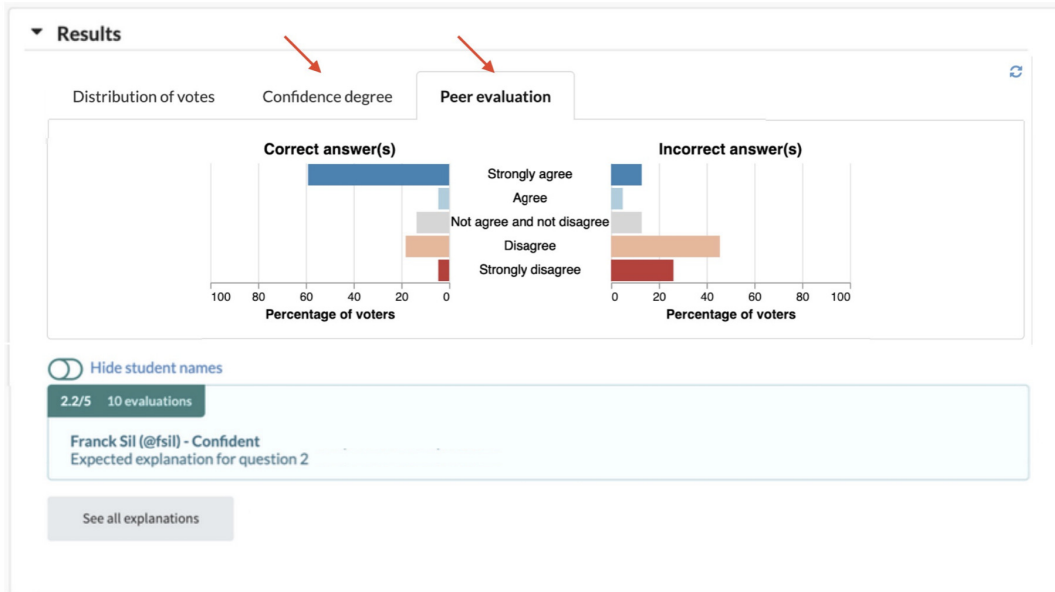


Figure 7.10: Histograms added to the results frame of the teacher interface.

7.2 Implementation of Explainable Recommendations

However, enabling such features do no guarantee that teachers will make a decision in accordance with the new model. Consequently, we must enable a feature that provides teachers with recommendations for orchestration. Such recommendations imply that our analysis on the benefits of our recommendation model is heavily dependent on the way teachers will perceive them. Related works about *explainability* emphasized the importance of being able to justify recommendations provided to users of a system [124] in order to earn their trust. Therefore, we propose to provide explanations alongside the recommendations since it is expected to make teachers trust the system more [147].

7.2.1 Content of the Explanations

Recommendations could be explained by comparable items that the user is familiar with [130]. In our case, such items would be other sequences. Specifically, the relevant sequences are the ones we analysed to design our orchestration model. Therefore, the associated explanation to a recommendation could be that we observed other similar sequences in our study, and that those sequences did not lead to the best learning results.

However, our indicators are based on latent variables that are hard to explain and therefore have no intuitive meaning [32]. In other words, simply providing teachers with the values of indicators in natural language as explanations would not be a satisfying explanation. As an example, when teachers are recommended to provide learners with hint before engaging them in the confrontation phase, the explanation "*There are less than 30% of correct answers and learners confidence degrees are consistent.*" is hard to understand because the consistency of confidence degree is based on latent variables. As a solution, we can leverage the properties for particular values of our indicators to design more understandable explanations.

Our second concern is that such an explanation leans heavily on the quantitative aspect of our study. Some studies argued that explanations based on probability and statistical generalisation do not satisfy users [105]. Consequently, we decided to provide teachers with explanations by adding qualitative arguments supported by graphical representations. The qualitative explanations we can use are the ones from the literature from which our hypotheses of Section 6.2 originated from. As an example, when there are a lot of correct answers, we should provide teachers with this information and with comparable sequences, but we should also provide the following qualitative explanation: *Learners have enough knowledge to answer correctly alone.*

Based on these proposals, Table 7.1 and 7.2 summarise the explanations for every relevant case of our orchestration model.

Indicator	Case	Property	Qualitative explanation
p1	< 30%	Less than 30% of the answers are correct.	Learners might not have enough knowledge about the topic.
	> 70%	More than 70% of the answers are correct.	Learners might have enough knowledge to answer correctly alone.
ρ_{conf}	< 0	The confidence degrees reported by students are not consistent.	Learners who provided a correct answer are less confident than those who provided an incorrect one, which means that the rationales might not be correctly evaluated by peers during phase 2.
	> 0	The confidence degrees reported by students are consistent.	Learners who provided a correct answer are more confident than those who provided an incorrect one, which means that the rationales might be correctly evaluated by peers during phase 2.
	= 0	Learners who provided a correct answer have an identical mean confidence degree than those who provided an incorrect one.	Learners who provided a correct answer are as confident as those who provided an incorrect one, which means that the rationales might be correctly evaluated by peers during phase 2.

Table 7.1: Explanations at the end of the first vote.

Indicator	Case	Property	Qualitative explanation
ρ_{conf}	< 0	The confidence degrees reported by students are not consistent.	This means that rationales for incorrect answers are associated to higher confidence degrees than the others.
	> 0	The confidence degrees reported by students are consistent.	This means that rationales for correct answers are associated to higher confidence degrees than the others.
	$= 0$	The confidence degrees reported by students are consistent.	This means that rationales for incorrect answers are associated to confidence degrees that are, in average, equal to those associated to rationales for correct answers.
ρ_{peer}	< 0	Rationales associated to a correct answer have a lower mean grade than those associated to an incorrect one.	This means that correct answers are less popular than incorrect ones.
	> 0	Rationales associated to a correct answer have a lower mean grade than those associated to an incorrect one.	This means that correct answers are more popular than incorrect ones.
	$= 0$	Rationales associated to a correct answer have an identical mean grade than those associated to an incorrect one.	This means that correct answers are as popular as incorrect ones.
d	< 0	The second vote has less correct answers than the first one.	This means that the evaluation phase was harmful to learners.
	$= 0$	The second vote has as many correct answers than the first one.	This means that the evaluation phase was not beneficial to learners.
	> 0	The second vote has more correct answers than the first one.	This means that the evaluation phase was beneficial to learners.

Table 7.2: Explanations for the discussion phase.

7.2.2 Presentation of the Explanations

Even though explanations are expected to satisfy users' curiosity, they can actually have undesired outcomes such as suppressing such curiosity or reinforcing flawed mental models [71]. This can happen when the explanations provide too much details [71]. Therefore, we propose that teachers choose the level of details regarding the explanations for recommendation by making them optional. As shown in Figure 7.11, a recommendation can be provided alongside a short explanation. If the curiosity of the teacher is not satisfied, she can click on "Read more" to be provided with a more detailed explanation. Figure 7.12 shows an example of such explanation.

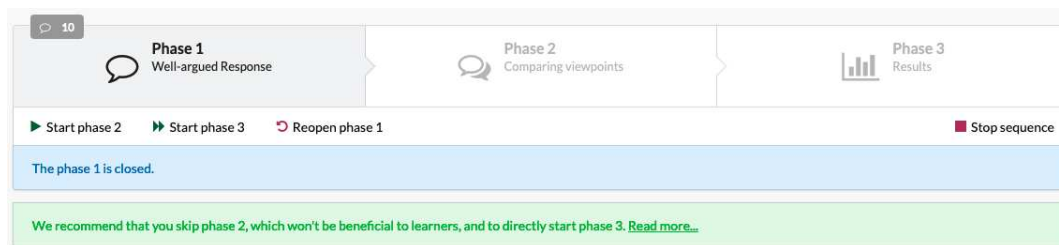


Figure 7.11: Recommendation: 1st level of detail for the explanations

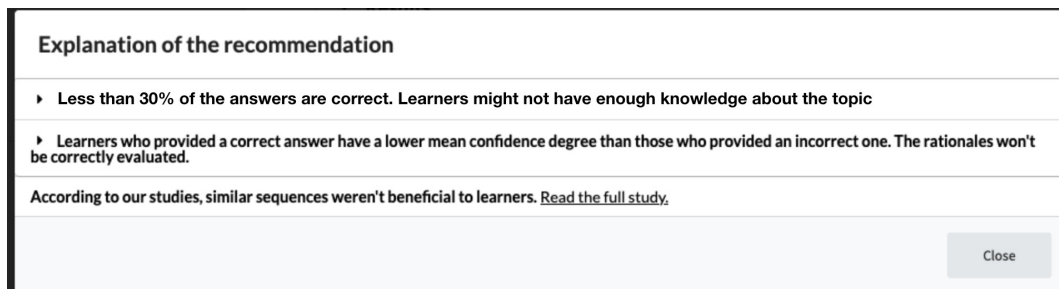


Figure 7.12: Recommendation: 2nd level of detail for the explanations.

By clicking on the various statements of this textual explanation, teachers can be shown the histogram we designed to summarise the relevant indicator. As an example, Figure 7.13 shows the diagram that is meant to illustrate that $\rho_{conf} < 0$.

Finally, if the curiosity of the teacher is still not satisfied, she can click on "Read the full study" to read the paper we published to justify our model [8].

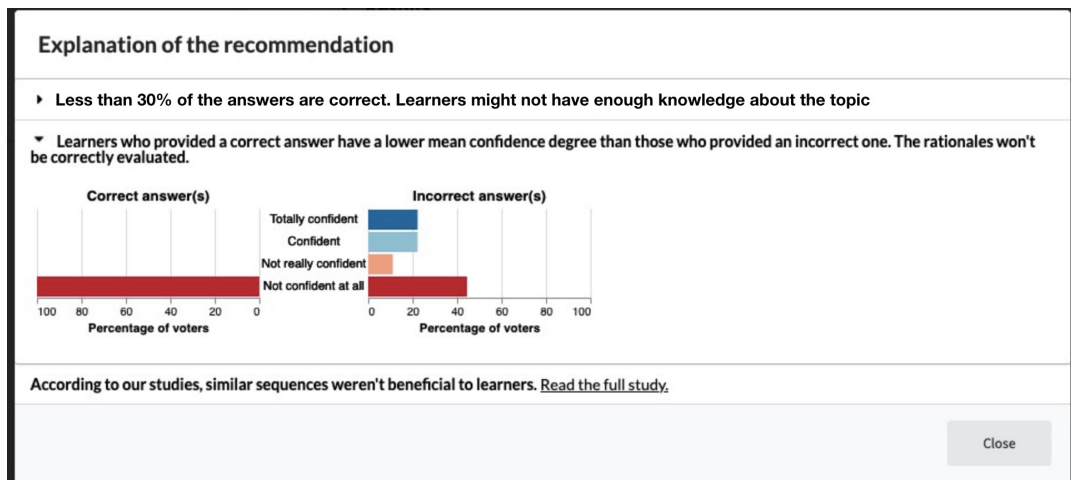


Figure 7.13: Recommendation: 3rd level of detail for the explanations.

7.3 Action Tracking

In order to facilitate our data analysis, we also implemented a tracking feature in our new version of Elaastic. Teachers' clicks, actions and decisions are collected and stored in the database. The new actions that are now recorded are the followings:

- Decisions when orchestrating the sequence such as starting, stopping or skipping a phase.
- Popups opened such as the explanations or the one that contains all learners' rationales.
- Navigation through the main result frame such as clicking on a tab to see a specific histogram or updating the results.
- Navigation through any popup such as clicking on a set of rationales to show or hide a content.

These actions are being saved in a format designed to facilitate conversion to xAPI format [14]. Figure 7.14 shows a simplified xAPI statement schema. We based our design on this figure and implemented a feature that saves the actions as described in Table 7.3: All the other elements of the "context" part of the xAPI schema can be retrieved from the sequence id.

