JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 6(1), 54 – 84. http://doi.org/10.18608/jla.2019.61.4

From Student Questions to Student Profiles in a Blended Learning Environment

Fatima Harrak[1], François Bouchet[2], Vanda Luengo[3]

Abstract

The analysis of student questions can be used to improve the learning experience for both students and teachers. We investigated questions (*N* = 6457) asked before the class by first-year medicine/pharmacy students on an online platform, used by professors to prepare for Q&A sessions. Our long-term objectives are to help professors in categorizing those questions, and to provide students with feedback on the quality of their questions. To do so, we developed a coding scheme and then used it for automatic annotation of the whole corpus. We identified student characteristics from the typology of questions they asked using the k-means algorithm over four courses. Students were clustered based on question dimensions only. Then, we characterized the clusters by attributes not used for clustering, such as student grade, attendance, and number and popularity of questions asked. Two similar clusters always appeared (lower than average students with popular questions, and higher than average students with unpopular questions). We replicated these analyses on the same courses across different years to show the possibility of predicting student profiles online. This work shows the usefulness and validity of our coding scheme and the relevance of this approach to identify different student profiles.

Notes for Practice

- Questions provide important insights into students' level of knowledge, but coding schemes are lacking to study this phenomenon.
- After providing a bottom-up coding scheme of student questions in a blended environment, we analyzed the relationship between the questions asked and the student profiles.
- Profiling students based on their questions over a year allows us to predict the profiles of future students to help the teacher understand who asks what.
- These results provide both a coding scheme that can be reused in various contexts involving questions, and a methodology that can be replicated in any context where students ask many questions, in particular to help the teacher in prioritizing them according to their own criteria.
- Teachers need to focus on the nature of questions asked by their students, because they can reveal information about their profile (attendance, activity, etc.).

Keywords

Clustering, question coding scheme, student behaviour, blended learning.

Submitted: 15.06.2018 — **Accepted**: 24.01.2019— **Published:** 15.04.2019

Corresponding author[1] Email: fatima.harrak@lip6.fr Address: Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, F-75005 Paris, France

[2] Address: Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, F-75005 Paris, France

[3] Address: Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, F-75005 Paris, France

1. Introduction

Student questions play an important role in the learning process as they are meaningful for both learning and teaching science (Chin & Osborne, 2008). Questions may help learners to acquire the knowledge and improve their understanding (Chin & Brown, 2000), as well as to stimulate them to consider different perspectives of an issue, thus enhancing the quality of their learning. Graesser, Person, and Huber (1992) suggested that one of the mechanisms that generate questions stems from a need to correct declarative knowledge deficits. Student questions not only help their learning process, but also serve as a pedagogical tool for the teacher. Questions that students ask can help the teacher better understand student thinking, thereby enabling instructional decisions better tailored to their needs (White & Gunstone, 1992; Etkina & Harper, 2002). For example, knowing

the difficulties students are facing helps a teacher to provide analogies, clarification, examples, and questions that assist students in understanding the content, and making more targeted decisions (Harper, Etkina, & Lin, 2003; Colbert, Olson, & Clough, 2007).

Student questions can reflect active learning, knowledge construction, curiosity, and the depth of the learning process (Graesser, Ozuru, & Sullins, 2009). The analysis of these questions may also indicate the type and nature of questions asked by successful and unsuccessful students. Therefore, student questions not only provide feedback to the teacher about student understanding but, according to these questions, the teacher often has an "intuition" about a student's level and profile in the classroom (e.g., if they are retaking the year or attending the course regularly). Although student questions in the classroom can sometimes be both infrequent and unsophisticated, online environments might remove many of the barriers that prevent students from asking questions in the classroom (Otero & Graesser, 2001). It is also relevant in the MOOC context where various research has tried to extract information from student messages (Elgort, Lundqvist, McDonald, & Moskal, 2018). However, in a university context, dealing with student questions asked online can be problematic from the teacher's point of view, one of the main issues being the volume of questions asked, preventing them from answering every question in class. Another typical issue is the difficulty for the teacher to link a student's profile to the questions asked, leading the teacher to empirically choose the questions to answer in priority, instead of being able to use a more advanced strategy (e.g., to not answer only questions from good students, to evaluate differently a question coming from a student not usually attending the course or from a student retaking the class).

More particularly, we are interested in whether the type of questions asked by students online can be related to their performance and their overall learning behaviour. We investigate this question in the context of a hybrid curriculum (such as Liu et al., 2016), in which every week students must ask questions regarding the online material they studied at home (following a flipped classroom approach), before the class, to help professors prepare their Q&A session. In our context, one week before their course, teachers currently receive via email a list of questions asked by students (a so-called "wall of questions"), sorted only by the number of votes for each of them, making it challenging for them to choose the set of questions to address (Harrak, Bouchet, & Luengo, 2017; Harrak, Bouchet, Luengo, & Gillois, 2018). Our overall goal, therefore, is to provide teachers with additional information to help them choose questions for their Q&A session, according to the strategy they see fit, thus "breaking down the wall of questions." Thus the purpose of this paper is to identify the relationship between students' questions and their profiles, based on existing data, and to use this analysis in predicting future student profiles in the next iterations of the course.

The remainder of this paper will be structured as follows: in section 2, we review the different taxonomies of student questions and address the research previously carried out on the relationship between a student's questions and their behaviour. In section 3, we introduce in more detail our context and the dataset used in this paper. In sections 4 and 5, we present the context of the use case and the methodology followed to define a coding scheme of student questions and to automatize the tagging of questions according to this coding scheme. Then, in section 6 we identify a set of available characteristics about the students (performance, attendance, votes, etc.) and study the links between the questions they asked and their characteristics using a clustering approach. Following this, we investigate the possibility of reliably predicting student profiles online, concluding with some perspectives for future research.

2. Related Work

2.1. Nature and Type of Student Questions

Chin and Osborne (2008) present an overview of the studies conducted on the characterization of the types of questions that students ask. Scardamalia and Bereiter (1992) distinguished between text-based questions and knowledge-based questions. Text-based questions refer to questions that students ask in response to reading a text, while knowledge-based questions are generated spontaneously to make sense of the world and to deepen a concept. They found that knowledge-based questions hold greater educational potential than text-based questions. Although this distinction seems relevant to us, we did not have access to transcriptions of the videos and slides that students interacted with before asking their questions online. It was, therefore, difficult to consider identifying automatically whether a question was text-based or not. Furthermore, Chin and Kayalvizhi (2002) proposed a typology of questions that distinguish investigable and non-investigable questions to guide students in generating questions that they can investigate on their own. Investigable questions refer to those to which students can find the answers by designing and performing hands-on investigations themselves, including comparison, cause-and-effect, prediction, design-and-make, exploratory, descriptive, and validation of mental model questions. Non-investigable questions do not lend themselves to hands-on investigations; they include basic or complex information questions. Although our goal is not to prescribe students to ask a certain type of question, this distinction could be relevant to our work; however, without the teacher's help we cannot immediately identify whether questions are amenable to practical inquiry or not. Bouchet

(2009) proposed a categorization of requests made in French to an assistant conversational agent, which distinguished between direct and indirect assistance requests. Although it is one of the few resources available in French, it appeared to be too coarse-grained for our needs here.

Another possible way to classify student questions can be according to the level of thought (or cognitive processes) required for answering them. Graesser and Person (1994) developed a taxonomy of questions asked during tutoring sessions to be used for automatic question generation. Although their taxonomy could be relevant to our work, some categories included high quality "deep-reasoning questions" associated to patterns of reasoning that are difficult to identify automatically (specifically, questions such as those described as "antecedent," "consequence," "goal orientation," or "instrumental/procedural and enablement"). Another well-known taxonomy regarding questions and education is Bloom's taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956), which was originally created for the purpose of formulating questions by the teacher. Anderson and Krathwohl (2001) revised this taxonomy to accommodate for a more highly differentiated range of cognitive processes subsumed under six major categories: remember, understand, apply, analyze, evaluate, and create. Recent research (Supraja, Hartman, Tatinati, & Khong, 2017) used a reduced version of Bloom's taxonomy to link actionable feedback to a learner's performance on assessment items. Because of its origins, this taxonomy tends to be more appropriate for teachers' questions (e.g., assessment) than student ones.

Pizzini and Shepardson (1991) developed another coding scheme that categorizes student questions under three types using cognitive levels as a criterion: 1) input: recall information; 2) processing: draw relationships among data; and 3) output: hypothesize, evaluate. The taxonomy has been developed to compare the quantity and quality of student questions in problem-solving instruction versus teacher-directed lab instruction. However, with only three levels, this scheme is not sufficiently descriptive of the various, nuanced cognitive levels.

Another student question classification was proposed by Pedrosa de Jesus, Teixeira-Dias, and Watts (2003), who defined two main categories: 1) confirmation questions that seek to clarify information and ask for an example or a definition, and 2) transformation questions that involve some reorganization of student understanding. According to the authors, both categories are necessary and complement each other; the appropriate type of question depends on the nature of the situation. In the same direction, Watts, Gould, and Alsop (1997) classified questions under three categories corresponding to different moments in the process of conceptual change: 1) consolidation (confirming explanations and consolidating new ideas); 2) exploration (seeking to expand knowledge and testing constructs); and 3) elaboration (reconciling different understandings, resolving conflict).

Marbach-Ad and Sokolove (2000b) proposed a taxonomy of questions to help university biology students identify good questions. The taxonomy proposed to categorize questions into four major types: 1) those based on a misunderstanding or misconception; 2) those about a simple or complex definition; 3) those involving students using information beyond that given in the textbook; and 4) those involving students using higher-level thinking skills. However, the main problem of this taxonomy consisted in the difficulty of automating the identification of questions into these categories, which is essential in our context.

In all the studies discussed so far, student questions were asked in the context of formal instruction in the classroom and the taxonomies proposed are mostly context-dependent and rarely provide an extensive set of keywords to allow for practical automatic identification. Therefore, we decided it would be more relevant to our end-goal to define a new coding scheme using a bottom-up data-driven approach, where keywords and expressions are identified as the scheme is defined, and to contrast it *a posteriori* to some of the aforementioned works (*cf.* Table 4 in section 4.2).

2.2. Relationship between Student Questions and Student Behaviour

Researchers have studied students' questioning behaviour in a variety of educational settings, such as classroom, tutoring, and online learning environments (Li et al., 2014). There have also been some recent studies that focused on detecting questions from recorded audio (Cook, Olney, Kelly, & D'Mello, 2018) instead of text. The analysis of learning questions has been used for a wide variety of purposes in order to improve teaching effectiveness and student learning. Much research investigated the relationship between student questions and a variety of variables, such as achievement level, engagement, learning approaches, learning styles, and nature of instruction.