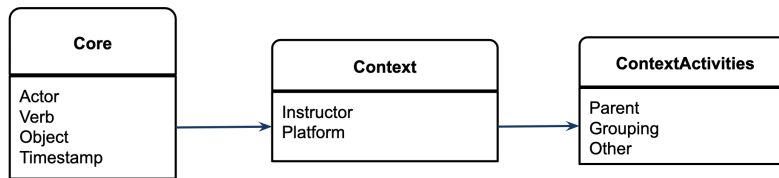


Figure 7.14: Simplified xAPI statement schema.

xAPI equivalent	Data label	examples
actor	user id	12, 18, 200
	role	Teacher, learner
verb	action	open, close, start, skip, stop, click, update
object	object	rationales popup, results, phase 2
timestamp	date and time	2022-05-24 16:36:01, 2015-12-01 00:00:45
context	sequence id	21, 23, 1

Table 7.3: Format of the actions saved.

7.4 Additional documentation

Finally, to improve usability for teachers, we added documentation within the platform. As Figure 7.15 shows, a new button appears on the left menu. By clicking on this button, two options appear. Teachers can trigger an onboarding [99] sequence by clicking on "Quick start", or go to the online documentation by clicking on "Online help" (see Figure 7.16).

In this chapter, we presented how the new orchestration model has been implemented within the new version of Elaastic. We involved final users in some of critical parts of the implementation approach with a small focus group, while taking into account insights from the literature. In the next chapter, we analyse data issued from the usage of Elaastic 5.0.

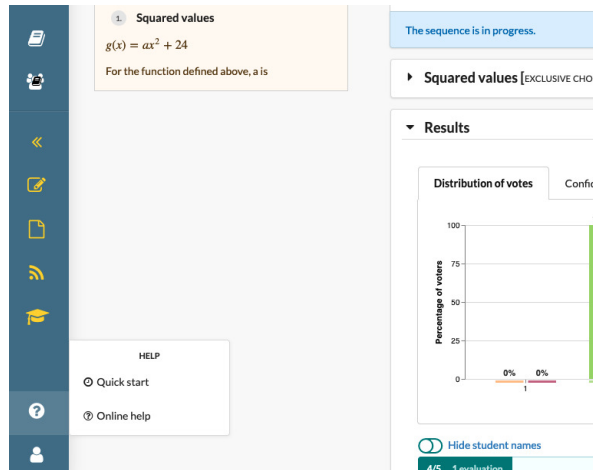


Figure 7.15: Help button on the left menu.

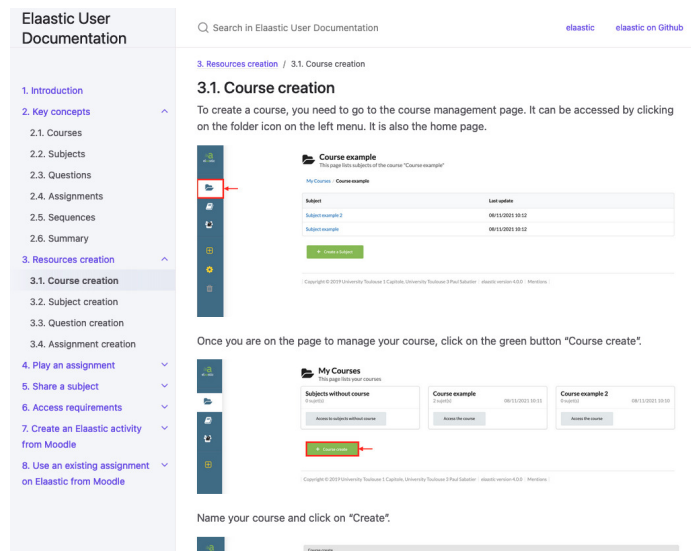


Figure 7.16: Online documentation.

Chapter 8

Evaluation

Contents

8.1	Expected Values of the Orchestration Model . . .	94
8.2	Data Collection	96
8.2.1	Description of the Dataset	96
8.2.2	Description of the Sample	97
8.3	Data Analysis	98
8.3.1	Teachers and Recommendations	98
8.3.2	Comparing s_{2022} with expected values	100
8.3.3	Comparing s_model_{2022} with expected values . . .	101
8.4	Verifying the Findings from the First Iteration	103
8.4.1	Proportion of Correct Answers during the First Vote	103
8.4.2	Learners Confidence Degree and Proportion of Correct Answers during the First Vote	104
8.4.3	Peer Grading Phase and Sequence Benefits.	105
8.4.4	Peer Gradings Phase and Learner’s Confidence Degree during the First Vote	105
8.5	Discussion	108

Back to our research questions (see Section 4.1), we provided answers to RQ1 by identifying meaningful information that can be inferred from the analysis of data gathered from Elaastic. We also proposed answers to RQ2 by designing orchestration practices based on this meaningful information. After having implemented our recommendation for teachers orchestration within Elaastic, we collected data from teachers usage of Elaastic 5.0 and analyse them. The next step of our research is to evaluate the effectiveness

of the orchestration recommendations regarding the benefits of a formative assessment sequence. Therefore, Section 8.1 computes the expected values of our orchestration model. More precisely, based on the sample we analysed in Chapter 6 that we named s_{2019} , we want to compute the average benefits of our model (i.e. the mean effect sizes of the sequences) as well as the percentage of beneficial sequences that our model leads to. These expected values will be our reference point for our analysis of the data collected from Elaastic 5.0 and described in Section 8.2. We analyse these data in Section 8.3 and 8.4. Finally, Section 8.5 discusses the results.

8.1 Expected Values of the Orchestration Model

Since our orchestration model was designed based on findings that where meant to improve sequences benefits, we want to focus on the effect of such model on our already existing sample that we analysed in Chapter 6. We will compute the distribution of outcomes (percentage of beneficial, not beneficial and harmful sequences) as well as the average benefits (effect size d). We name such outcomes and benefits the *expected values*.

In order to calculate the expected values of our orchestration model, we analysed some sequences from s_{2019} . More precisely, we analysed the subgroup of s_{2019} that contains all the sequences of s_{2019} that are in accordance with the model. Sequences in accordance with the model are sequences where the model would not have skipped the phase 2 according to the relevant indicators. We name the resulting subgroup of 71 sequences s_model_{2019} . Figure 8.1 shows the distribution of the outcome of sequences for s_{2019} and s_model_{2019} . Based on this figure, we can see that s_model_{2019} has better outcomes (i.e. more beneficial sequences and less harmful and non beneficial sequences). Figure 8.2 compares both sample through the kernel density of the effect size d . This figure shows that s_model_{2019} has overall better benefits than s_{2019} as expected. To verify if this difference is significant, we need to run a statistical test. Since s_model_{2019} is a subgroup of s_{2019} , the adequate statistical test is not the comparison of the mean effect sizes of s_{2019} and s_model_{2019} . Instead, we need to compare the mean of s_model_{2019} and the sequences of s_{2019} that are not in s_model_{2019} . This is because the test of whether a subgroup differs from the whole group is identical to the test of whether the subgroup differs from the remainder of the group [39]. A wilcoxon test that compared the mean effect sizes of both groups returned a p-value equal to $5.55e - 5$ and a 95% confidence interval equal to $[0.15 : 0.44]$.

Our results suggest that the model improves sequences outcomes and benefits. More precisely, our expected values are 76% of beneficial sequences

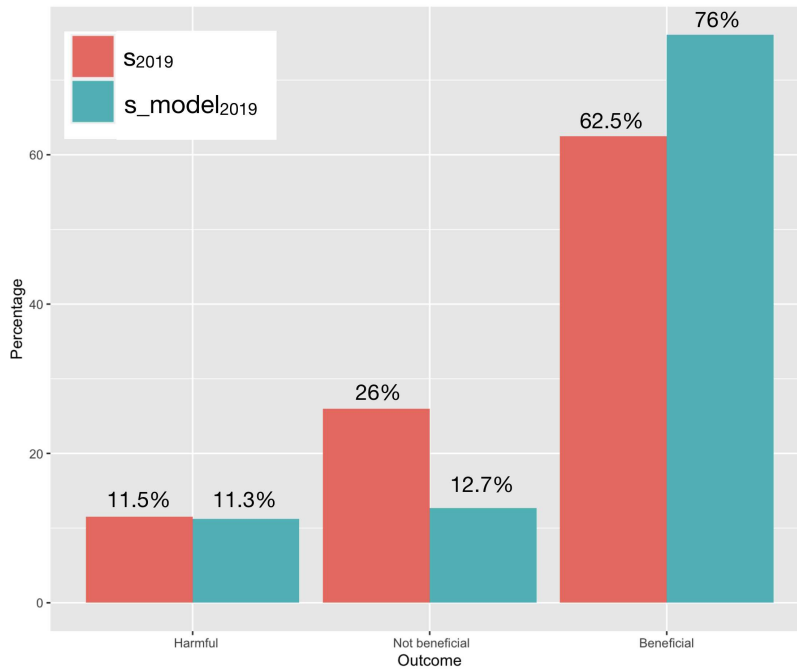


Figure 8.1: Outcome of sequences for s_{2019} and $s_{model2019}$.

and a mean effect size d equal to 0.35.

We can analyse the data collected from Elastic 5.0 to see how the authentic usage of Elastic 5.0 is close to these expected values. However, such a usage and results heavily depends on teachers' willingness to follow the recommendations or not. Therefore, we propose the following hypothesis:

"H.1: Teachers follow the explainable recommendations that are based on the orchestration model."

Furthermore, based on the expected benefits that we computed, we analyse the sequences that we obtained from teachers usage of our orchestration model and compute the *actual value*. By "actual value", we mean to designate the distribution of outcomes of these newly obtained sequences as well as their benefits. Consequently, we want to verify this second hypothesis:

"H.2: Actual values of sequences where teachers are provided with recommendations based on the orchestration model are not significantly lower than the expected values."

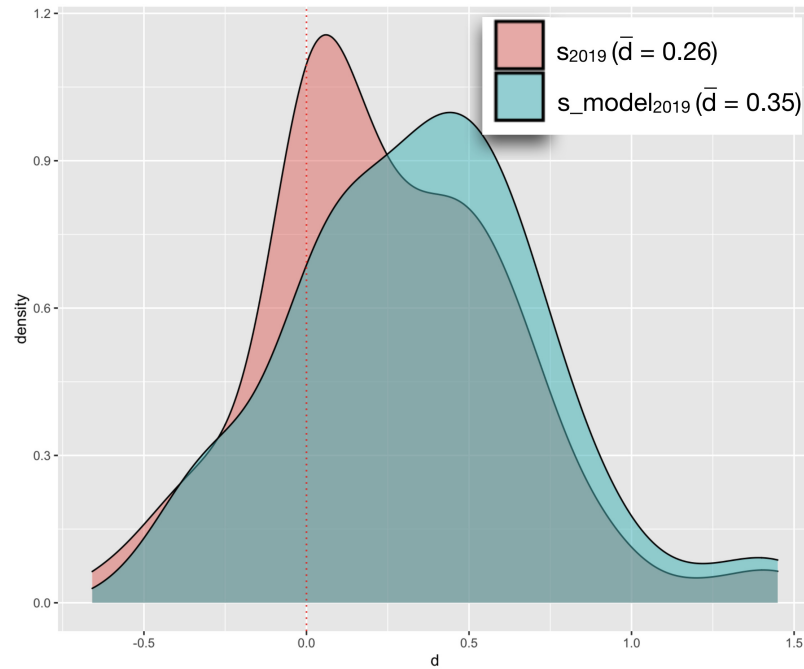


Figure 8.2: Kernel densities of s_{2019} and $s_{model_{2019}}$.

8.2 Data Collection

This section describes the data we collected since we implemented the new orchestration model into Elastic. We describe our dataset in subsection 8.2.1 and our selected sample in subsection 8.2.2.

8.2.1 Description of the Dataset

The dataset has the same properties as the one described in subsection 6.1.1. However, it is extended with additional event logs. Figure 8.3 is a simplified UML class diagram that provides a summary of this new dataset. In this figure, the red frame represents the new elements we introduced. Since Elastic 5.0 was deployed, we managed to collect 212 sequences conducted by 19 teachers, where 584 learners provided 2,629 answers and performed 23,637 peer ratings mainly in secondary education.