Harper et al. (2003) investigated the relationship between the types of questions that students asked on a weekly written report and their conceptual understanding of various topics. However, there was no significant correlation between the number of questions asked and conceptual achievement. Work has also been done by Cao, Tang, and Hu (2017) to investigate students' questioning behaviours on MOOC discussion threads. The authors used the quantity of questions, classification, and variations over time, but they were not able to identify clear relationships between the features they considered and the threads categories.

Chin and Brown (2002) focused on the relationship between student questions, the nature of their thinking, and actions undertaken during the knowledge construction process. They showed that the kind of questions that students ask depends on the way they approach their learning tasks. Indeed, the students' "basic information" questions that focused on facts and

procedures (typical of a shallow learning approach) generated little productive discussion. In contrast, "wonderment questions" that focused on comprehension, prediction, anomaly detection, application, and planning (characterizing a deep learning approach) led students to engage in more thoughtful ideas and group talk. Likewise, the study conducted by Li et al. (2014) examined the frequency of questions asked during collaborative problem solving and depth of questions in terms of both task difficulty and game phase. They found that the players in an online serious game environment for learning asked more shallow questions than deep ones. Similarly, Pedrosa, Almeida, and Watts (2004) studied the relationships between student questioning and different learning styles (diverger, converger, assimilator, accommodator) based on Kolb's theory (1984, 1985) using four case studies in a university chemistry course. However, they failed to identify any influence of the learning styles considered on the kinds of questions that students asked.

A study done by Marbach-Ad and Sokolove (2000a) compared active versus traditional learning. They found that undergraduate biology students who learned using an active learning approach (using interactive instructional approaches in co-operative learning groups) were able to ask better, higher-level questions after reading chapters from the textbook than those taught in a traditional lecture format. Therefore, the results show that student questions were dependent on the nature of instruction.

Cluster analysis can also be used to help researchers analyze characteristics of learning behaviour and develop profiles that are grounded in learner activities (Antonenko, Toy, & Niederhauser, 2012). Sindhgatta, Marvaniya, Dhamecha, and Sengupta (2017) used hierarchical agglomerative clustering to identify groups of questions from forums to provide lecturers with information about frequent questions asked. Their approach was based on the similarity of the words in the question, using a combination of both lexical and semantic similarities. Closer to our work, using natural language processing techniques, Kim, Shaw, and Ravi (2010) helped students participating in online discussions by retrieving useful messages, and then used supervised machine learning approaches to classify student messages as questions or answers to analyze the distribution of student participation by gender. A slightly different approach has been used by Ishola and McCalla (2017) in order to predict prospective helpers who can provide timely answers to questions. They used strategies based on frequency, knowledgeability, eagerness, willingness, and recency besides timeliness measures. Their results are promising for providing just-in time help to lifelong learners in online forums.

Overall, these studies tend to consider a given student behaviour or characteristic (learning style, level of understanding, etc.) and look for differences in the questions asked by various student profiles. Although similar, our approach takes a different direction as we first extract a student profile based only on the questions they have been asking, and then look for correlations between question profiles and other features (level, course attendance, etc.). Moreover, another difference in our approach is that after an exploratory analysis we try to predict student profiles online, i.e., using a model trained on data from students who took the course in previous years. More precisely, our goal was to answer four research questions:

RQ1: Can we propose a relevant coding scheme to analyze student questions in terms of intentions?

RQ2: How well can we automatize the annotation of student questions according to this coding scheme?

RQ3: Once annotated according to the defined coding scheme, can questions asked by students be used to identify student profiles that are substantially different in terms of performance and various aspects of course behaviour (attendance, number of questions, and votes)?

RQ4: Can we use a model trained over a previous session of a course to reliably predict student profiles in the following sessions of that course?

3. Context and Data Set Used

The dataset we considered is comprised of questions asked in 2012–2013 by first-year medicine/pharmacy students at a major public French university (Université Joseph-Fourier). Not all of them asked questions, but 1608 students were enrolled that year. The Faculty of Medicine and Pharmacy has a specific hybrid training system (part of the work is done remotely and the other part in the classroom) for their first-year students. The first year (called PACES) takes place over a year divided into two semesters. Each semester ends (in January and May) with a competitive exam on the content studied during the period. A proportion of the students who fail the competitive exam are allowed to retake the year only once. Each course is composed of four to six 4-week sequences on the PACES[1] platform, as illustrated in Figure 1.

In every sequence, the first week consists of studying the course on a DVD-ROM or on the website Medatice (slides and videos from the professor). The second week is dedicated to the Online Formulation of Questions (OFQ) for teachers related exclusively to the multimedia courses studied the previous week. Students can see questions asked by other students in their group and vote for them if they also want an answer to that question, but they cannot answer them. At the end of the week, the

[1] paces.medatice-grenoble.fr

questions are sent to the adjunct professors intervening during the third week, who use them to structure their interactive classroom teaching sessions. During these sessions, the professor answers some of the questions asked online by students. The fourth week is dedicated to tutoring sessions in order to test the knowledge acquired during the training sequence through self-testing using multiple choice questions (MCQ), which are then corrected by an adjunct professor. There are two tutoring sessions of two hours/week (about 48 sessions/year and 40 MCQ/sessions). Each student can individually check their grades and rankings relative to the overall promotion. These tutoring results allow students to assess the quality of their learning.

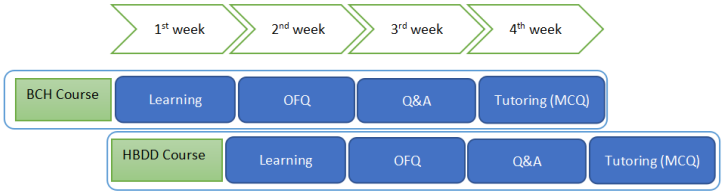

Figure 1: Learning activities over a sequence (4 weeks) / course

For each of the 13 courses, therefore, we have four to six sets of questions (one set per sequence) asked by 429 students (6457 questions overall) during the second week of each period. Not all courses received the same number of questions: courses from the first semester received more questions, in particular because some students are forced to quit at the end of the first semester based on their results in the exam. The question distribution per course is shown in Table 1. It is worth noting that overall only about 25% of students asked at least one question. This phenomenon could be explained by the fact that instructors encourage students to vote on existing questions instead of asking new ones, not only to force them to read what others have asked, but also to reduce the number of questions they receive by mail for their Q&A session. We also hypothesize that not all instructors rely on the online platform as much as they should to choose questions to address during their Q&A session, which could partially explain why fewer students ask questions on the platform over time.

In addition to questions, we also have access to some information about the 1608 students enrolled. We were able to define a set of nine variables relative to each student and each course, which are summarized in Table 2.

Table 1: Distribution of questions asked/course

BCH	BPH	HBDD	BCE	ANT	PHS	SSH	ICM	MAT	Spec.
19%	17%	15%	11%	10%	9%	8%	6%	3%	1%

Note: BCH = Biochemistry, BPH = Biophysics, HBDD = Histology & developmental biology, BCE = Cellular biology, ANT = Anatomy, PHS = Physiology, SSH = Health, human, society, ICM = Knowledge of drugs, MAT = Mathematics, Spec. (Specialty)= Pharmacy, Odontology, Maieutic

Table 2: Description of the 9 variables available for each student for each course.

Variables	Description
AvgGrd*	Average grade on the tutoring sessions of each course (out of 20)
FinGrd*	Final grade in the final exam on each course (out of 20)
GlbAtt	Global attendance ratio over the 2 semesters (from 0 [never there] to 1 [always there])
CouAtt	Attendance ratio for each course (from 0 [never there] to 1 [always there])
NbQst	Number of questions asked in each course
NbVotRec	Number of votes received from other students on their questions in each course (0 if no questions asked)
NbVotDone	Number of votes on other students' questions in each course
RetStu	Whether the student was retaking the year (1 for students whose rank in the final exam is inferior to 200th, 0 otherwise)
SucStu	Whether the student was eventually successful after the final exam (1 for students whose rank in the final exam is inferior to 200th, 0 otherwise)

* When a student was a no-show to a MCQ at the end of a sequence or for the final exam at the end of the course, their corresponding grade was set to 0, in accordance with university policy.

4. Question Coding Scheme Definition

4.1. Method

To answer RQ1, we considered a sample of 800 questions (12% of the corpus size) from two courses (biochemistry [BCH], histology & developmental biology [HBDD]), considered by the pedagogical team to be among the most difficult ones and which had the highest number of questions asked (*cf.* Table 1). This sample was randomly divided into four sub-samples of 200 questions to apply four successive categorization steps: 1) discovery, 2) consolidation, 3) validation, and 4) evaluation.

The **discovery step** consisted of empirically grouping sentences with similarities to extract significant concepts. Although students were instructed to ask simple questions (i.e., centred on a single topic, avoiding questions such as "Could you re-explain X? Also, Y wasn't clear"), we found that a significant subset of the questions (40%) could be broken down into several independent questions. Once sentences were segmented into so-called simple questions, we grouped questions that appeared similar in structure (e.g., "what is X?" and "what is Y?") and in meaning (e.g., "what is X? and "could you define X?"). Groups of questions were then given "labels" (e.g., "definition of a concept") that could be grouped into higher level categories. Then we identified mutual exclusion between labels (e.g., a simple question cannot be both a verification and a request for re-explanation) and, conversely, labels that could co-occur (e.g., both a verification and a re-explanation could be relative to the correction of an exercise). This led us to define a set of what we called *dimensions*, which are made of sets of labels of similar but mutually exclusive types of questions (in the previous example, one cannot at the same time verify the validity of an assertion and ask for a re-explanation of a concept). Each of these individual labels ("verification," "re-explanation," etc.) are the *values* that can be associated to a dimension. At the same time as a dimension value was identified, the human annotator identified the *keywords* or idiomatic expressions that were indicative of that dimension value (e.g., in Dimension1, for the dimension value "Re-explain," some of the keywords identified would be "re-explain," "restate," "redefine," "retry," "repeat," "revise," "get back on," etc.). In summary, the coding scheme is made of dimensions that form a set of values with a list of associated keywords for each. Each simple question can then be associated with an *annotation* in that coding scheme by choosing, for each dimension, one and only one value, based on the associated keywords. A question annotation can thus be seen as a vector with N values, N being the number of dimensions in the coding scheme (*cf.* Figure 2 for an example of an annotation showing how certain keywords are used to represent a question as a vector).

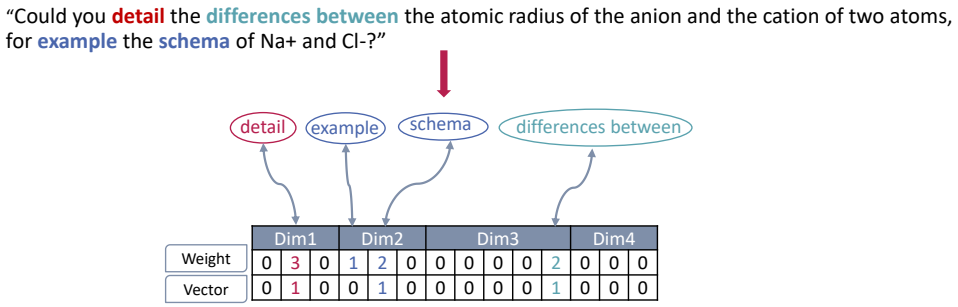

Figure 2: Example (translated) of an automatic annotation of a question using weighted keywords.

The **consolidation step** consisted in annotating the second sub-sample to validate the dimensions and values previously identified. This led to various adjustments to the dimensions to make sure they were indeed independent from each other (e.g., addition of the value "correction" in Dim2, not previously identified). In parallel, the identified dimensions were reviewed and validated by an expert professor teaching in the PACES context, who found the categories were potentially relevant to analyze student questions.

In the **validation step**, we performed a double annotation to validate the generality of our categories on the third sub-sample of 200 sentences. First, the 200 sentences were manually segmented, providing 238 segments with a single question in them. Then two human annotators used as a unique reference the coding scheme created at the end of the previous step to annotate each of these segments. There were four dimensions in this consolidation step: Dim1 (relative to the question type), Dim2 (relative to the explanation modality), Dim4 (optional, only annotated if the question is a verification, relative to the nature of what is being checked) and another dimension (relative to the occurrence or not of mistakes in sentences such as grammatical errors, words missing, orthographic mistakes, etc.). This latter dimension is not reported here because it was excluded due to the difficulty of automating it. The dimension later called "Dim3" did not exist yet in this step. First, the human annotators made two separate and independent annotations on each dimension, and their agreement was evaluated using Cohen's kappa (K_1=0.72, K_2=0.62 where K_1 and K_2 correspond respectively to the kappa for Dim1 and Dim2). For Dim4, because of its optional status, the two annotators did not necessarily annotate the same questions: one annotator annotated 82

questions, the other 68 questions, with an overlap of 68 questions. The kappa value calculated over those 68 questions was 0.66. Then they met to discuss and resolve the disagreements, which mostly corresponded to ambiguous cases. This led to a final refinement of the categories (e.g., separation of categories Dim1 and Dim4, addition of category Dim3). Eventually, the entire sample (600 sentences) was re-annotated on four dimensions (Dim1 to 4) to consider the changes and to provide a grounded truth to which the automatic annotation could be compared. This final version of the coding scheme with four dimensions is provided in Table 3.

Finally, in the **evaluation step**, the last sub-sample of 200 segments was annotated manually by the two expert annotators (with an increased kappa of 0.83 on Dim1, 0.76 on Dim2, and 0.47 on Dim3). This sub-sample, not used for the training of the automatic annotator, was used for its testing (*cf.* section 5).

4.2. Results and Discussion

The main result of this step was the creation of the question coding scheme provided in Table 3. This coding scheme is presented here in English, with a non-exhaustive list of French keywords translated as well as possible to provide an idea of their nature. But we also established some rules from the manual annotation step that were useful to automatically annotate the rest of the data:

- All questions must be simple (i.e., combined questions must be segmented into several simple ones).
- The annotation must be unique on each dimension (i.e., a simple question can correspond to the category "deepen a concept" or "re-explain a concept," but not both), but it is possible to have no annotation on some dimensions (e.g., a question can be neither about an example, nor a schema, nor a correction).
- Dimension 4 (verification) is annotated only for the questions identified as verification on dimension 1. This dimension is the most difficult one to identify automatically and might ideally require some semantic analysis.

Table 3: Proposed coding scheme created (translated from French) from manual annotation

Dim1	**Question type**	**Description**	**Keywords**
1	Re-explain / redefine	Ask for an explanation already provided in the course material	Re-explain, restate, revise, repeat, retry...
2	Deepen a concept	Broaden knowledge, clarify an ambiguity, or request for a better understanding	Explain, detail, precise, develop, describe...
3	Validation / verification	Verify or validate a formulated hypothesis	Can we, is it...
Dim2	**Explanation modality / Subject of question**	**Description**	**Keywords**
1	Example	Example application (course/exercise)	Example…
2	Schema	Schema application or an explanation about it	Schema, representation…
3	Correction	Correction of an exercise in course/exam	Correction, response…
Dim3	**Explanation type**	**Description**	**Keywords**
1	Define	Define a concept or term	Define, signify, mean…
2	Manner (how?)	The manner in which to proceed	How…
3	Reason (why?)	Ask for the reason	Why, reason…
4	Roles (utility?)	What's the use / function	Role, utility, function…
5	Link between concepts	Verify a link between two concepts, define it	Matches, difference between, relation, similar…
Dim4	**Verification type (optional)**	**Description**	**Keywords**
1	Mistake / contradiction	Detect mistake/contradiction in course or explanation of teacher	Contradiction, mistake…
2	Knowledge in course	Verify knowledge	(*)
3	Exam	Verify information about an exam	Exam, should we know…

(*) There are no specific keywords in this subcategory because the subject matter would require semantic analysis. This subcategory is annotated, therefore, only if the question of verification is not a verification of mistake or exam (Dim4-1 and Dim4-3 are not annotated).

Although developed independently, we noticed *a posteriori* overlap between some of the aforementioned state of the art taxonomies and our coding scheme. In particular, dimension 1 matches well categories from several taxonomies, such as "Re-

explain/redefine" (Dim1-1), which is included in "Non investigable" (Chin & Kayalvizhi, 2002) and "Transformation" (Pedrosa et al., 2003). The explanation modality (Dim2) seems less common, with only the notion of "example" included in the category "Confirmation" (Pedrosa et al., 2003) and "Example" (Graesser & Person, 1994). Likewise, only the categories "define" and "link between concepts" in dimension 3 are included respectively in "Non investigable" and "Investigable" (Chin & Kayalvizhi, 2002). We note also that some categories in dimension 4 can be found in other taxonomies, such as the category "mistake/contradiction," which corresponds to the category "Elaboration" (Watts et al., 1997), and "knowledge in course" which is included in "Confirmation" (Pedrosa et al., 2003). Table 4 summarizes these overlaps.

Table 4: Summary of similarities between existing question taxonomies and the proposed coding scheme

Dim1	Question type	(Chin & Kayalvizhi, 2002)	(Watts et al., 1997)	(Pedrosa et al., 2003)	(Graesser & Person, 1994)
1	Re-explain / redefine	Included in Non-investigable	–	Included in Transformation	–
2	Deepen a concept	–	Included in Exploration	Included in Transformation	–
3	Validation / verification	Investigable (partially)	Consolidate (partially)	Confirmation (partially)	Verification (partially)
Dim2	**Explanation modality / Subject of question**				
1	Example	–	–	Included in Confirmation	Included in Example
2	Schema	–	–	–	–
3	Correction	–	–	–	–
Dim3	**Explanation type**				
1	Define	Included in Non-investigable	–	Included in Confirmation	Included in Definition
2	Manner (how?)	–	–	–	–
3	Reason (why?)	–	–	–	–
4	Roles (utility?)	–	–	–	–
5	Link between concepts	Included in Investigable	–	–	Comparison (partially)
Dim4	**Verification type (optional)**				
1	Mistake / contradiction	–	Included in Elaboration	Included in Confirmation	–
2	Knowledge in course	Included in Non-investigable	–	Included in Confirmation	–
3	Exam	–	–	–	–

5. Automated Annotation

5.1. Automated Expert Rule-Based Annotation

5.1.1. Method

To answer RQ2 and to annotate the whole corpus (and in the long term, to use it online to analyze the questions collected), we developed a tool to identify the keywords, give them a weight, and segment the questions, as detailed below.

First, given a student question, we identify the keywords representative of each value in each dimension using string matching and a set of regular expressions expanding the keywords identified during the manual annotation (*cf.* Figure 2). The question is then automatically segmented based on a system for sentence boundary detection (Kiss & Strunk, 2006), which is one of the few systems that works for French. This system, which relies on an unsupervised approach, has been largely tested on different languages and on different text genres. It achieves good results without any further amendments or language-specific resources. Although, some student questions can be poorly written and incorrectly formulated, the segmentation method appeared to be working well enough in this context. It is also worth noting that in practice, when manually annotating

segments, neither human expert found a situation where they considered that the segment should have been segmented more than it was. Then for each segment of the question, for each dimension, we tag the segment in that dimension according to the value that has the most keywords associated to it (e.g., for dimension 1, a segment with two keywords associated to the value "re-explain" and one keyword associated to the value "validation" would be tagged as a "re-explain" question). Finally, the segment identified by the values associated to each dimension, is represented as a binary annotation vector (as seen in Figures 2 and 3).

An issue that can arise in this process is the existence of an ambiguity within a dimension (e.g., in the example in Figure 3, both categories dim3-1 and dim3-4 are annotated in the same question). To solve this issue, the automatic annotator uses a set of weights associated to each keyword of each dimension (e.g., explain: 7; what/how: 3), and defined using the set of 600 questions. Practically, those weights were determined in two steps: first, the expert annotators empirically associated a weight between 1 and 9 to each keyword, depending on whether they thought they were very marginally (1), significantly (5), or very significantly (9) associated to a given dimension. Second, the automatic annotator was used on the 600 manually annotated questions, and weights were manually adjusted (adding or removing 1) on some keywords for questions for which the manual and automatic annotations were different, iterating until full agreement was obtained on almost all segments from the 600 questions. In the example shown in Figure 2, the introduction of weights helps the annotator choose between "example" and "schema" for the value associated to dimension 2. The automatic annotation is thus sensitive to weight variations, nonetheless after the weight adjustments, only 1% of the questions were still considered ambiguous (*cf.* Figure 3, the words "definition" and "functions" had the same weight within dimension 3). Finally, we used the sample of 200 questions from the evaluation step (*cf.* section 4.1), which were not used for the weight adjustment phase, to evaluate the automatic annotator and to calculate kappa values per dimension.