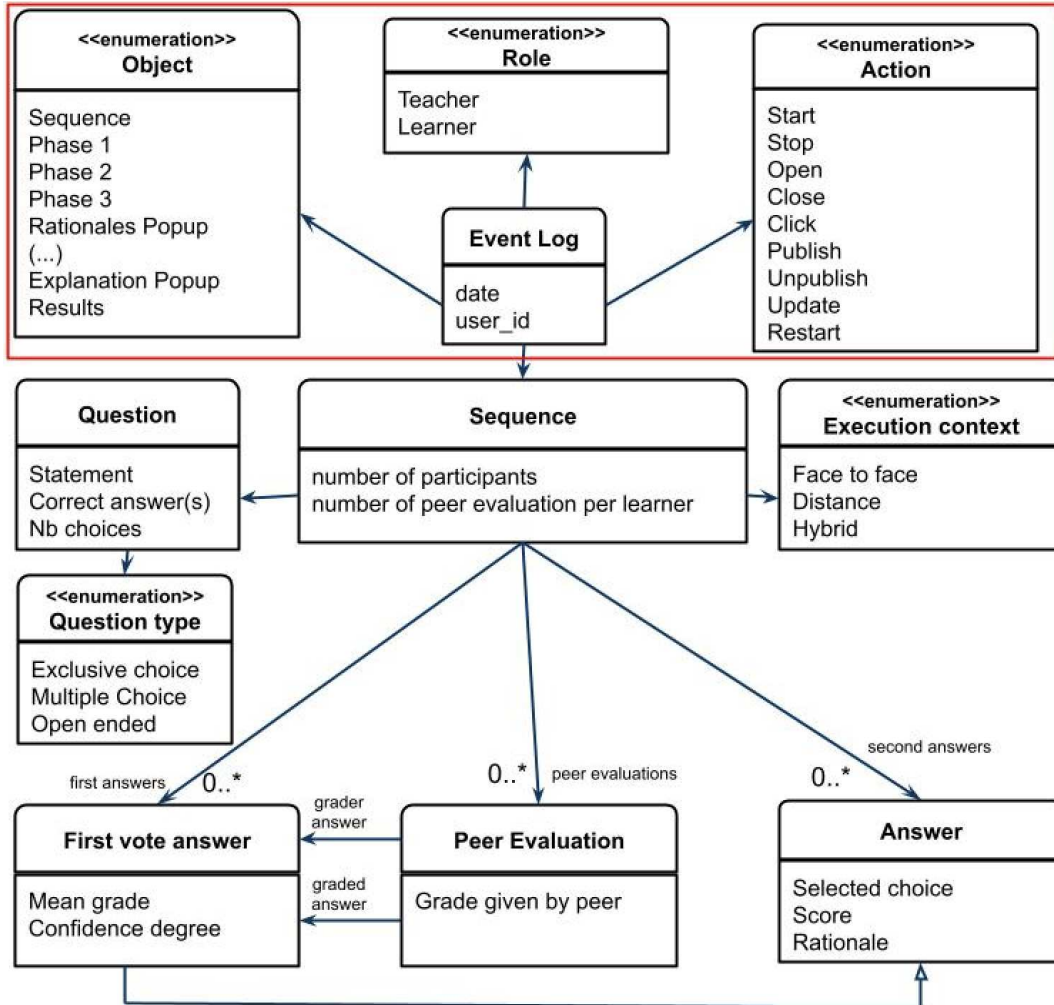


Figure 8.3: Summary of the dataset of Elastic 5.0 .

8.2.2 Description of the Sample

In order to conduct our evaluation analysis, the whole dataset has once again been filtered in order to reduce influential external factors and outliers according to the same criteria as the empirical study mentioned in Part III (see Section 6.1.2). Figure 8.4 summarises the filter criteria as well as the amount of sequences that resulted from each one. We ended up with 118 sequences conducted by 5 teachers, where 436 learners provided 1,988 answers and performed 6,036 peer ratings. In 4 of these sequences, the confrontation step and second vote step were skipped. Therefore, some indicators can not be computed for them such as the effect size d and ρ_{peer} . Table 8.1 summarises the topics of the 118 sequences of the sample, named s_{2022} . Let us

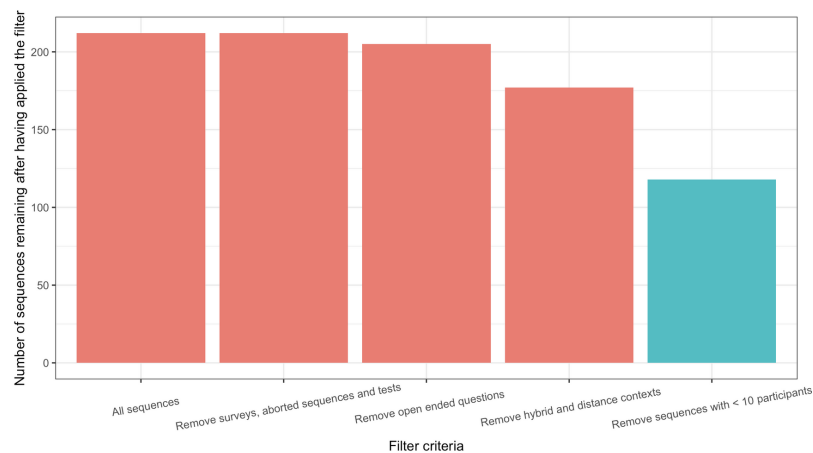


Figure 8.4: Summary of the filtering criteria applied to our dataset. The blue bar represents the 118 sequences of the final sample for our data analysis.

Topic	Number of sequences
Chemistry	68
Sociology	33
Biology	13
History	2
Mathematics	2

Table 8.1: Summary of our 118 sequences per topic.

also note that the data collected are from usages of Elaastic 5.0 in secondary education whereas our model was designed based on sequences collected in higher education. Such a difference is taken into account when discussion the results.

Figure 8.5 summarises the effect size of all the sequences of s_{2022} . The sample contains 72 beneficial sequences, 28 non beneficial sequences, 14 harmful sequences and 4 sequences where phase 2 was skipped.

8.3 Data Analysis

8.3.1 Teachers and Recommendations

In this section, we want to verify the hypothesis H.1 (i.e., teachers follow the recommendations for orchestration). We focus on recommendations provided to teachers at the end of phase 1 (since the other ones have no influences on sequences benefits), which include:

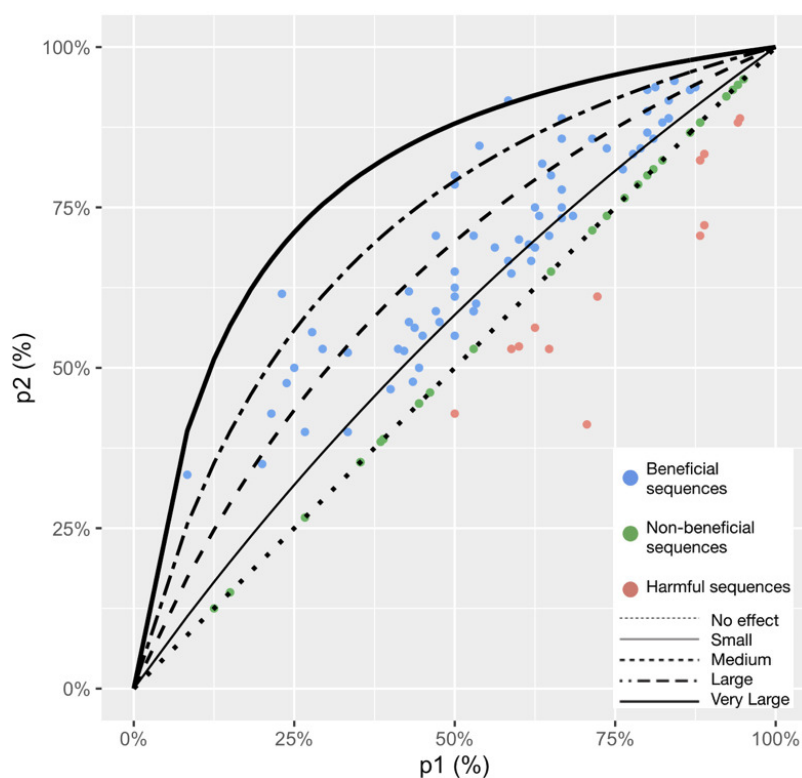


Figure 8.5: Our new sample compared to isovalues of the Cohen's d effect size: no effect ($d = 0$), small ($d = 0.2$), medium ($d = 0.5$), large ($d = 0.8$), and very large ($d = 1.2$). Each point is a sequence from s_{2022} .

- Teachers are recommended to provide hints to learners before proceeding with phase 2.
- Teachers are recommended to skip phase 2 and directly proceed with the discussion phase.
- Teachers are presented with no recommendations, which implies that they can proceed with phase 2.

Based on the tracking feature we implemented, we detected whether teachers followed our recommendations at the end of phase 1. We can not discuss here hints delivered by teachers to learners, because those oral interactions are not captured by Elaastic 5.0. Let us note that, at the time of implementation, we chose $[30\% : 70\%]$ as our ideal interval for $p1$ instead of $[30\% : 80\%]$. However, we made sure that such difference with the interval did not impact the results presented in the remaining of this chapter. Teachers were thus recommended to skip phase 2 when:

Recommendation	Teachers decision	
	Skip phase 2	Start phase 2
Skip phase 2	2	47
Start phase 2	2	67

Table 8.2: Contingency table of recommendations regarding phase 2, compared to teachers' final decision.

- $p1 > 0.7$.
- $p1 < 0.3$ and $\rho_{conf} < 0$.

Table 8.2 summarises teachers' decision depending on whether they were recommended to skip phase 2 or not. Teachers act in accordance with the recommendation 69 times out of 118 sequences ($\approx 58\%$). Out of the 49 sequences where the recommendation to skip phase 2 appeared, they followed it 2 times ($\approx 4\%$). When they were not provided with any recommendation (which implies that they can start phase 2), they act in accordance with the model 67 times out of 69 ($\approx 97\%$). We also ran a χ^2 test to verify whether there is a relationship between teachers' decision and the recommendation or not. We obtained a p-value equal to 1. This result strongly suggests that there is no relationships between recommendations and teachers' decisions.

As a conclusion, recommendations provided to teachers at the end of phase 1 do not influence their final decision. After an informal discussion with some teachers, it would appear that most of them actually did not see the recommendations. This is reinforced by the fact that, according to the tracking of teachers actions, none of them clicked on the "Know more..." button next to these recommendations.

8.3.2 Comparing s_{2022} with expected values

This section aims to verify hypothesis H.2. In this analysis we removed sequences where phase 2 was skipped because the outcomes and benefits can not be computed without phase 2 (because there is no second vote to compute d). In order to compare the expected values with the actual values, we compared expected values based on s_{2019} with s_{2022} . Figure 8.6 shows the distribution of the outcome of s_{2022} compared to the expected values. Figure 8.7 compares both sample through the kernel density of the effect size d . Both figures suggest that the expected values are better than the actual ones. A mean comparison of d through a Wilcoxon test return a p-value equal to 0.02 and a 95% confidence interval equal to $[0 : 0.25]$.

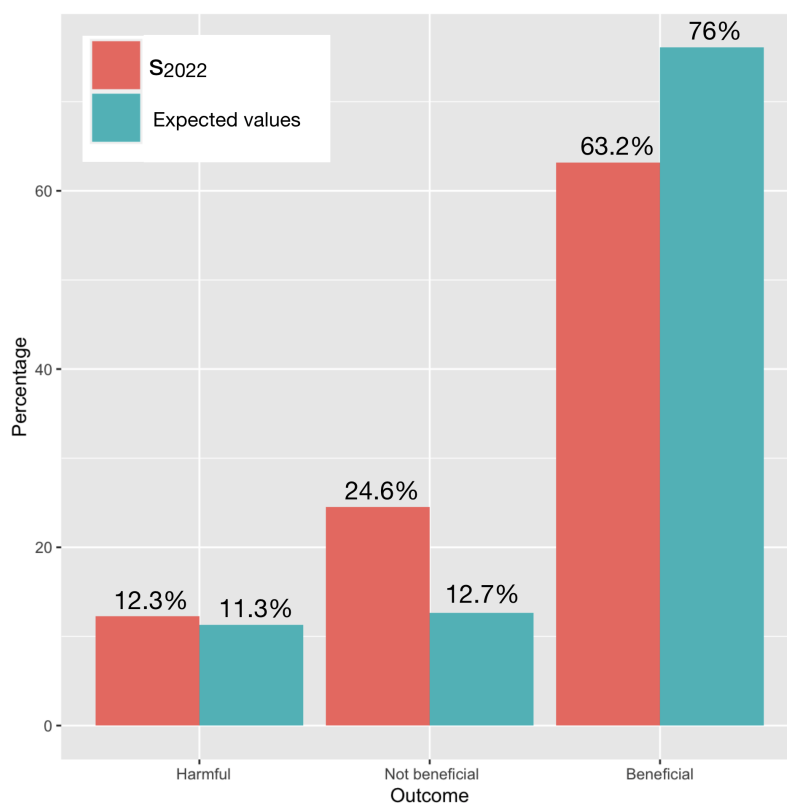


Figure 8.6: Outcome of sequences from s_{2022} compared to expected values.