Practically, we implemented the annotator using the Natural Language Toolkit (NLTK), a major library in Python to handle natural language data. In particular we used the "Punkt Sentence tokenizer" module for French to divide each sentence into a list of segments, and each segment into a list of words. This tokenizer, already trained for the French language, uses the aforementioned unsupervised algorithm to build a model for abbreviated words, collocations, words with characters, and to mark the beginning and the end of a sentence.

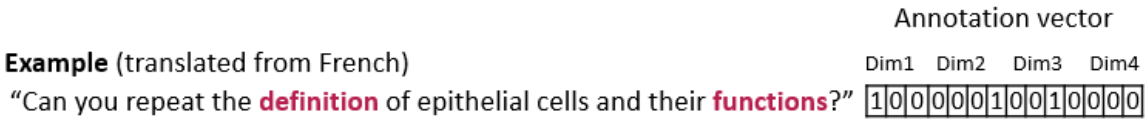

Figure 3: Example (translated) of question ambiguity of the automatic annotation

5.1.2. Results and discussion

The kappa values per dimension are given in Table 5 for the annotations coming from both expert annotators.

Table 5: Kappa values between automatic and manual annotation

Dimensions	Dim1	Dim2	Dim3	Dim4
Kappa between expert 1 & rule-based automatic annotator	0.66	0.50	0.63	0.55
Kappa between expert 2 & rule-based automatic annotator	0.76	0.69	0.70	0.65

We consider that the kappa values obtained (between 0.60 and 0.70) are high enough to apply the automatic annotator to the full corpus, even if this type of decision is always partially arbitrary (Artstein & Poesio, 2008). Moreover, those values include unannotated sentences (tagged as "0" by the annotator in all dimensions because they did not match any of the known patterns). For instance, without unannotated questions, the kappa values on dimension 1 are between 0.80 and 0.86, for both human annotators.

Overall, we have therefore positively answered RQ1 and RQ2: not only have we successfully defined a coding scheme of questions relevant for our goal, but we have been able to automatize the annotation of questions using this coding scheme, thus allowing us to automatically annotate all the sentences in the corpus (and potentially annotate future student questions).

5.2. Automated Machine Learning-Based Annotation

5.2.1. Method

Two steps were performed manually in the methodology described before: 1) keyword identification, and 2) associating a set of keywords to a value for each dimension. Therefore, we wanted to examine whether these steps could be fully automated as

well without degrading the performance of the system.

To automate keyword identification, we used an algorithm to extract keywords from the questions, with a "bag of words" type of approach. First, we preprocessed all questions in the corpus, using the classical steps: tokenization, stemming, punctuation removal. We then extracted all the unigrams and bigrams and counted their occurrences in the sample of 600 questions from the validation step. Then, each of the 723 segments was represented by a word vector ("1" if the word is in the segment, "0" otherwise). We finally reduced the number of keywords automatically extracted to keep the most important and significant ones using a feature selection technique (removing less frequent and correlated unigrams and bigrams). In this approach, no stop words were removed before the unigram and bigram extraction. Then if a unigram was removed, we kept the bigrams containing it. This decision is justified by the fact that some of the manually identified keywords relied on two words that could not be meaningful separately (e.g., "one" and "need" — in French "il" and "faut") and would hence be filtered out, although the associated bigram is highly relevant ("faut-il," which could be translated as "does one need to").

To automatically associate keywords to the values associated to each dimension (e.g., "re-explain"), we tried different machine learning algorithms on the segments represented as vectors annotated: 1) with the manually extracted keywords, and 2) with the set of automatically extracted keywords (*cf.* Figure 4). We tried four classification techniques (decision tree, K-NN, rule induction, and naïve Bayes) for each dimension separately. For each classifier, the input was the vectors representing the segments in terms of keywords, and the label to predict was the value associated to the segment in that dimension. We considered all the questions manually annotated from the validation and evaluation steps as labelled data. Then, we trained the classifiers on the 723 segments corresponding to 600 questions and evaluated their performance on an independent sample of 200 segments, to ensure a good estimation of the performance on unseen data. Finally, we calculated the kappa values between the values found by the classification model for that dimension, and the corresponding values found by the manual annotation. Three combinations were considered for comparison: 1) manual keyword extraction and expert rule-based annotation, 2) manual keyword extraction and ML-based annotation, 3) automatic keyword extraction and ML-based annotation. We did not consider the case of automatic keyword extraction with expert rule-based annotation, which would have required recoding a new expert rule-based annotator (as it heavily relies on the set of keywords).

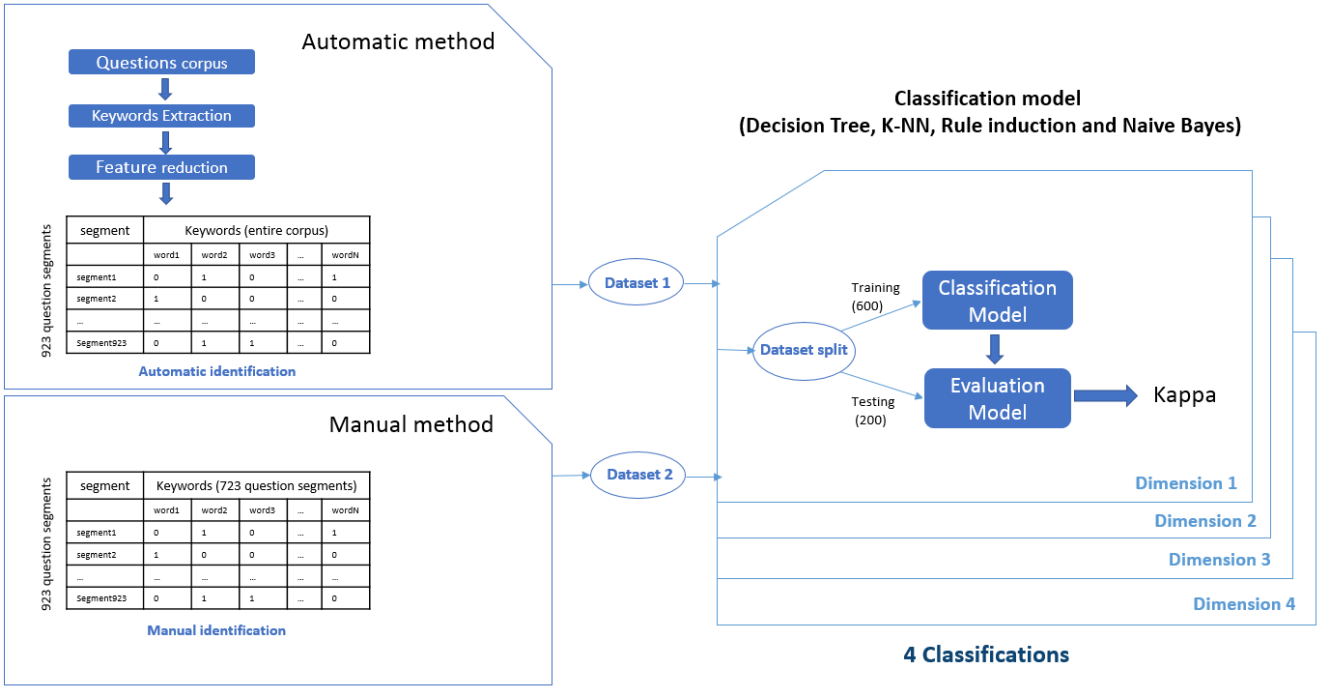

Figure 4: The automatic process to associate keywords to values in each dimension.

5.2.2. Results and discussion

The automatic keyword extraction method led us to a set of 268 keywords (to compare with 149 keywords selected by manual identification). Among these keywords, 50 were common to both datasets (e.g., "role," "why," "can we"), but the majority were found only in the automatic keywords list (e.g., "method," "also," "enzyme").

The kappa values found with the Decision Tree and K-NN algorithms (the two giving the best results) are provided in Table 6, for each dimension, and compared with the results obtained with the previous expert rule-based annotator.

Table 6: Kappa values between different combinations of annotation and keyword extraction methods with the reference manual annotation

Keyword extraction method	Annotation method	Dim1	Dim2	Dim3	Dim4
Manual	Expert rule-based annotator	0.76	0.69	0.70	0.65
Automatic	ML-based annotator (Decision Tree)	0.75	0.72	0.65	0.51
Manual	ML-based annotator (K-NN)	0.71	0.63	0.56	0.62

We note that the expert rule-based annotator outperforms both machine learning-based annotators on dimensions 1, 3, and 4, whereas the fully automatic annotator performed slightly better on dimension 2. The fully automatic annotator also performs better than the mixed one (manual keywords extraction + ML) for the four dimensions.

Based on these results, we chose to use the expert rule-based annotator for the rest of this paper because of its overall superior performance in our context. Although it is not important in our context, it is worth noting that due to the domain-dependent keywords identified (e.g., "enzyme"), one could question the capability of the fully automatic annotator to generalize well to questions similar in structure but from another (non-medical) domain. This is probably in part due to the modest size of the corpus of questions. We are therefore in one of the situations of NLP where the limited size of the data available still allows expert-based systems to outperform statistical approaches.

6. Identifying Links Between Questions and Behaviour

To address RQ3 — to know if the questions asked by a student can be informative of their characteristics — we needed to identify characteristics generic enough not to be specific to only one particular course. We decided to consider the four courses that had received the most questions: BCH, HBDD, BCE, and ANT (the latter being the only course to have received a large number of questions in the second semester). Before comparing different categories of students asking questions, we thought it could be relevant to first consider how students who asked questions differed from those who did not, in order to contrast elements characteristic of question-asking behaviour from those associated with the mere fact of asking questions or not.

This led us to split RQ3 into four questions:

RQ3.1: Is there a link between students asking questions and their learner characteristics?

RQ3.2: Are there questions that are characteristic of student performance?

RQ3.3: Are there temporal patterns of questions that are characteristic of student performance?

RQ3.4: Is there a link between the questions asked by students and their learner characteristics?