As a conclusion, our new implementation of Elaastic have significantly lower values than expected. This result is not surprising, as teachers did not follow the recommendations (see Section 8.3.1). Another analysis that does not account for teachers decision to follow the recommendations need to be conducted. Consequently, in the next section, we explore the sequences of our sample that were delivered according to the model.

8.3.3 Comparing s_model_{2022} with expected values

According to Table 8.2, 69 sequences of the whole sample were delivered in accordance with the decision recommended by the model. However, for 2 of these 69 sequences, phase 2 was skipped and their benefits could not be computed. We name the resulting sample of 67 sequences s_model_{2022} .

Figure 8.8 shows the distribution of the outcome of sequences for s_model_{2022} compared to expected values. Based on this figure, we can see that s_model_{2022} is close to the expected values. More precisely expected values have a similar proportion of beneficial sequences. However, when it comes to harmful and

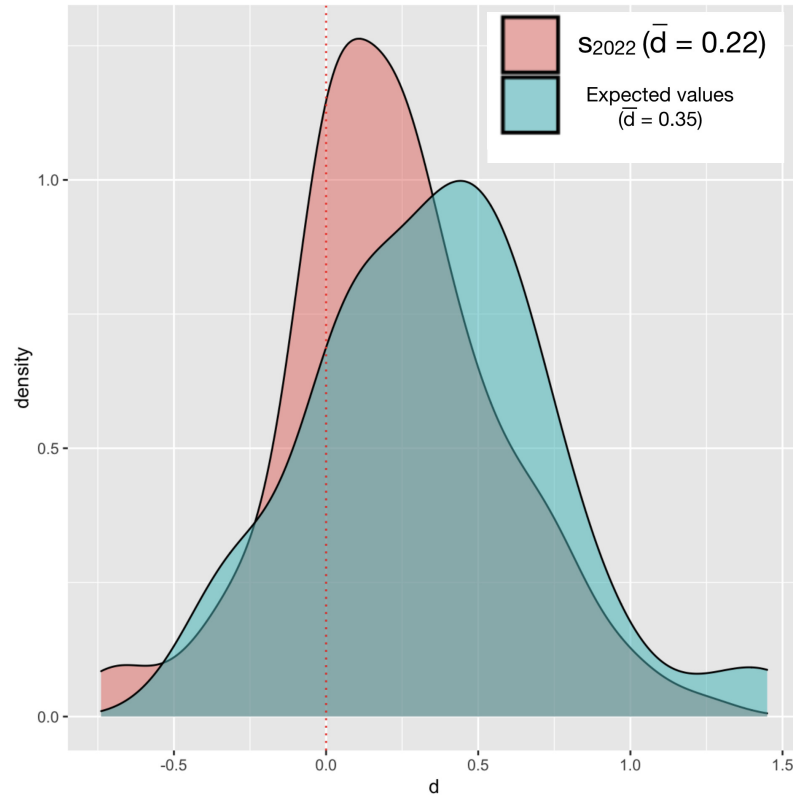


Figure 8.7: Kernel densities of the effect size d of s_{2022} compared to expected values.

non beneficial sequences, the similarity is not as obvious. Figure 8.9 compares both sample through the kernel density of the benefits. This figure shows that expected values has overall better benefits than s_model_{2022} . A wilcoxon test comparing the means benefits of s_model_{2019} and s_model_{2022} returned a p-value equal to 0.11 and a 95% confidence interval equal to $[0 : 0.21]$ which is not significant enough. In other words, s_model_{2019} and s_model_{2022} have non significantly different results.

In conclusion, sequences where teachers act in accordance with the recommendations, have overall better results than all the sequences. This means that, if teachers had followed all the recommendations, they would have significantly more beneficial sequences as well as higher average benefits. More precisely, the actual results that they would have obtained would be close to the theoretical expected results that we obtained when designing the model.

The next section focuses on cross validating the finding from the first iteration.

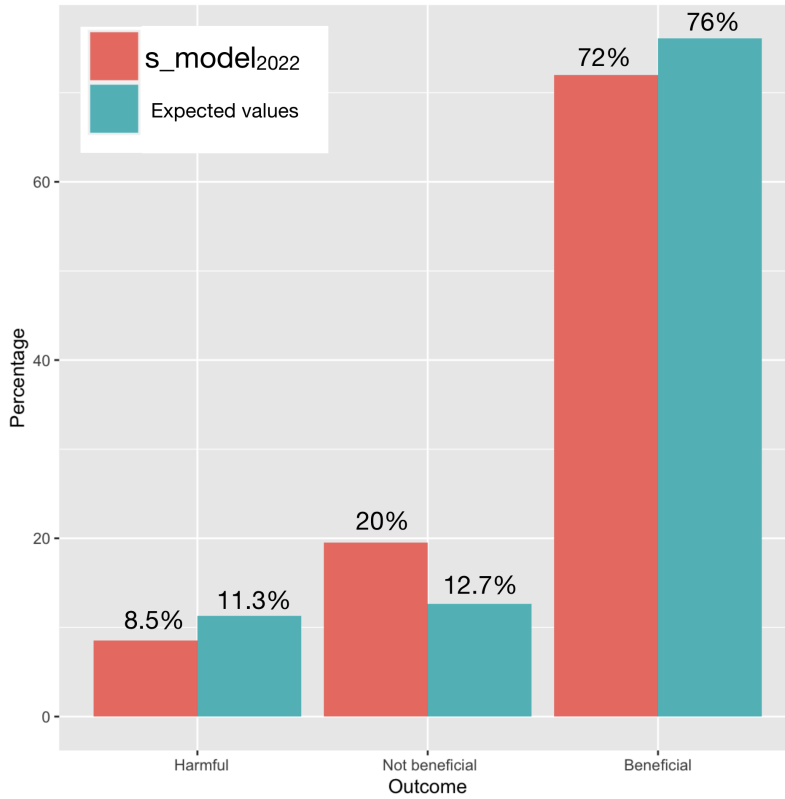


Figure 8.8: Outcome of sequences from s_model_{2022} compared to expected values.

8.4 Verifying the Findings from the First Iteration

In this section, we want to take advantage of this new sample to verify whether the findings from which our orchestration model originated are also verified with this new sample. The exact findings that we want to perform are the ones from subsections 6.2.2, 6.2.3, 6.2.4 and 6.2.5 since they are the ones that led to recommendations for orchestration designed to have an influence on sequences benefits and outcomes.

8.4.1 Proportion of Correct Answers during the First Vote

The first finding we want to verify is the one mentioned in Section 6.2.2, namely, the correlation between d and $|p1 - 0.5|$. Results appear in Fig-

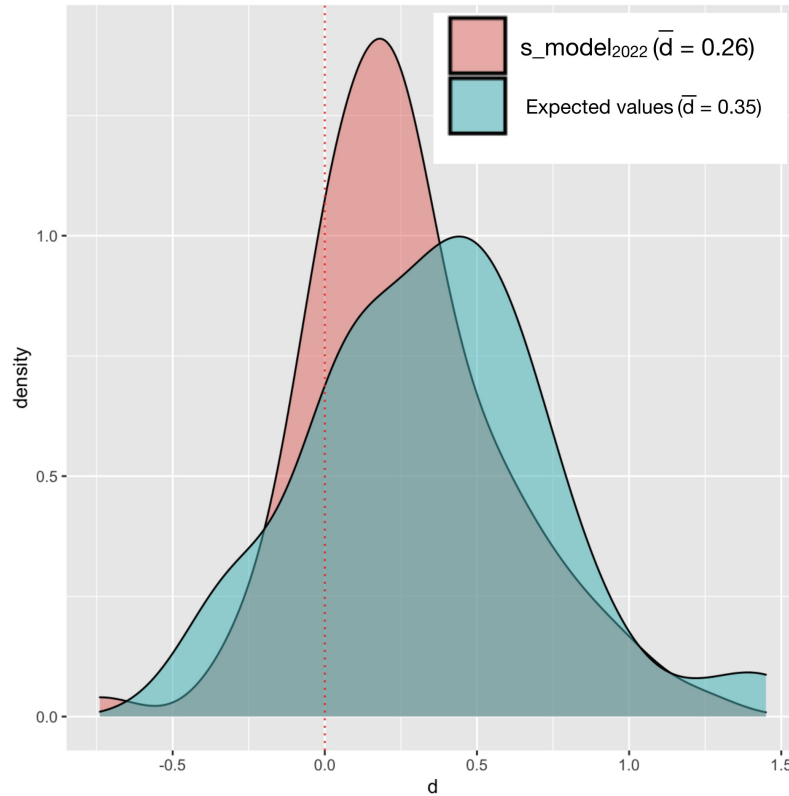


Figure 8.9: Kernel densities of the effect size d of s_model_{2022} compared to expected values.

Figure 8.10, and we computed the Spearman correlation between $|p1 - 0.5|$ and d . It is equal to -0.17 with a p -value = 0.07 and a 95% confidence interval equal to $[-0.35;-0.01]$, which suggests that the hypothesis is verified but the correlation is weaker and not significant.

8.4.2 Learners Confidence Degree and Proportion of Correct Answers during the First Vote

The next finding is the one from subsection 6.2.3 which suggested leveraging ρ_{conf} and $p1$ in order to design recommendations for teachers. Figure 8.11 shows the benefits of sequences depending on the proportion of correct answers at the end of first vote and the consistency of confidence degree. Based on the mean effect sizes observed on this figure, ρ_{conf} seems to play a significant role when $p1 < 30\%$. The effect size is normally distributed (as computed by the Shapiro-Wilk test which returned a p -value equal to 0.15) and both groups have homogeneous variances (F test returned a p -value equal

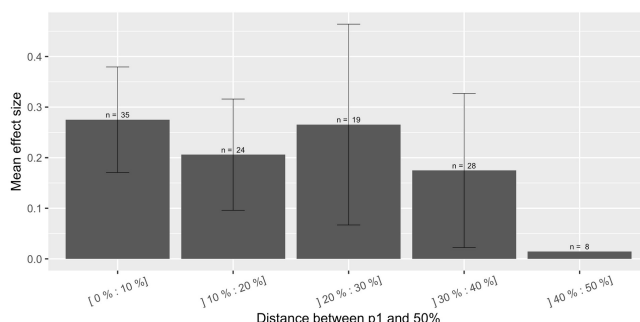


Figure 8.10: The effect size d of the sequences from s_{2022} depending on the distance between $p1$ and 50% with 95% confidence intervals. The last bar has too few sequences to compute a meaningful confidence interval.

to 0.91). We can therefore use a t-test to compare the mean benefits of sequences grouped by consistency of confidence degree ($\rho_{conf} < 0$ or not) when $p1 < 30\%$, which returned a significant difference (p-value = $7e - 3$ and 95% confidence interval = $[0.1 : 0.6]$). This finding thus applies to our sample as well, and is even more significant than it was for s_{2019} .

8.4.3 Peer Grading Phase and Sequence Benefits.

In Section 6.2.4, we explored the relation between ρ_{peer} and d . Figure 8.12 shows a plot diagram of the same analysis applied to the new sample. The Spearman correlation between ρ_{peer} and d is 0.15 with a p-value = 0.12 and a 95% confidence interval equal to $[-0.04 : 0.32]$, which is inconclusive for this finding.

8.4.4 Peer Gradings Phase and Learner's Confidence Degree during the First Vote

Finally, we want to verify the finding from Section 6.2.5. We identified a correlation between the consistency of confidence degree (ρ_{conf}) and the consistency of peer gradings (ρ_{peer}).

Figure 8.13 is a plot diagram of ρ_{peer} according to ρ_{conf} for s_{2022} . The Spearman correlation between ρ_{conf} and ρ_{peer} is 0.33 (p-value = $3e - 4$ and 95% confidence interval = $[0.16 : 0.49]$), which supports our finding.

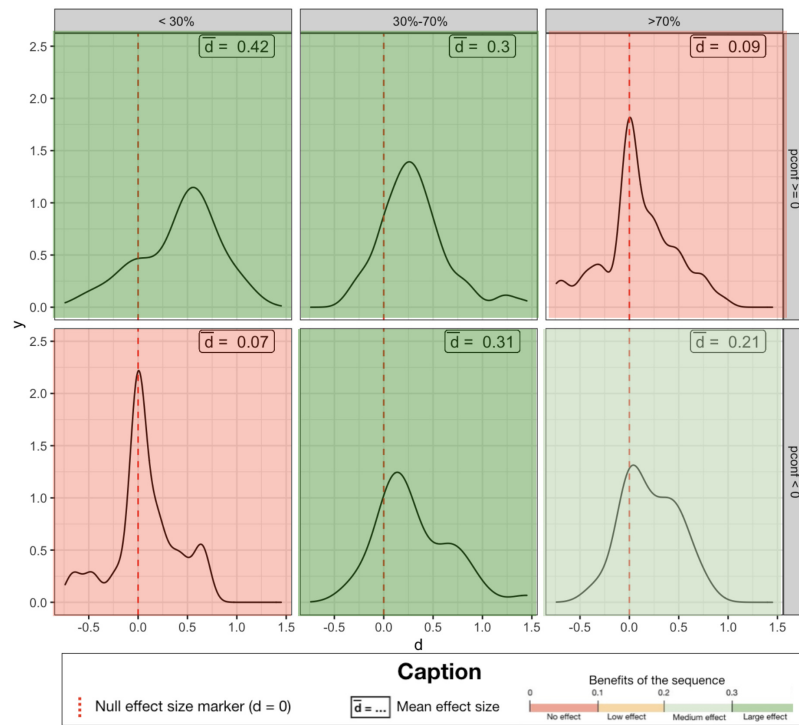


Figure 8.11: Kernel density estimates and mean of the effect size (d) of our sequences depending on the confidence consistency (ρ_{conf}) and the proportion of correct answers at the first vote (p_1).

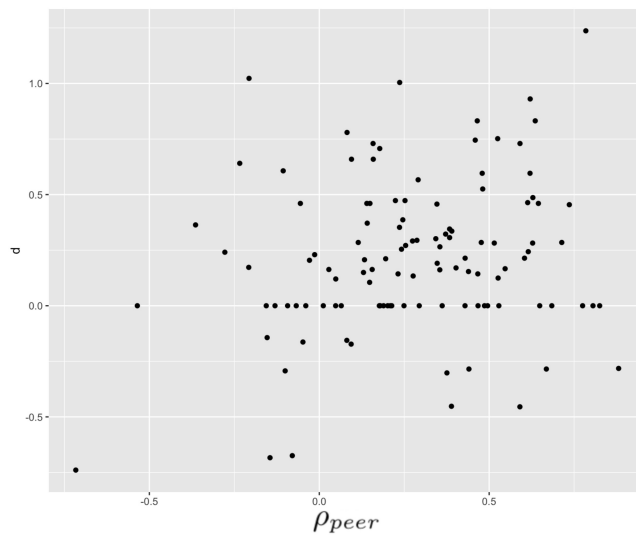


Figure 8.12: Effect size d depending on the consistency of peer gradings ρ_{peer} . Each point is a sequence from s_{2022} .

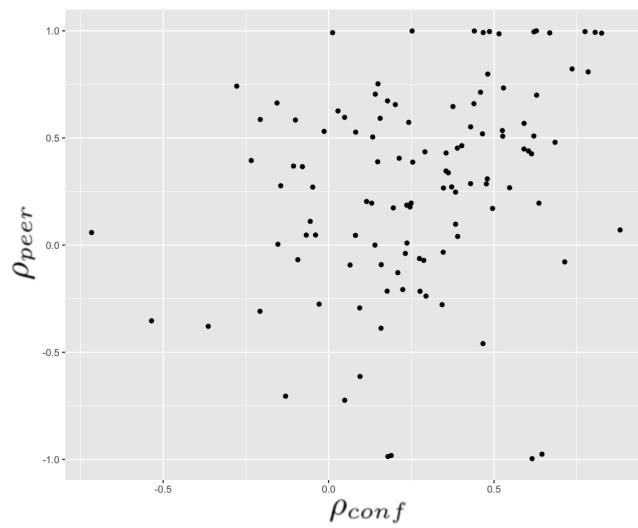


Figure 8.13: Peer grading consistency ρ_{peer} depending on the confidence consistency ρ_{conf} for s_{2022} . Each point is a sequence from s_{2022} .

8.5 Discussion

In this chapter, we analysed the data obtained from the usage of Elaastic 5.0 and obtained results that suggest that this new version of Elaastic does not improve sequences outcomes and benefits. However, we provided evidence that shows that this result can be explained by teachers not seeing (and therefore not following) the recommendation that are provided to them. This suggests that our implementation of explainable recommendations is not effective. An informal discussion with some teachers suggests that orchestrating a two-votes-based sequences with Elaastic is a very heavy task on a cognitive level. This would explain why teachers from our sample which, in addition to that, were mainly newcomers to Elaastic, did not notice the recommendations. Based on cognitive load theory [144], we believe that future works should focus on a version of Elaastic that would prevent teachers from being overloaded with information.

Consequently, we performed another analysis without accounting for sequences where teachers were not in accordance with the recommendation. This analysis provided evidences that, if teachers had followed the recommendations, the outcomes and benefits of sequences would have been close to their theoretical expected values. As a consequence, our orchestration model can actually improves benefits for learners compared to the basic version of two-votes-based process. More precisely, as contribution of this thesis, we provided evidences that show that, when teachers make decision in accordance with our orchestration model (i) sequences benefits are significantly increased (ii) the outcomes of sequences in term of benefits is improved.

In addition to that we performed analysis that were intended to verify the robustness of our orchestration model by verifying our findings from Chapter 6. This was done because the two samples had different characteristics. The sample from our first empirical study was from higher education and addressed mainly computer science related topics whereas the sample we collected for the evaluation of our orchestration model was from secondary education and had a more balanced representation of its topics and addressed notably chemistry and biology. With this analysis, we provided evidences that show that our contributions from the first iteration fits secondary education as well as evidences that show that our contributions from the first iteration fits other topics. However, we could not draw any conclusion regarding a finding based on our sample collected for evaluation of our orchestration model. More precisely, the correlation between consistency of peer grading and benefits of a sequence could not be verified in the latest sample. We believe that this might be due to the difference of level since a previous study suggests that the impact of peer assessment on student

performance vary from a level education to another [51]. Another empirical review of very few studies led to preliminary results that suggest that anonymous peer review activities lead to better performance in higher education than in the secondary education [113]. However, these studies do not provide strong enough evidences. More generally, since most of the study on peer assessment occur in higher education, there is a gap of knowledge regarding such effects in secondary education [152].

Chapter 9

Conclusion

9.1 Summary

We introduced formative assessment and demonstrated its effectiveness for improving learning results based on literature. Then, we emphasized feedback's importance within formative assessment practices and stated that conducting such practices is a challenging task for teachers. As a consequence, we addressed the following research questions:

- **RQ1 - Which useful information can be inferred from the analysis of data gathered from a tool used in authentic contexts?**
- **RQ2 - How can such information contribute to improve formative assessment processes orchestration?**

To answer these questions, we proposed to conduct learning analytics related research. As a consequence, we collected data issued from a formative assessment tool that proposed an implementation of a family of processes we identified and called the two-votes-based processes. With these data, we conducted analysis that resulted in findings about formative assessment. Such findings are contributions that address RQ1. Regarding RQ2, based on these findings, we managed to propose recommendations for engineers designing formative assessment systems (particularly two-votes-based formative assessment systems). We also designed recommendations for orchestration intended to help teachers in their practice of formative assessment and we compiled them to propose an orchestration model. We put this model to the test by implementing it and collecting data again. Even though the analysis of these data provided encouraging results, it provided evidences that some improvement can still be made to make the formative assessment system

significantly more beneficial to learners. More precisely, formative assessment system implementing our recommendations are required to improve the human-computer interaction related feature regarding recommendations so that teachers are more likely to follow them.

9.2 Future Works

9.2.1 Short Term: New Version of Elaastic

As short terms future works, we would like to conduct analysis with the data collected after we changed our ideal interval for recommendations from [30% : 70%] to [30% : 80%]. Indeed, the ideal interval that we identified when implementing Elaastic 5.0 was [30% : 70%] but new evidences led us to believe that [30% : 80%] was a better suited interval.

Even though our model returned satisfying results regarding the effect sizes (e.g. the benefits) of the sequences in general, it would appear that it mainly reduces the occurrence of non beneficial sequences instead of reducing the occurrence of harmful sequences as well (see Figure 8.1). This might suggest that our model encourages learners to change their initial answer much more than it prevents them from changing from wrong to right. A previous study from Smith analyses learners response switching in Peer instruction context and suggests that is correlated with their self-efficacy and the difficulty of the question [104]. These two statements can be at the core of a future analysis.

In addition to that, the next intervention to measure learners benefits should focus on improving Human-Computer Interaction related issues regarding our recommendations for orchestration. As demonstrated in Section 8.3.1, it would appear that teachers do not see the recommendations that are provided to them, which prevent them from maximizing the benefits of their sequences. As a consequence, we want to improve the way recommendations are presented to teachers by adding more visual artefacts that would support such recommendations. Figure 9.1 is a mockup of our proposition for a possible next iteration of data collection and analysis. As shown in this figure, two additions are made to recommendations. The first one is the addition of a little light bulb icon that is expected to catch teachers' eyes. The second addition is a disrecommendation. More precisely, we added "*(not recommended)*" on the label of the button on which we do not recommend teachers to click. This is based on Zhang's study [162] which proposed to present disrecommendations by telling the users what not to do beyond simply telling them what to do.

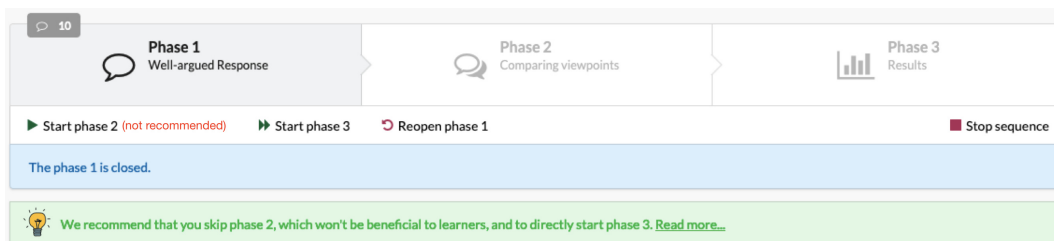


Figure 9.1: New design of recommendations for teachers.

9.2.2 Mid-term: Expansion of the new process

Deeper studies have also to be performed to adapt our new orchestration model to various execution contexts. Indeed, as explained earlier (see Section 5.1), Elaastic has been designed for remote usage as well, whether in synchronous mode such as live virtual classroom, or in asynchronous mode such as homeworks. This means that some indicators (such as the consistency of confidence degree and the proportion of correct answers at the end of the first vote) can not be interpreted the same way. We began to conduct some analysis to propose alternatives to ρ_{conf} and ρ_{peer} that fit asynchronous contexts. Such analysis can be found in Appendix A.3 and led to encouraging results. We believe that an asynchronous-adapted model should not focus on the current state of the sequence depending on indicators based on an overview of all learners' performance, but instead on a contextual version of these indicators that focuses on each learner individually.