6.1. Comparing Students Who Asked Questions With Those Who Did Not

Method: For each of the four courses considered, we split the students between those who asked questions (group Q) for that course, and those who did not (group NQ). We then analyzed these two groups according to seven of the nine variables introduced in Table 2 (NbQst and NbVotRec are, by definition, different and thus irrelevant to consider here). For the two variables relative to grades (AvgGrd and FinGrd) and the two variables relative to attendance (GlbAtt and CouAtt), which are ratios, not all distributions followed a normal law ($p < .05$ in some cases when testing with Shapiro-Wilk), which led us to perform Mann-Whitney U tests instead of t-tests. We also used Mann-Whitney U tests for the variable relative to the votes done (NbVotDone), which is an ordinal variable. We report an estimated effect size, calculated as $r^2 = \eta^2 = \frac{Z^2}{n}$, where Z stands for the z-score associated to the test p-value, and n the number of students in that group (Fritz, Morris, & Richler, 2012). For the variable indicating whether the student was retaking the year (RetStu), and the one indicating whether the student was eventually accepted to move on to year 2 after the final exam (SucStu), which are categorical variables, the samples were large enough to consider using a Chi-square test instead of Fisher's test, using Yates' correction for continuity when one cell of the contingency table had a count inferior to 5, and report effect size using corrected Cramér's V (more conservative than the original), noted $\tilde{V}$ (Bergsma, 2013).

Results: In terms of grades, there was no statistically significant difference for AvgGrd in any of the four courses ($p > .05$). For FinGrd, only ANT had a statistically significant difference ($U = 66393$, $p = .020$, $\eta^2 = .003$), the other two courses with data available did not.

In terms of attendance, for GlbAtt there was only a statistically significant difference for BCE ($U = 54755$, $p < .001$, $\eta^2 = .024$). For CouAtt however, there were statistically significant results for all courses: BCH ($U = 213974$, $p < .001$, $\eta^2 = .035$), HBDD ($U = 71554$, $p < .001$, $\eta^2 = .097$), BEH ($U = 110005$, $p < .001$, $\eta^2 = .021$), and ANT ($U = 88238$, $p < .001$, $\eta^2 = .042$).

In terms of number votes done on other students' questions (NbVotDone), the results are statistically significant for all courses: BCH ($U = 270020$, $p < .001$, $\eta^2 = .246$), HBDD ($U = 225962$, $p < .001$, $\eta^2 = .197$), BCE ($U = 134731$, $p < .001$, $\eta^2 = .215$), and ANT ($U = 84866$, $p < .001$, $\eta^2 = .285$).

In terms of retaking students, the Chi-square test revealed a significant difference between the groups in BCH ($\chi^2(1) = 35.67$, $p < .001$, $\tilde{V} = .177$), HBDD ($\chi^2(1) = 23.25$, $p < .001$, $\tilde{V} = .144$), and BCE ($\chi^2(1) = 9.11$, $p < .010$, $\tilde{V} = .088$), but not for ANT. In terms of successful students, the Chi-square test revealed a significant difference between the two groups in all courses: BCH ($\chi^2(1) = 25.21$, $p < .001$, $\tilde{V} = .122$), HBDD ($\chi^2(1) = 29.05$, $p < .001$, $\tilde{V} = .132$), BCE ($\chi^2(1) = 49.72$, $p < .001$, $\tilde{V} = .175$), and ANT ($\chi^2(1) = 86.96$, $p < .001$, $\tilde{V} = .232$).

Table 7 provides a summary of the descriptive statistics for both groups in each course (proportion for RetStu and SucStu, median, 1st and 3rd quartiles for other variables because of the non-normal distribution), with indication of statistically significant results.

Table 7: Summary of median, 1st and 3rd quartiles of dependent variables (proportion of Retstu and Sucstu) for each group and each course

			AvgGrd			FinGrd			GlbAtt			CouAtt			RetStu	NbVotDone			SucStu
	Grp	**N**	**Q1**	**Md**	**Q3**	**Q1**	**Md**	**Q3**	**Q1**	**Md**	**Q3**	**Q1**	**Md**	**Q3**	**P**	**Q1**	**Md**	**Q3**	**P**
BCH	Q	244	6.67	8.83	11.83	5.5	8.5	12	0.90	0.98	1	1	1**	1	21%**	1	2**	9	22%**
	NQ	1372	5.50	8	10.83	3.75	6.75	10.25	0.76	0.95	1	0.50	1**	1	42%**	0	0**	0	11%**
HBDD	Q	201	7.75	10.5	14	6.75	9.75	12.5	0.93	0.98	1	1	1**	1	23%**	0	3**	7	24%**
	NQ	1410	6	9.25	12.25	4	7.75	10.75	0.76	0.95	1	0.50	1**	1	42%**	0	0**	0	11%**
BCE	Q	114	7.25	10	12.5	N/A	N/A	N/A	0.86	0.98**	1	1	1**	1	26%**	0	1**	4	33%**
	NQ	1486	5	7.75	10.75	N/A	N/A	N/A	0.76	0.95**	1	0.40	1**	1	41%**	0	0**	0	11%
ANT	Q	75	8.6	12	14.6	9.5	13.5**	15.5	0.98	1	1	1	1**	1	40%	0	0**	3	47%
	NQ	1528	5	8	11.80	3.50	7.25**	12.19	0.76	0.95	1.00	0	0.60**	1	40%	0	0**	0	11%

* $p<.01$; ** $p<.001$ — in bold, the highest value between NQ and Q; N/A = no data available for the course.

Discussion: Overall, we see that students who asked questions tended to have higher final grades, but that was significant only for the second-semester course (ANT). Considering the difference becomes significant only in the second semester, we could hypothesize (which would have to be verified) that the differences between students who ask questions and those who do not become stronger over time. Asking questions, however, was not associated with student performance in the MCQ session during the class. Asking questions was associated to a lower global attendance only in BCE, indicating possibly that this particular course was more attractive to the average student, as there is no clear pattern otherwise. Similarly, asking questions was logically associated to a higher attendance in the course (as the answers are provided during it), except for HBDD where the pattern is reversed for a reason we could not identify. Interestingly, students who voted the most were also the ones asking questions. One of the limits of this result is that we lack log data to distinguish among students who did not ask any questions those who actually spent time on the online platform from those who did not (we know all students connected and at least loaded the question page because it is a prerequisite for them to be able to see their grades). Thus, we cannot verify whether there is a subgroup of "lurkers" who may only be voting and never asking questions by themselves. We also observe a higher proportion of retaking students among those not asking questions, at least in the first semester, possibly because they may consider it less relevant to them. Finally, students asking questions consistently succeed better at the end of the year — although we cannot say whether excellent students naturally ask more questions or whether the fact of asking questions leads to better results. However, the coding scheme previously established offers a way to analyze in a finer-grained manner the students in the Q group, to see if we can distinguish them by the types of questions they ask.

6.2. Are the Proportions of Question Types Characteristic of Student Performance?

Method: One of the most obvious characteristics is the student's level, as indicated by their grade in the final exam at the end of the year (FinGrd). We decided to distinguish two types of students: the good students (whose rank in the final exam is less than 200th — which corresponds to the students allowed to move up to the second year of PACES) and the average ones (whose rank is between 200th and 600th). The students with a rank higher than 600th (the low-level students with poor grades) generally did not ask enough questions to be considered here. Moreover, teachers usually consider that students beyond 600th do not really have a chance to pass the final exam, even with their assistance, and they generally want to focus more on average students. Our hypothesis relative to RQ3.2 was that good and average students would be asking significantly different

questions.

Results: Unfortunately, an exploration of the data for each dimension for all 13 courses did not reveal any clear trend (*cf.* Figure 5 for the results over 13 courses — the results were extremely similar for the four main courses [BCE, BCH, HBDD, ANT]), which quickly led us to reply negatively to RQ3.2 without the need for advanced statistical comparisons between the two populations.

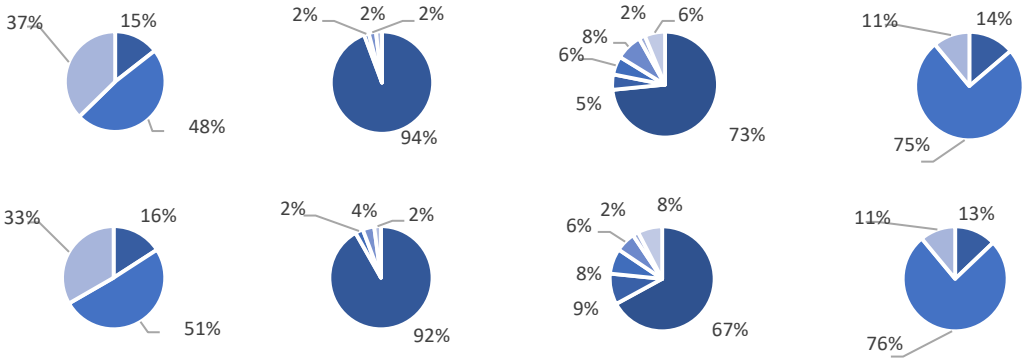

Figure 5: Proportions of questions on the four dimensions across all the 13 courses; (top row = good students, bottom row = average students) (colours: from dark blue [0 or 1] to light blue [the maximum value for that dimension]).

6.3. Are the Dynamics of Questions Asked Indicative of Student Performance?

Method: Our second hypothesis was that the dynamics of questions asked by the students could be indicative of their level. For instance, we could assume that good students would tend to ask more complex questions towards the end of the course, because they have already acquired a solid understanding of the basic concepts during the first weeks. To investigate the RQ3.3, we started by a visual data exploration, comparing the proportion of questions asked by students each week in each dimension. An example of such a visualization is shown in Figure 6. In particular, we were looking for opposite patterns between good and average students. In our previous example, we could hypothesize that, contrary to good students, the proportion of complex questions asked by average students would not increase towards the last weeks. Table 8 presents a systematic overview of the difference in trends between the categories of students (only dimensions showing differences in trends between good and average students were reported here).

Results: Unfortunately, the results showed that even if some dimensions seemed to be important in distinguishing good from average students (for instance dimension 4–3, i.e., verification question about the exams), and that there may exist ways to distinguish good from average students based on the type of questions they asked (e.g., an increase in questions about a deeper explanation for good students in ANT, whereas a decrease was observed for average students), the trends were different across the courses. Therefore, the answer to our RQ3.3 was that such patterns may exist, and it is then necessary to consider not only the raw proportion of questions of each type, but also their dynamics across the course. However, this dynamic seems to be very course-related, which led us to perform our following analyses separately on each course, instead of trying to group questions across the courses.