9.2.3 Long term: Additional exploratory studies

Finally, long term future works may explore more data. In this thesis, we focused on result across a sequence in order to provide teachers with feedback. However, other studies could focus on learners by analysing their performance and behaviour along many sequences and providing them with feedback designed to improve self regulation [35]. We also began some prior works to focus on questions and their statement. More precisely, we wanted to detect which question are most reused by teachers and therefore perceived positively by them [6].

The other future work that we plan to do is the one regarding relevancy of rationales. Rationales are a crucial part of the process and we currently provide no indicator that is related to it. Prior works proposed an indicator to measure the relevance of rationales in Peer Instruction context [59] by averaging the cosine similarity of each rationale with all the others ones. We conducted some analysis to validate such indicator and obtained encouraging

results that would require teachers validation. We detail such an analysis in Appendix B. With this new indicator, we might be able to improve the algorithm that determines the rationales attributed to peer for evaluation during the peer grading phase. We may also be able to propose a semi-automatic classification of learners rationales. More precisely, identifying irrelevant rationales would improve the process by (i) improving the peer grading phase by preventing irrelevant rationales from being graded (ii) preventing teachers from focusing on irrelevant rationales during the debriefing phase (iii) prompting teachers to intervene by identifying the author of an irrelevant rationale.

Bibliography

Bibliography

- [1] Aakerlind, G. Academic growth and development-How do university academics experience it?. *Higher Education*. **50**, 1-32 (2005)
- [2] Adelman, N., Donnelly, M., Dove, T., Tiffany-Morales, J., Wayne, A. & Zucker, A. The integrated studies of educational technology: Professional development and teachers' use of technology. *Arlington, VA: SRI International*. (2002)
- [3] Alamri, R. & Alharbi, B. Explainable student performance prediction models: a systematic review. *IEEE Access*. (2021)
- [4] Anderson, T. & Shattuck, J. Design-based research: A decade of progress in education research?. *Educational Researcher*. **41**, 16-25 (2012)
- [5] 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)
- [6] Andriamiseza, R. Évaluer la réutilisabilité d'une question: une utilisation des learning analytics dans un contexte d'évaluation formative. *8e Rencontres Jeunes Chercheurs En EIAH 2021*. (2020)
- [7] Andriamiseza, R., Silvestre, F., Parmentier, J. & Broisin, J. Data-informed Decision-making in TEFA Processes: An Empirical Study of a Process Derived from Peer-Instruction. *Proceedings Of The Eighth ACM Conference On Learning@ Scale*. pp. 259-262 (2021)
- [8] Andriamiseza, R., Silvestre, F., Parmentier, J. & Broisin, J. Recommendations for Orchestration of Formative Assessment Sequences: a Data-driven Approach. *European Conference On Technology Enhanced Learning*. pp. 245-259 (2021)
- [9] Andriamiseza, R., Silvestre, F., Parmentier, J. & Broisin, J. Vers la conception de feedback pour enseignants dans un contexte d'évaluation formative à grande échelle: une approche analytique. *10e Conférence Sur Les Environnements Informatiques Pour L'Apprentissage Humain*. pp. 46-57 (2021)
- [10] Araceli Ruiz-Primo, M. & Furtak, E. Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment*. **11**, 237-263 (2006)

- [11] Araujo, T., Helberger, N., Kruikemeier, S. & De Vreese, C. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*. **35**, 611-623 (2020)
- [12] Awedh, M., Mueen, A., Zafar, B. & Manzoor, U. Using Socratic and Smartphones for the support of collaborative learning. *ArXiv Preprint ArXiv:1501.01276*. (2015)
- [13] Baird, J., Hopfenbeck, T., Newton, P., Stobart, G. & Steen-Utheim, A. Assessment and learning: State of the field review. *Knowledge Center For Education, Oslo*. (2014)
- [14] Bakharia, A., Kitto, K., Pardo, A., Gašević, D. & Dawson, S. Recipe for success: lessons learnt from using xAPI within the connected learning analytics toolkit. *Proceedings Of The Sixth International Conference On Learning Analytics & Knowledge*. pp. 378-382 (2016)
- [15] 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)
- [16] Bell, N., Perret-Clermont, A. & Grossen, M. Sociocognitive conflict and intellectual growth. *Peer Conflict And Psychological Growth. New Directions For Child Development*. **29** pp. 41-54 (1985)
- [17] Bell, P. & Linn, M. Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal Of Science Education*. **22**, 797-817 (2000)
- [18] 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)
- [19] Black, P., Harrison, C. & Lee, C. Assessment for learning: Putting it into practice. (McGraw-Hill Education (UK),2003)
- [20] Black, P., Wiliam, D.: Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice* **5**(1), 7–74 (Mar 1998)
- [21] 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)

- [22] Bradley, R. & Terry, M. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*. **39**, 324-345 (1952)
- [23] Brink, M. & Bartz, D. Effective use of formative assessment by high school teachers. *Practical Assessment, Research, And Evaluation*. **22**, 8 (2017)
- [24] Brooks, B., Gilbuena, D., Krause, S. & Koretsky, M. Using word clouds for fast, formative assessment of students' short written responses. *Chemical Engineering Education*. **48**, 190-198 (2014)
- [25] Brooks, Bill J and Koretsky, Milo D: The Influence of Group Discussion on Students' Responses and Confidence during Peer Instruction. *Science* **323**(5910), 122–124 (2009)
- [26] Calderon, A. Massification of higher education revisited. *Melbourne: RMIT University*. (2018)
- [27] Caldwell, J. Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sciences Education*. **6**, 9-20 (2007)
- [28] Cambre, J., Klemmer, S. & Kulkarni, C. Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. *Proceedings Of The 2018 CHI Conference On Human Factors In Computing Systems*. pp. 1-13 (2018)
- [29] 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)
- [30] Chen, X., Bennett, P., Collins-Thompson, K. & Horvitz, E. Pairwise ranking aggregation in a crowdsourced setting. *Proceedings Of The Sixth ACM International Conference On Web Search And Data Mining*. pp. 193-202 (2013)
- [31] Chen, G., Shah, D. & Others Explaining the success of nearest neighbor methods in prediction. (Now Publishers,2018)
- [32] Cheng, W., Shen, Y., Huang, L. & Zhu, Y. Incorporating interpretability into latent factor models via fast influence analysis. *Proceedings Of The 25th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. pp. 885-893 (2019)

- [33] Chi, M.T.H., Wylie, R.: The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist* **49**(4), 219–243 (Oct 2014)
- [34] Chowdhary, K. Natural language processing. *Fundamentals Of Artificial Intelligence*. pp. 603-649 (2020)
- [35] Clark, I. Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*. **24**, 205-249 (2012)
- [36] Clark, S.: Enhancing active learning: Assessment of poll everywhere in the classroom. Tech. rep., University of Manitoba (2017)
- [37] Elaastic: <https://github.com/elaastic/elaastic-questions-server>, last consulted: 2021-01-28.
- [38] Tsaap-Note: <https://github.com/TSaaP/tsaap-notes>, last consulted: 2021-01-28.
- [39] Source: https://www.ibm.com/my-support/s/question/0D50z00006PsEenCAF/comparing-subgroup-means-against-the-total-whole-mean?language=en_US, last consulted: 2022-09-19.
- [40] Crouch, C.H., Mazur, E.: Peer instruction: Ten years of experience and results. *American journal of physics* **69**(9), 970–977 (2001)
- [41] Clow, D. The learning analytics cycle: closing the loop effectively. *Proceedings Of The 2nd International Conference On Learning Analytics And Knowledge*. pp. 134-138 (2012)
- [42] Cuban, L., Kirkpatrick, H. & Peck, C. High access and low use of technologies in high school classrooms: Explaining an apparent paradox. *American Educational Research Journal*. **38**, 813-834 (2001)
- [43] Curtis, D.A., Lind, S.L., Boscardin, C.K., Dellenges, M.: Does student confidence on multiple-choice question assessments provide useful information? *Medical education* **47**(6), 578–584 (2013)
- [44] Daniel, B. Artificial reality: The practice of analytics and big data in educational research. *Big Data*. (2019)
- [45] Das, S., Alsalhanie, K., Nauhria, S., Joshi, V., Khan, S. & Surender, V. Impact of formative assessment on the outcome of summative assessment—a feedback based cross sectional study conducted among basic science medical students enrolled in MD program. *Asian Journal Of Medical Sciences*. **8**, 38-43 (2017)

- [46] Davis, M.: Technology fed growth in formative assessment. *Education Week* p. 11 (2015)
- [47] Dawson, S., Bakharia, A., Heathcote, E. & Others SNAPP: Realising the affordances of real-time SNA within networked learning environments. (Networked Learning,2010)
- [48] Dillenbourg, P. Design for classroom orchestration, position paper. *Design For Classroom Orchestration. Computers & Education.* (2012)
- [49] Dmoshinskaia, N., Gijlers, H., de Jong, T.: Learning from reviewing peers' concept maps in an inquiry context: Commenting or grading, which is better? *Studies in Educational Evaluation* **68**, 100959 (2021)
- [50] Doshi-Velez, F. & Kim, B. A roadmap for a rigorous science of interpretability. *ArXiv: Abs/1702.08608.* (2017)
- [51] 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)
- [52] Elias, T. Learning analytics. *Learning.* pp. 1-22 (2011)
- [53] 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)
- [54] Everett, B.: An introduction to latent variable models. Springer Science & Business Media (2013)
- [55] Ferguson, R. Learning analytics: drivers, developments and challenges. *International Journal Of Technology Enhanced Learning.* **4**, 304-317 (2012)
- [56] France, L., Heraud, J., Marty, J., Carron, T. & Heili, J. Monitoring virtual classroom: Visualization techniques to observe student activities in an e-learning system. *Sixth IEEE International Conference On Advanced Learning Technologies (ICALT'06).* pp. 716-720 (2006)
- [57] Frömer, R., Nassar, M., Bruckner, R., Stürmer, B., Sommer, W. & Yeung, N. Response-based outcome predictions and confidence regulate feedback processing and learning. *Elife.* **10** pp. e62825 (2021)

- [58] Furtak, E., Kiemer, K., Circi, R., Swanson, R., León, V., Morrison, D. & Heredia, S. Teachers' formative assessment abilities and their relationship to student learning: findings from a four-year intervention study. *Instructional Science*. **44**, 267-291 (2016)
- [59] Gagnon, V., Labrie, A., Bhatnagar, S., Desmarais, M.C.: Filtering non-relevant short answers in peer learning applications. In: EDM (2019)
- [60] Ghasemi, A., Zahediasl, S.: Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism* **10**(2), 486 (2012)
- [61] Gikandi, J., Morrow, D. & Davis, N. Online formative assessment in higher education: A review of the literature. *Computers & Education*. **57**, 2333-2351 (2011)
- [62] Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M. & Kagal, L. Explaining explanations: An approach to evaluating interpretability of ml. *ArXiv Preprint ArXiv:1806.00069*. (2018)
- [63] Gormally, C., Evans, M. & Brickman, P. Feedback about teaching in higher ed: Neglected opportunities to promote change. *CBE—Life Sciences Education*. **13**, 187-199 (2014)
- [64] Hattie, J. Measuring the effects of schooling. *Australian Journal Of Education*. **36**, 5-13 (1992)
- [65] Hattie, J. & Timperley, H. The power of feedback. *Review Of Educational Research*. **77**, 81-112 (2007)
- [66] Hattie, J.: Visible learning for teachers: Maximizing impact on learning. Routledge, 711 Third Avenue, New York, NY 10017 (2012)
- [67] Heritage, M. Formative Assessment and Next-Generation Assessment Systems: Are We Losing an Opportunity?.. *Council Of Chief State School Officers*. (2010)
- [68] Herlocker, J., Konstan, J. & Riedl, J. Explaining collaborative filtering recommendations. *Proceedings Of The 2000 ACM Conference On Computer Supported Cooperative Work*. pp. 241-250 (2000)
- [69] Herlocker, J. Understanding and improving automated collaborative filtering systems. (University of Minnesota,2000)