6.4. Are Student Characteristics Related to the Questions Asked?

Our third and final hypothesis was that student questions may indicate something subtler than their mere level (good versus average), such as whether they were regularly missing courses, or whether they were taking this course for the second time. Moreover, a combination of questions and not only a single dimension may need to be considered. To answer RQ3.4, we performed an analysis in two steps, similar to [3], where we started by extracting clusters of students, and then tried to identify the characteristics distinguishing each of them.

6.4.1. Clustering students based on their questions

Method: First, we performed four separate clustering analyses using the k-means algorithm (with k between 2 and 10) and the agglomerative hierarchical clustering algorithm over four datasets: students who asked questions in the BCH (1227 questions by $N_1 = 244$ students), HBDD (979 questions by $N_2 = 201$ students), BCE (685 questions by $N_3 = 114$ students), and ANT courses (649 questions by $N_4 = 75$ students). We performed the clustering using as features for each student the proportion of questions asked in each dimension (e.g., the proportion of questions with value 1 in dimension 1) asked (a) overall, (b)

during the first half of the course and (c) during the second half of the course. We also considered the overall proportion of questions asked in the first half and second half of the course (44 features overall as mentioned in Figure 7). Distinguishing (b) and (c), in addition to (a), allowed us to take into account the dynamics of questions (as suggested by our answer to RQ3.3), in addition to the overall distribution.

Table 8: The evolution of students along the weeks of each course/dimension

Course	Dim	Good	Average	Opposite trends
ANT	1–2	Increase	Decrease	1 week
5 weeks	2–2	Decrease/increase	Stable/decrease	2 weeks
	3–5	Increase	Decrease/stable	2 weeks
	4–3	Stable	Increase	1 week
BCE	2–2	Decrease/increase	Increase/stable	2 weeks
5 weeks	3–5	Increase/decrease	Decrease/increase	2 weeks
	4–3	Decrease/increase	Increase/decrease	2 weeks
BCH	1–2	Increase/decrease	Decrease/increase	4 weeks
6 weeks	1–3	Increase/decrease	Decrease/increase	6 weeks
	2–2	Decrease/increase	Increase/decrease	4 weeks
	3–5	Increase/decrease	Decrease/increase	2 weeks
	4–1	Decrease/increase	Increase/decrease	6 weeks
	4–3	Increase/decrease	Decrease/increase	5 weeks
	4–2	Increase/decrease	Decrease/increase	6 weeks
HBDD	3–5	Increase/decrease	Decrease/increase	3 weeks
4 weeks	4–3	Increase/decrease	Decrease	1 week

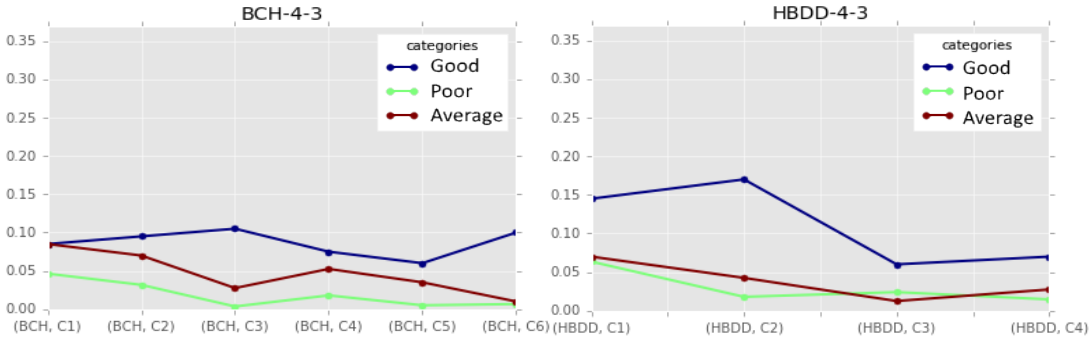

Figure 6: Pattern difference of good vs. average students on dim. 4–3 for BCH (left)/HBDD (right) (X-axis: questions weeks of each course; Y-axis: proportions of questions asked).

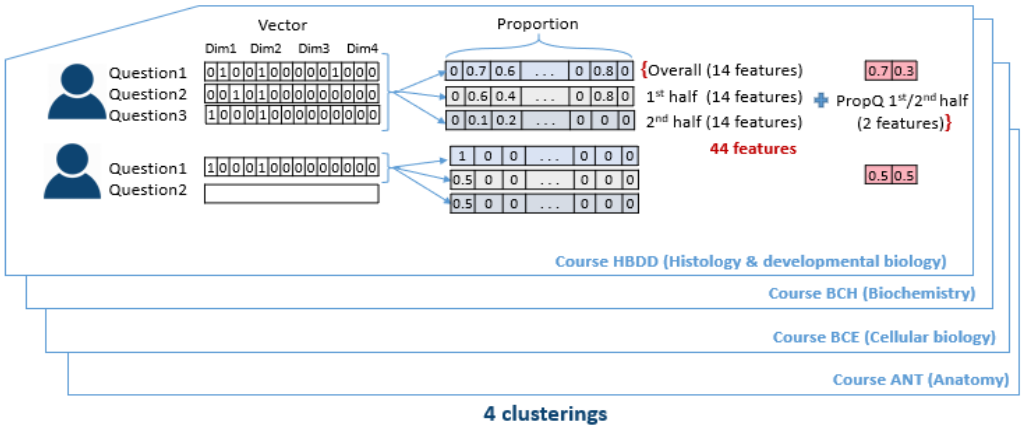

Figure 7: Student clustering features.

We have also tried to reduce the number of features for clustering using a feature selection technique (removing useless and correlated features) to show up most important and discriminant features for all courses. Although a Principal Component Analysis could also lead to that result, the difficulty of interpreting the dimensions extracted is an issue when the aim is to present them to teachers — moreover, when trying that approach, the cluster quality was not improved. Because of the differences across courses, we performed this on each course individually (e.g., the feature prop_d1-1 is more important for BCH than BCE especially in the first half of the course — *cf.* Table 13 in the appendix). We compared the quality of the various clustering results found using the silhouette coefficient, which is a measure of consistency within the clusters (between –1 and 1).

Results: For k-means, we obtained four clusters for BCH and HBDD and three clusters for BCE and ANT, which centroids are provided in Table 13 in the appendix. The number of clusters found was similar to the agglomerative hierarchical clustering approach by looking for gaps in the dendrograms. However, the quality of clusters obtained with hierarchical clustering was lower than the one obtained with k-means (silhouette coefficients varied between .12 and .19 for hierarchical clustering vs .19 and .23 for k-means). Although feature selection gives new clusters and characteristics across the courses, the silhouette coefficients are not better on average either. Due to the lack of better-defined clusters and the additional difficulty of reporting different dimensions and student characteristics for each course, we therefore focus in the remainder of this paper on the clusterings obtained with k-means using 44 features, without feature reduction.

6.4.2. Cluster characterization

Method: The second step consisted of characterizing the clusters by considering the student attributes (not used for clustering) mentioned in section 3, similarly to the methodology followed in section 6.1. The characterization of clusters was done manually because we were not certain to find similar characteristics across the courses. For the two variables relative to grades (AvgGrd and FinGrd) and the two variables relative to attendance (GlbAtt and CouAtt), which are ratios, not all distributions followed a normal law ($p < .05$ in some cases when testing with Shapiro-Wilk), leading us to perform Kruskal-Wallis H tests on ranks instead of one-way ANOVAs for the clusterings associated to the four courses considered. We also used Kruskal-Wallis tests for the variables relative to the questions (NbQst) and votes (NbVotRec and NbVotDone), which are ordinal variables. When a result was found to be statistically significant, we ran follow-up comparisons using Dunn's test with Holm-Bonferroni correction. We also report effect size for the Kruskal-Wallis H tests, using Cohen's formula for eta-squared measure (Cohen, 1988), i.e., $\eta_H^2 = \frac{H-k+1}{n-k}$, where k stands for the number of clusters and n the number of students in that cluster. For the two other remaining variables (RetStu and SucStu), which are categorical, we performed a Chi-square test in conditions similar to those described in section 6.1. Cluster naming has been chosen *a posteriori* to match the similar characteristics identified further.

Results: In terms of grades, for BCH there was a statistically significant difference between clusters for the average grade ($\chi^2(3) = 17.20$, $p < .001$, $\eta^2 = .059$) and for the final grade ($\chi^2(3) = 20.71$, $p < .001$, $\eta^2 = .074$). This was also the case for HBDD (AvgGrd: $\chi^2(3) = 25.11$, $p < .001$, $\eta^2 = .112$, and FinGrd: $\chi^2(3) = 28.95$, $p < .001$, $\eta^2 = .132$). For BCE, there was no statistically significant difference between clusters for AvgGrd and no FinGrd data available. For ANT, there was a statistically significant difference between clusters for the average grade ($\chi^2(2) = 7.84$, $p = .020$, $\eta^2 = .081$) and for the final grade ($\chi^2(2) = 13.50$, $p = .001$, $\eta^2 = .160$).

In terms of attendance, for BCH there was a statistically significant difference for CouAtt ($\chi^2(3) = 10.51$, $p = .015$, $\eta^2 = .031$), but not for GlbAtt. For HBDD, the test showed a statistically significant difference for GlbAtt ($\chi^2(3) = 9.33$, $p = .025$, $\eta^2 = .041$) but not for CouAtt. For BCE, there was a significant difference for CouAtt ($\chi^2(2) = 13.69$, $p = .001$, $\eta^2 = .105$), but not for GlbAtt. For ANT, there was a statistically significant difference for GlbAtt ($\chi^2(2) = 6.19$, $p = .045$, $\eta^2 = .066$) and CouAtt ($\chi^2(2) = 12.04$, $p = .002$, $\eta^2 = .139$).

In terms of number of questions/votes, for BCH there was a statistically significant difference for NbQst ($\chi^2(3) = 42.12$, $p < .001$, $\eta^2 = .163$) and NbVotRec ($\chi^2(3) = 12.06$, $p = .038$, $\eta^2 = .037$), but not for NbVotDone. For HBDD, the differences were also statistically significant for NbQst ($\chi^2(3) = 33.20$, $p < .001$, $\eta^2 = .153$) and NbVotRec ($\chi^2(3) = 16.76$, $p < .001$, $\eta^2 = .070$), but not for NbVotDone. For BCE, a difference for NbQst ($\chi^2(2) = 9.85$, $p = .007$, $\eta^2 = .071$) but none for NbVotRec and NbVoteDone. For ANT, no statistically significant differences were found for NbQst, NbVotRec, and NbVotDone.

In terms of retaking students, the Chi-square test revealed a significant difference between the clusters in BCH ($\chi^2(3) = 14.43$, $p = .002$, $\tilde{V} = .217$), HBDD ($\chi^2(3) = 23.72$, $p < .001$, $\tilde{V} = .322$), and no difference in BCE and ANT. In terms of successful students, the Chi-square test revealed a significant difference between the clusters in BCH ($\chi^2(3) = 21.47$, $p < .001$, $\tilde{V} = .276$), HBDD ($\chi^2(3) = 19.14$, $p < .001$, $\tilde{V} = .284$), and no difference in BCE and ANT.

Table 9 provides a summary of the descriptive statistics (proportion for RetStu and SucStu, median, 1st and 3rd quartiles for other variables) for each cluster in each course, and Table 10 provides a summary of the aforementioned inferential statistics results, according to follow-up post hoc comparisons.

Table 9: Summary of median, 1st and 3rd quartiles of dependent variables (proportion of Retstu and Sucstu) for each cluster and each course

	Clu.	N	RetStu		SucStu		AvgGrd			CouAtt			FinGrd			GlbAtt			NbQst			NbVotDone			NbVotRec		
			P		P		Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3	Q1	Md	Q3
BCH	A	44	14%		14%		5.67	7.8	10.25	0.83	1.0	1.0	4.12	6.5	10.5	0.74	0.95	1.0	1.0	1.0	2.0	1.0	3.0	7.5	0.25	2.0	4.38
	B	89	22%		27%		7.0	9.42	12.33	1.0	1.0	1.0	6.5	8.88	12.38	0.93	1.0	1.0	2.0	4.0	9.0	0.0	3.0	10.25	0.5	1.83	3.76
	C	77	14%		10%		6.5	8.45	10.67	0.83	1.0	1.0	5.5	8.0	10.5	0.91	0.98	1.0	1.0	2.0	4.0	1.0	2.0	7.0	0.67	2.0	3.65
	D	34	44%		47%		8.04	10.75	13.13	1.0	1.0	1.0	8.0	12.0	14.25	0.9	0.98	1.0	2.0	5.5	13.25	0.0	1.0	3.5	0.0	0.86	1.42
HBDD	A	59	7%		7%		7.0	8.25	10.88	1.0	1.0	1.0	5.0	7.38	10.31	0.89	0.95	0.98	1.0	1.0	3.0	1.0	3.0	6.0	1.0	3.0	8.0
	B	74	27%		32%		8.75	12.25	14.69	1.0	1.0	1.0	8.0	11.0	12.75	0.93	1.0	1.0	2.0	3.0	9.75	0.25	3.0	7.5	1.0	2.89	5.0
	C	31	16%		16%		7.88	10.25	12.54	1.0	1.0	1.0	7.75	10.0	12.75	0.95	1.0	1.0	1.0	2.0	3.0	0.0	1.0	7.0	0.42	1.6	3.5
	D	37	49%		41%		9.65	13.38	15.56	1.0	1.0	1.0	8.25	12.0	14.12	0.95	1.0	1.0	2.0	5.0	7.0	0.0	3.0	8.0	0.0	1.0	2.17
BCE	A	26	15%		19%		6.25	8.6	10.75	0.8	1.0	1.0	N/A	N/A	N/A	0.91	0.98	1.0	1.0	4.0	9.25	0.0	1.5	7.0	0.49	1.02	3.19
	B	52	31%		31%		7.19	10.3	12.8	1.0	1.0	1.0	N/A	N/A	N/A	0.44	0.94	1.0	1.0	1.5	2.0	0.0	1.0	2.0	0.0	1.5	3.2
	D	36	28%		47%		8.55	11.2	12.6	1.0	1.0	1.0	N/A	N/A	N/A	0.93	1.0	1.0	1.0	2.5	7.5	0.0	1.5	3.0	0.38	1.0	2.0
ANT	A	15	20%		27%		4.29	6.0	13.45	0.8	1.0	1.0	6.25	9.0	12.5	0.91	0.95	1.0	1.0	2.0	3.0	0.0	0.0	3.5	0.0	0.5	1.0
	B	30	43%		50%		8.2	11.4	14.0	1.0	1.0	1.0	11.5	14.5	15.5	0.98	1.0	1.0	3.0	4.5	9.0	0.0	0.0	2.0	0.14	0.53	0.95
	D	30	47%		53%		11.2	12.7	14.9	1.0	1.0	1.0	12.5	14.0	16.38	0.98	1.0	1.0	2.25	5.5	11.5	0.0	1.0	3.0	0.0	0.29	0.59

N/A: No data available for the course

Table 10: Pairwise differences for AvgGrd, FinGrd, GlbAtt, CouAtt, NbQst, NbVotRec, NbVotDone

Clust.	Clust.	AvgGrd			FinGrd				GlbAtt	CouAtt				NbQst			NbVotRec	
		BCH	HBDD	ANT	BCH	HBDD	BCE	ANT	HBDD	ANT	BCH	BCE	ANT	BCH	HBDD	BCE	BCH	HBDD
A	B	–	***	–	*	***	N/A[1]	*	*	–	–	*	–	***	***	**	–	–
A	C	–	–	N/A[2]	–	***	N/A[1]	N/A[2]	–	–	–	N/A[2]	N/A[2]	*	–	N/A[2]	–	–
A	D	**	***	*	***	*	N/A[1]	***	*	–	*	***	**	***	***	–	*	***
B	C	–	–	N/A[2]	–	–	N/A[1]	N/A[2]	–	–	–	N/A[2]	N/A[2]	***	**	N/A[2]	–	–
B	D	–	–	–	–	–	N/A[1]	–	–	–	–	–	*	–	–	–	**	*
C	D	**	–	N/A[2]	**	–	N/A[1]	N/A[2]	–	–	–	N/A[2]	N/A[2]	**	**	N/A[2]	*	–

6.5. Discussion

Figure 8 presents a synthesis of the results, described in more detail in this section. Cluster A represents 18% to 29% of students and is characterized by grades lower than average (both in that course and overall), and lower course and global attendance — sometimes significantly lower — than that of students in other clusters. The students in this cluster are mostly (86% on average) taking the course for the first time and they succeed less (17% on average, which almost corresponds to the proportion of retaking students). They asked fewer questions than average, but their questions received more votes than average and were therefore fairly popular. Their questions are mainly about re-explanation or deepening a concept, specifically requests for definition (dim1-1, dim1-2, and dim3-1) and often asked during the first half of the course. This cluster therefore corresponds to struggling passive students who need basic re-explaining, which many students could benefit from.

	Struggling passive	Nitpicking active	Actively understanding
	Cluster A	**Cluster D**	**Cluster B**
#students	18-29%	14-25%	36-40%
Grades	lower	higher	higher
Attendance	low	high	high
#questions asked	low	high	high
#votes	popular	unpopular	popular
% Retaking students	low	high (42%)	average (31%)
% Successful students	low	high (47%)	average (35%)
Questions type	Re-explanation & definition	Verification: mistake or contradiction	Verification of knowledge & link between concepts
Majority of questions asked	1st half	2nd half	1st half

Figure 8: Summary of variables similar/different for three similar clusters across the courses.

At the other end of the spectrum, cluster D represents 14% to 25% of students in the first semester (and 40% in the second semester, an increased proportion that can be linked to a decrease in activity from other students). Cluster D is characterized by course and final grades significantly higher than others (and almost always statistically significantly higher than students from cluster A) and attendance at most classes. Another distinctive characteristic of this cluster is the fact it contains an important proportion of students taking the course for the second time (42% on average) and succeeding more (47% on average). They ask more questions than average but their questions are less popular, with fewer votes overall. We may assume that these are very precise questions that require a good understanding of the content of the course and are thus not deemed as important by other students. Indeed, they mainly asked questions of verification to check for a mistake or contradiction in the course (dim1-3 and dim4-1) and fewer questions about the link between concepts (dim3-5). Interestingly, when comparing the proportion of questions asked in the first versus the second half of the class, they are the only students who asked more questions in the second half, presumably because the concepts presented at the beginning were easier to understand. This cluster therefore corresponds to *nitpicking active students*, who understand well the basics and point out potential mistakes on advanced concepts — sometimes even to confuse others, according to the pedagogical team (because of the competitive nature of the courses).

Cluster B represents 36% to 40% of students, whose grades, attendance, and number of questions asked are similar to cluster D. However, the proportion of students taking the course for the second time and the proportion of students allowed to move up to the second year is lower than D (31% and 35% on average respectively). They voted more and their questions were more popular. Overall, their questions are mostly on verification of knowledge and the links between concepts (dim1-3, dim4-2, and dim3-5), but only in the first half of the class. They correspond to students who are in the process of *actively building* their understanding of the course.

Cluster C (only found in the BCH and HBDD courses) tends to be an intermediate cluster lying between clusters A and D.

When comparing those clusters to students who did not ask questions (group NQ in 6.1), we see that it is the closest to cluster A in terms of grades. This group of students not asking questions is also characterized by the lowest proportion of successful students (11% on average), showing that no matter the type of question asked, it seems to be always better than not

asking questions. They also have the lowest proportion of course attendance, which can be interpreted in one of two ways: (a) they do not have any questions because the material is easy for them and they consider the Q&A session unnecessary, or (b) not having asked a question themselves, they are not motivated to attend and therefore miss an opportunity to improve their understanding. A complementary analysis revealed that 90% of the students in this group did not ask a single question in any course. This, combined with their overall lower performance and the fact that they vote less, seems to give more weight to interpretation (b).

Therefore, overall, we can say that our clustering (based on question dimensions only) and the follow-up inferential statistical analyses (using student characteristics) allow us to positively answer RQ3.4: the proportion and dynamics of questions asked by students is strongly linked to some of their characteristics (their performance, their attendance, the number of questions they asked, the number of votes they received, the number of times they voted on others' questions, whether they were following this course for the first time, and if they succeeded). Moreover, although the four clusterings were performed separately, and contrary to our previous analyses, the fact that three very similar clusters systematically appeared in the four courses considered suggests that the relationships identified are not dependent on the courses themselves.