- [70] Hestenes, D., Wells, M. & Swackhamer, G. Force concept inventory. *The Physics Teacher*. **30**, 141-158 (1992)
- [71] Hoffman, R., Mueller, S., Klein, G. & Litman, J. Metrics for explainable AI: Challenges and prospects. *ArXiv Preprint ArXiv:1812.04608*. (2018)
- [72] Hoffman, R., Mueller, S., Klein, G. & Litman, J. Metrics for Explainable AI: Challenges and Prospects.
- [73] 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)
- [74] Jhangiani, R. The impact of participating in a peer assessment activity on subsequent academic performance. *Teaching Of Psychology*. **43**, 180-186 (2016)
- [75] Johnson, C., Hill, L., Lock, J., Altowairiki, N., Ostrowski, C., Santos, L. & Liu, Y. Using design-based research to develop meaningful online discussions in undergraduate field experience courses. *International Review Of Research In Open And Distributed Learning: IRRODL*. **18**, 36-53 (2017)
- [76] Jiang, Y. Exploring teacher questioning as a formative assessment strategy. *RELC Journal*. **45**, 287-304 (2014)
- [77] Joshi, A., Kale, S., Chandel, S. & Pal, D. Likert scale: Explored and explained. *British Journal Of Applied Science & Technology*. **7**, 396 (2015)
- [78] Kamata, A. & Bauer, D. A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*. **15**, 136-153 (2008)
- [79] Kendall, M. The treatment of ties in ranking problems. *Biometrika*. **33**, 239-251 (1945)
- [80] Khizar, N., Daud, Z. & Asad, Z. AN OPINION BASED STUDY REGARDING THE ROLE OF FORMATIVE ASSESSMENT IN ESL LEARNING. *PalArch's Journal Of Archaeology Of Egypt/Egyptology*. **18**, 708-717 (2021)
- [81] Knight, P. Summative assessment in higher education: practices in disarray. *Studies In Higher Education*. **27**, 275-286 (2002)

- [82] Kollar, I., Hämäläinen, R., Evans, M., De Wever, B. & Perrotta, C. Orchestrating CSCL-more than a metaphor?. (2011)
- [83] Krause, J., O’Neil, K. & Dauenhauer, B. Plickers: A formative assessment tool for K–12 and PETE professionals. *Strategies*. **30**, 30-36 (2017)
- [84] 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)
- [85] Kulkarni, C., Bernstein, M. & Klemmer, S. PeerStudio: rapid peer feedback emphasizes revision and improves performance. *Proceedings Of The Second (2015) ACM Conference On Learning@ Scale*. pp. 75-84 (2015)
- [86] Lasry, N.: Clickers or flashcards: Is there really a difference? *The Physics Teacher* **46**(4), 242–244 (2008)
- [87] Lasry, N., Mazur, E., Watkins, J.: Peer instruction: From harvard to the two-year college. *American journal of Physics* **76**(11), 1066–1069 (2008)
- [88] Li, H., Xiong, Y., Hunter, C.V., Guo, X., Tywoniw, R.: Does peer assessment promote student learning? a meta-analysis. *Assessment & Evaluation in Higher Education* **45**(2), 193–211 (2020)
- [89] Li, H. How is formative assessment related to students’ reading achievement? Findings from PISA 2009. *Assessment In Education: Principles, Policy & Practice*. **23**, 473-494 (2016)
- [90] Li, L., Zhang, Y. & Chen, L. Generate Neural Template Explanations for Recommendation. *Proceedings Of The 29th ACM International Conference On Information & Knowledge Management*. pp. 755-764 (2020,10)
- [91] Linn, M., Clark, D. & Slotta, J. WISE design for knowledge integration. *Science Education*. **87**, 517-538 (2003)
- [92] Lipton, Z. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.. *Queue*. **16**, 31-57 (2018)
- [93] Long, P. & Siemens, G. Penetrating the fog: analytics in learning and education. *Italian Journal Of Educational Technology*. **22**, 132-137 (2014)
- [94] Lyster, R. & Saito, K. Oral feedback in classroom SLA: A meta-analysis. *Studies In Second Language Acquisition*. **32**, 265-302 (2010)

- [95] Maheswaran, D. & Chaiken, S. Promoting systematic processing in low-motivation settings: Effect of incongruent information on processing and judgment.. *Journal Of Personality And Social Psychology*. **61**, 13 (1991)
- [96] Martinez, M.E., Lipson, J.I.: Assessment for learning. *Educational Leadership* **46**(7), 73–75 (1989)
- [97] 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)
- [98] McMillan, J.H., Hearn, J.: Student self-assessment: The key to stronger student motivation and higher achievement. *Educational Horizons* **87**(1), 40–49 (2008)
- [99] Meilandt, A. & Sell, L. *Improving User Onboarding Experience with Contextual Design*. (2021)
- [100] Meltzer, D.E., Manivannan, K.: Transforming the lecture-hall environment: The fully interactive physics lecture. *American Journal of Physics* **70**(6), 639–654 (2002)
- [101] Mense, E., Lemoine, P., Garretson, C. & Richardson, M. The development of global higher education in a world of transformation. *Journal Of Education And Development*. **2**, 47 (2018)
- [102] Millecamp, M., Broos, T., De Laet, T. & Verbert, K. DIY: learning analytics dashboards. *Companion Proceeding Of The 9th International Conference On Learning Analytics & Knowledge (LAK'19)*. pp. 947-954 (2019)
- [103] Miller, T. Formative computer-based assessment in higher education: The effectiveness of feedback in supporting student learning. *Assessment & Evaluation In Higher Education*. **34**, 181-192 (2009)
- [104] Miller, K., Schell, J., Ho, A., Lukoff, B. & Mazur, E. Response switching and self-efficacy in Peer Instruction classrooms. *Physical Review Special Topics-Physics Education Research*. **11**, 010104 (2015)
- [105] Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*. **267** pp. 1-38 (2019)

- [106] Montebello, M., Pinheiro, P., Cope, B., Kalantzis, M., Amina, T., Sears-Smith, 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)
- [107] Morris, R., Perry, T. & Wardle, L. Formative assessment and feedback for learning in higher education: A systematic review. *Review Of Education*. **9**, e3292 (2021)
- [108] Muijs, D.: Doing quantitative research in education with SPSS. Sage Publications (2004)
- [109] NiChroinin, D. & Cosgrave, C. Implementing formative assessment in primary physical education: teacher perspectives and experiences. *Physical Education And Sport Pedagogy*. **18**, 219-233 (2013)
- [110] David J Nicol and James T Boyle. 2003. Peer instruction versus class-wide discussion in large classes: A comparison of two interaction methods in the wired classroom. *Studies in higher education* 28, 4 (2003), 457–473.
- [111] Nussbaum, M., Alvarez, C., McFarlane, A., Gomez, F., Claro, S. & Radovic, D. Technology as small group face-to-face Collaborative Scaffolding. *Computers & Education*. **52**, 147-153 (2009)
- [112] Olsson, U.: Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**(4), 443–460 (1979)
- [113] Panadero, E. & Alqassab, M. An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation In Higher Education*. (2019)
- [114] Panadero, E. & Lipnevich, A. A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*. **35** pp. 100416 (2022)
- [115] 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)
- [116] Kathryn E Perez, Eric A Strauss, Nicholas Downey, Anne Galbraith, Robert Jeanne, and Scott Cooper. 2010. Does displaying the class results affect student discussion during peer instruction? *CBE—Life Sciences Education* 9, 2 (2010), 133–140.

- [117] Potter, T., Englund, L., Charbonneau, J., MacLean, M.T., Newell, J., Roll, I., et al.: Compar: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry* **5**(2), 89–113 (2017)
- [118] Rakhimov, I. & Kankarej, M. Forecasting the number of students in general education in University College using mathematical modelling. *Journal Of Mathematical Sciences: Advances And Applications*. **32** pp. 57-71 (2015)
- [119] Ramchander, M. & Naude, M. The relationship between increasing enrolment and student academic achievement in higher education. *Africa Education Review*. **15**, 135-151 (2018)
- [120] Reeves, T., Herrington, J. & Oliver, R. Design research: A socially responsible approach to instructional technology research in higher education. *Journal Of Computing In Higher Education*. **16**, 96-115 (2005)
- [121] B4MATIVE! project: <http://www.atief.fr/projet/b4mative>, last consulted: 2022-04-12.
- [122] Elaastic: <https://elaastic.irit.fr>, last consulted: 2022-04-11.
- [123] Reimann, P. Design-based research. *Methodological Choice And Design*. pp. 37-50 (2011)
- [124] Ribera, M. & Lapedriza, A. Can we do better explanations? A proposal of User-Centered Explainable AI. *Joint Proceedings Of The ACM IUI 2019 Workshops*. pp. 8 (2019)
- [125] Ricketts, C. & Wilks, S. Improving student performance through computer-based assessment: Insights from recent research. *Assessment & Evaluation In Higher Education*. **27**, 475-479 (2002)
- [126] Robbins, N., Heiberger, R. & Others Plotting Likert and other rating scales. *Proceedings Of The 2011 Joint Statistical Meeting*. **1** (2011)
- [127] Rudolph, J. A brief review of Mentimeter—A student response system. *Journal Of Applied Learning & Teaching*. **1**, 35-37 (2018)
- [128] Sadler, D. Formative assessment and the design of instructional systems. *Instructional Science*. **18**, 119-144 (1989)
- [129] Sawilowsky, S. New effect size rules of thumb. *Journal Of Modern Applied Statistical Methods*. **8**, 26 (2009)

- [130] Schafer, J., Konstan, J. & Riedl, J. Recommender systems in e-commerce. *Proceedings Of The 1st ACM Conference On Electronic Commerce*. pp. 158-166 (1999)
- [131] Scheeler, M., Ruhl, K. & McAfee, J. Providing performance feedback to teachers: A review. *Teacher Education And Special Education*. **27**, 396-407 (2004)
- [132] Schein, A., Popescul, A., Ungar, L. & Pennock, D. Methods and metrics for cold-start recommendations. *Proceedings Of The 25th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 253-260 (2002)
- [133] Selwyn, N. Should robots replace teachers?: AI and the future of education. (John Wiley & Sons,2019)
- [134] Schildkamp, K., Kleij, F., Heitink, M., Kippers, W. & Veldkamp, B. Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal Of Educational Research*. **103** pp. 101602 (2020)
- [135] Silvestre, F., Tranier, J. & Andriamiseza, R. Elaastic. (IRIT-Institut de Recherche en Informatique de Toulouse; Université Toulouse . . . ,2012)
- [136] Silvestre, F., Vidal, P. & Broisin, J. Tsaap-Notes–An Open Microblogging Tool for Collaborative Notetaking during Face-to-Face Lectures. *2014 IEEE 14th International Conference On Advanced Learning Technologies*. pp. 39-43 (2014)
- [137] 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)
- [138] Silvestre, F., Vidal, P. & Broisin, J. Reflexive learning, socio-cognitive conflict and peer-assessment to improve the quality of feedbacks in online tests. *European Conference On Technology Enhanced Learning*. pp. 339-351 (2015)
- [139] 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 Université, LIP6Paris, France (2019)

- [140] Smith, M.K., Wood, W.B., Adams, W.K., Wieman, C., Knight, J.K., Guild, N., Su, T.T.: Why peer discussion improves student performance on in-class concept questions. *Science* **323**(5910), 122–124 (2009)
- [141] Smith, M., Wood, W., Adams, W., Wieman, C., Knight, J., Guild, N. & Su, T. Why Peer Discussion Improves Student Performance on In-Class Concept Questions. *Science*. **323**, 122-124 (2009,1)
- [142] 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)
- [143] Stevens, S. On the theory of scales of measurement. *Science*. **103**, 677-680 (1946)
- [144] Sweller, J. Cognitive load theory. *Psychology Of Learning And Motivation*. **55** pp. 37-76 (2011)
- [145] Tempelaar, D., Heck, A., Cuypers, H., Kooij, H. & Vrie, E. Formative assessment and learning analytics. *Proceedings Of The Third International Conference On Learning Analytics And Knowledge*. pp. 205-209 (2013)
- [146] Thurlings, M., Vermeulen, M., Bastiaens, T. & Stijnen, S. Understanding feedback: A learning theory perspective. *Educational Research Review*. **9** pp. 1-15 (2013)
- [147] Tintarev, N. & Masthoff, J. Explaining recommendations: Design and evaluation. *Recommender Systems Handbook*. pp. 353-382 (2015)
- [148] Tissenbaum, M. & Slotta, J. Supporting classroom orchestration with real-time feedback: A role for teacher dashboards and real-time agents. *International Journal Of Computer-Supported Collaborative Learning*. **14**, 325-351 (2019)
- [149] Topping, K. Peer assessment between students in colleges and universities. *Review Of Educational Research*. **68**, 249-276 (1998)
- [150] Tullis, J.G., Goldstone, R.L.: Why does peer instruction benefit student learning? *Cognitive Research: Principles and Implications* **5**(1), 15 (Dec 2020)

- [151] 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)
- [152] Van Zundert, M., Sluijsmans, D. & Van Merriënboer, J. Effective peer assessment processes: Research findings and future directions. *Learning And Instruction*. **20**, 270-279 (2010)
- [153] Vanlommel, K., Van Gasse, R., Vanhoof, J. & Van Petegem, P. Teachers' decision-making: Data based or intuition driven?. *International Journal Of Educational Research*. **83** pp. 75-83 (2017)
- [154] Verbert, K., Duval, E., Klerkx, J., Govaerts, S. & Santos, J. Learning Analytics Dashboard Applications. *American Behavioral Scientist*. **57**, 1500-1509 (2013,10)
- [155] 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) @article-wang2005design, title=Design-based research and technology-enhanced learning environments, author=Wang, Feng and Hannafin, Michael J, journal=Educational technology research and development, volume=53, number=4, pages=5–23, year=2005, publisher=Springer
- [156] Wang, F. & Hannafin, M. Design-based research and technology-enhanced learning environments. *Educational Technology Research And Development*. **53**, 5-23 (2005)
- [157] Weinstein, Y., Sumeracki, M. & Caviglioli, O. Understanding how we learn: A visual guide. (Routledge,2018)
- [158] White, B. & Frederiksen, J. Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition And Instruction*. **16**, 3-118 (1998)
- [159] Wiliam, D. What is assessment for learning?. *Studies In Educational Evaluation*. **37**, 3-14 (2011)
- [160] Xie, Y., Fang, M. & Shauman, K. STEM education. *Annual Review Of Sociology*. **41** pp. 331-357 (2015)
- [161] Yüksel, H. & Gündüz, N. Formative and summative assessment in higher education: Opinions and practices of instructors. *European Journal Of Education Studies*. (2017)

- [162] Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y. & Ma, S. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. *Proceedings Of The 37th International ACM SIGIR Conference On Research & Development In Information Retrieval*. pp. 83-92 (2014)
- [163] Zheng, L., Zhang, X., Cui, P.: The role of technology-facilitated peer assessment and supporting strategies: a meta-analysis. *Assessment & Evaluation in Higher Education* **45**(3), 372–386 (2020)

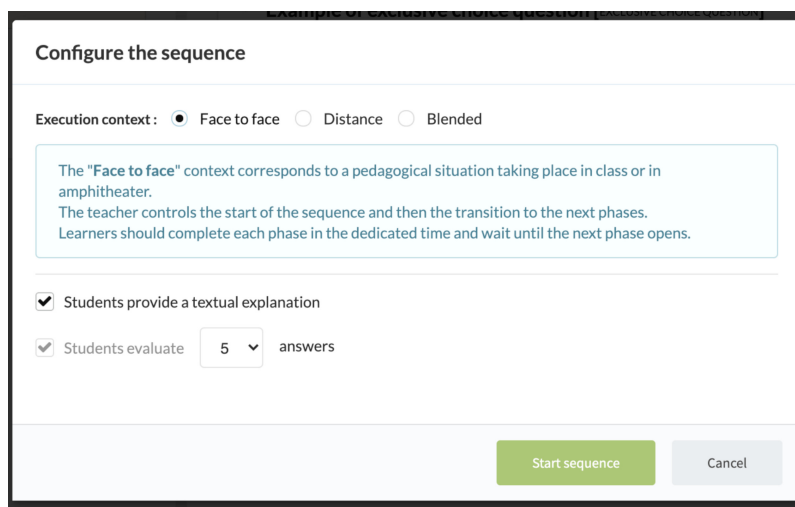
Appendix

Appendix A

Asynchronous Settings

A.1 Execution Contexts

A sequence can be run in asynchronous and synchronous contexts [139]. More precisely, when starting a sequence, teachers can choose between three execution contexts as shown in Figure A.1. The definitions for each execution



Configure the sequence

Execution context: Face to face Distance Blended

The "Face to face" context corresponds to a pedagogical situation taking place in class or in amphitheater.
The teacher controls the start of the sequence and then the transition to the next phases.
Learners should complete each phase in the dedicated time and wait until the next phase opens.

Students provide a textual explanation

Students evaluate answers

Figure A.1: Elastic: Configuration popup before teachers start a sequence

context are the following:

1. The "**Face to face**" context corresponds to a pedagogical situation taking place in class or in amphitheater. The teacher controls the start of the sequence and then the transition to the next phases. Learners should complete each phase in the dedicated time and wait until the next phase opens.

This execution context is the synchronous one, it can also be used for synchronous distance settings such as virtual classrooms.

2. The "**Distance**" context corresponds to a pedagogical situation for which learners are in a situation of autonomy. The teacher controls only the opening and closing of the sequence. Each learner has the opportunity to do one phase after the other at his own pace, and then immediately discover the results.

This execution context is the asynchronous one.

3. The "**Hybrid**" context corresponds to a pedagogical situation taking place at a distance followed by a presentation of the results in face-to-face. The teacher controls the opening of the sequence and the publication of the results. Learners can follow the first two phases at their own pace, but will not discover the results until they are published.

This execution context is the half synchronous and half asynchronous one.

A.2 The Cold Start Problem

Let us note that when it comes to distance and hybrid contexts, learners can provide a rationale for their second answer, which is not the case in face to face contexts. Another difference between the face-to-face contexts and the other is the occurrence of the "cold start problem". Such problem occurs when there is a lack of information, on users or items [132]. In our context, since a peer grading activity occurs in phase 2, some rationales written by learners must already exist. Although it is not an issue in face-to-face context due to the synchronisation of all learners (i.e. phase 2 starts when rationales have been provided), the very first learner who completed phase 1 of an asynchronous sequence can not be provided with any of his peers rationales to grade during phase 2. To work around this problem, a feature is made available in the question creation form [139]. This feature allows teachers to write rationales that will blend with learners' rationales during the sequence.

A.3 Alternative to ρ_{conf} and ρ_{peer}

We conducted deeper analysis to better understand the relation between ρ_{conf} and ρ_{peer} . More precisely, we have seen in Section 6.2.5 that across a sequence, learners peer evaluation are globally expected to be consistent

if their confidence degree is globally consistent. Our next step is to verify whether this finding applies on an individual level. Consequently, we made the hypothesis that a peer evaluation is more likely to be consistent when the learner who performed the evaluation is consistently confident. To this end, we proposed two equivalent to ρ_{conf} and ρ_{peer} but on an individual level. In order to compute each learner's consistency of confidence degree we proposed the following equation: $l_{conf} = correctness * confidence$. In this equation, the *correctness* is a binary variable that is equal to 1 if the learner answered correctly and is equal to -1 otherwise. The variable *confidence* is the confidence degree as reported by the learner on a Likert scale when submitting her answer. It ranges from 1 to 4 (Not confident at all, not really confident, confident and absolutely confident). The consistency of each evaluation is computed as follows: $l_{peer} = correctness * agreement$ with *agreement* being the level of agreement of a learner to an answer as reported on a Likert scale during the confrontation phase. It ranges from -1 to 1. More precisely, agreement is equal to -1 if learner strongly disagrees (1/5) or disagrees (2/5), 0 if the learner does not agree nor disagree (3/5) and if the learner agrees (4/5) or strongly agrees (5/5). We purposefully decided not to distinguish between strongly disagree and disagree as well as between strongly agree and agree because that level of precision regarding the level of agreement would require a deeper analysis of the rationale. The set of possible values for l_{conf} is finite and is equal to $\{-4; -3; -2; -1; 1; 2; 3; 4\}$ whereas the one for l_{peer} is equal to $\{-1; 0; 1\}$. Figure A.2 summarises l_{conf} and l_{peer} across the 4,072 evaluations of our sample. On this figure, it would

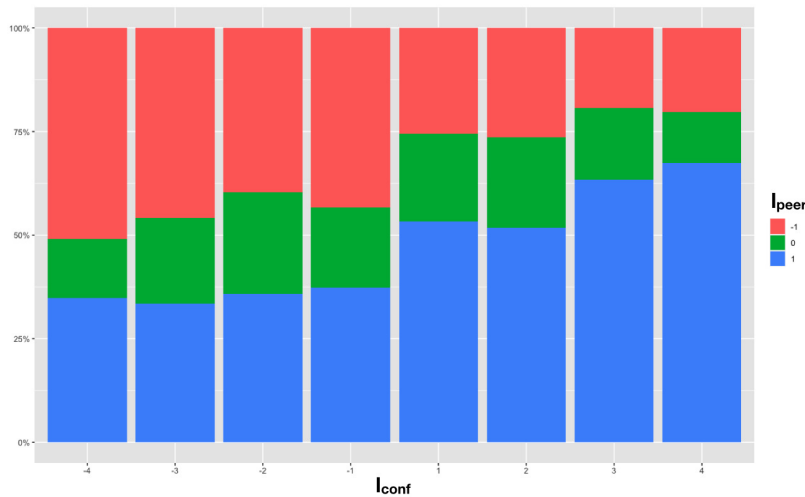


Figure A.2: l_{conf} depending on l_{peer} from s_{2019} .

appear that l_{peer} is more frequently equal to 1 and less frequently equal to -1

as l_{conf} increases. In other words, an evaluation is more likely to be consistent when the learner who performs it is consistently confident. As stated earlier, since both these equations are based on data reported on Likert scales, they must be considered as ordinal. Therefore we use a correlation based on rank that can handles ties. We computed the correlation of Kendall [79] and obtained a moderately high tau equal to 0.21 with a p-value $< 2.2e - 16$, and a 95% confidence interval equal to [0.19:0.23]. Such results can serve as a basis to adapt the recommendations model to different execution contexts.

Appendix B

Relevancy of Rationales

B.1 Cosine Similarity

The algorithm used here is the cosine similarity. More precisely, we average the cosine of each answer with all the other answers as follows:

$$c\bar{o}s_D(d_i) = \frac{\sum_{j \in D} |\cos(d_i, d_j)|}{|D|} \quad (\text{B.1})$$

In this equation, D is the set of all rationales from a sequence and d_i is the current rationale for which we compute the relevancy.

B.2 Early Results

Since rationales are written in any configuration of Elaastic, we analysed rationales from sequences conducted with any execution context and which contained any kind of question. We ended up with 190 sequences and 3,468 rationales. As an example, a sequence asked the following open-ended question to learners:

You want to evaluate the thermal exchange between air and a radiator at Mach = 0.01 in a 3D flow. Which lattice / technique will you use? Please justify your choice.

Table B.1 shows the 5 rationales with the lowest computed relevancy. Let us note that the typos were kept for authenticity purposes.

We can see that the two rationales with the lowest computed relevancy are the ones with one word only. Such rationales only give the answer and

Rationale	Average Cosine
D3Q39	0.08
D3Q27	0.08
D3Q27, D3Q39, D3Q103 for lattice	0.09
D3Q103 because it has been the only one presented that allows changes of temperature	0.17
I would use a hybrid model with an energy equation or D3Q103 if affordable.	0.18
D3Q103. Because it's a thermal calculation and 3D.	0.18
I will use a DDF technique in order to recover all needed equations at lower cost..	0.19

Table B.1: Average cosine of rationales ordered from lowest to highest.

do not provide any justification as a rationale is expected to do. However these rationales could already be detected as irrelevant based on the number of words. The third rationale with the lowest computed relevancy however is a rationale with no detailed justification that could not be classified as irrelevant based solely on the number of words.