7. Towards Predicting Student Profiles Online

We showed in the previous section that we were able to characterize students from their questions using clustering. Our goal is now to try to answer RQ4, which is to know if we can use the previous analyses to move towards prediction. This led us to investigate two options: first, performing another clustering on new data, but it was unlikely that we would obtain similar results when using only partial data in an online context. The second option was to apply the partition of the space learned on the year N (2012–2013 in the previous section) to the year N+1. The validity of this approach relies on the assumption that the nature of the questions asked in a given course does not vary much from one year to the next. To validate this hypothesis, we first compared the clusters of year N with the clusters of year N+1 (cf. Figure 9 — in blue and green). Then, we compared the application of the clustering of year N to the data of year N+1 (cf. Figure 9 — in orange), to the clustering performed on N+1 (which cannot be obtained before the end of the course, cf. Figure 9 — in green). The process of creating a predictive model through replicating the previous analyses on the next iterations of the course is summarized in Figure 9. Concretely, when using the k-means algorithm, applying clustering as a predictive model simply consists in classifying any new data point (here, a student) to a cluster based on the closest centroid (here using the same Euclidean distance used to create the clusterings originally). In other words, if a student question profile in year N+1 is closer to the centroid of cluster A in year N than to the centroids of clusters B, C, and D, then this student will be classified as part of cluster A.

More specifically, our goal is to answer two questions:

RQ4.1: When performing two separate clustering exercises on two separate datasets from two different years, is the intrinsic quality of clustering similar or does it vary? In other words, is it necessary to see if applying the predictive model of clustering from year N to the data from year N + 1 gives clusters of comparable quality to clustering done directly on year N + 1 (*cf.* Figure 9)?

RQ4.2: Are the clusters of student characteristics similar from one year to the next?

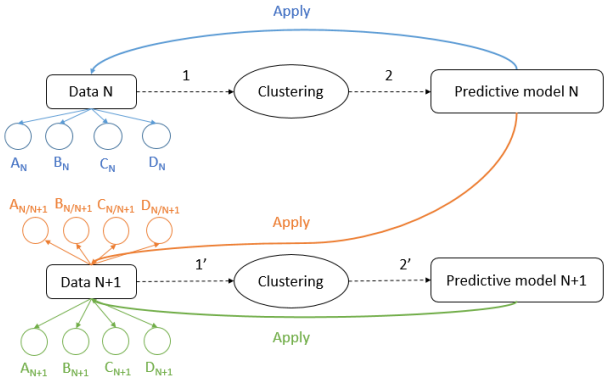

Figure 9: Clustering replication on next iterations.

7.1. What is the Difference in the Intrinsic Quality of Clusters across Different Years?

To answer this question, we used the silhouette coefficient to compare the quality of three clusterings: the previously computed clustering made on data from 2012 (*cf.* Figure 9 — arrow 1), a clustering obtained in a similar manner with k-means on data from 2013 (*cf.* Figure 9 — arrow 1'), and finally the data points labelled by the application of the predictive model of 2012

(*cf.* Figure 9 — in orange) on data from 2013. We computed the silhouette coefficient for each cluster and an overall average coefficient for each clustering (by construction, also a value between –1 and 1), where a high value indicates that across the clusters overall, students are closer to the cluster they have been assigned to than they are to the other clusters. The results of the silhouette coefficient for each cluster and each course for the three data partitions are given in Table 11.

The results of the silhouette coefficient showed that the predictive model of 2012 applied to the 2013 question dataset (in orange) and the clustering made directly on the 2013 dataset (in green) are similar in terms of cluster quality (but are of course different in terms of clusters obtained, although a comparison for each course revealed some similarities). Therefore, it seems that in our context using the predictive model resulting from the clustering performed over a given year to the data from the next year is as efficient, in terms of cluster quality, as performing clustering directly on the next year's dataset. This result is necessary but not sufficient, however, as we do not yet know whether the characteristics of those clusters (in terms of grades, attendance, etc.) are identical, which is mandatory if we want to reliably provide more information to the instructors about students asking those questions.

Table 11: Summary of silhouette coefficient for each cluster and each course of each clustering

Course	Clusters	Predictive model 2012 on data from 2013	Clustering 2012	Clustering 2013
BCH	Cluster_0	0.19	0.18	0.16
	Cluster_1	0.20	0.20	0.29
	Cluster_2	0.17	0.31	0.08
	Cluster_3	0.13	0.11	0.16
	Avg. coefficient	0.17	0.19	0.15
HBDD	Cluster_0	0.06	0.06	0.19
	Cluster_1	0.28	0.37	0.20
	Cluster_2	0.27	0.15	0.12
	Cluster_3	0.25	0.23	0.17
	Avg. coefficient	0.21	0.23	0.21
BCE	Cluster_0	0.21	0.27	0.12
	Cluster_1	0.12	0.19	0.15
	Cluster_2	0.15	0.17	0.09
	Cluster_3	–	–	0.26
	Avg. coefficient	0.19	0.22	0.20
ANT	Cluster_0	0.08	0.18	0.10
	Cluster_1	0.15	0.18	0.20
	Cluster_2	0.29	0.21	0.10
	Cluster_3	–	–	0.26
	Avg. coefficient	0.16	0.19	0.19

The results of the silhouette coefficient showed that the predictive model of 2012 applied to the 2013 question dataset (in orange) and the clustering made directly on the 2013 dataset (in green) are similar in terms of cluster quality (but are of course different in terms of clusters obtained, although a comparison for each course revealed some similarities). Therefore, it seems that in our context using the predictive model resulting from the clustering performed over a given year to the data from the next year is as efficient, in terms of cluster quality, as performing clustering directly on the next year's dataset. This result is necessary but not sufficient, however, as we do not yet know whether the characteristics of those clusters (in terms of grades, attendance, etc.) are identical, which is mandatory if we want to reliably provide more information to the instructors about students asking those questions.

7.2. Do Student Characteristics Vary from Year to Year?

To answer this question, we characterized the clusters of the 2012 model on data from 2013 according to the nine variables used for 2012 clustering. Then, we tried to see if there was a match between the clusters obtained on the 2013 dataset (in orange) with those identified on the 2012 dataset.

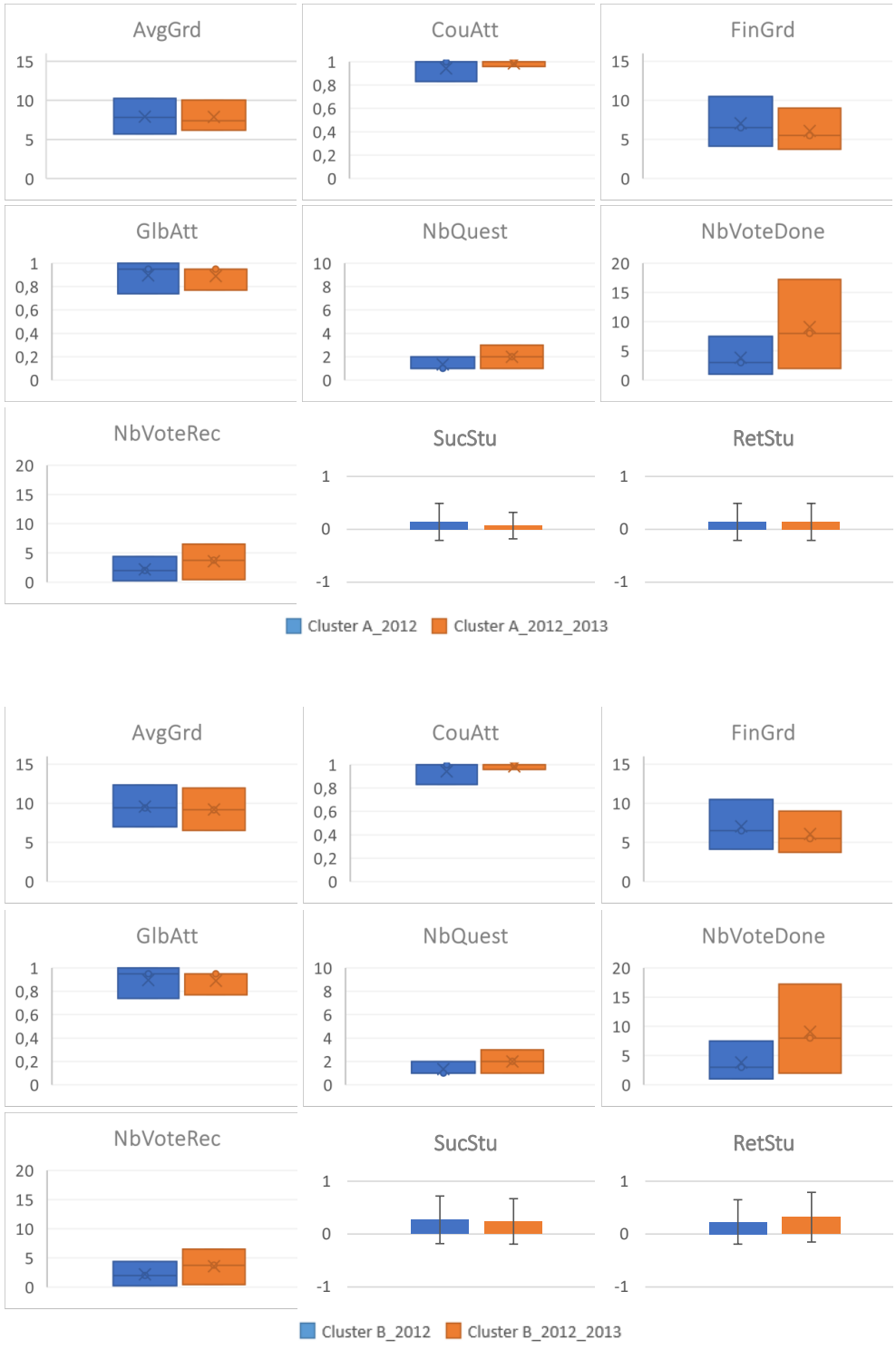

Figure 10: Comparison of 2012 and 2013 clusters in terms of median (centre of the bar), 1st quartile (bottom of the bar), and 3rd quartile (top of the bar) of dependent variables and mean and standard deviation of RetStu and SucStu for each cluster of BCH (see HBDD, BCE, and ANT in appendix).

The results show that the clusters obtained on the 2013 data matched those on the 2012 data quite well. The students identified in clusters A, B, and D on 2013 data had similar characteristics to those identified on 2012 data (in terms of performance, attendance, number of questions, and number of votes). Besides these similarities, the students of 2013 and 2012

share the same patterns of questions, i.e., the nature of questions asked by students in each cluster remains the same (even though centroids may vary). We have therefore answered RQ.4.1 and RQ4.2 positively.

7.3. Discussion

The results show that the predictive model is stable from one year to the next, even if the underlying clusters are not extremely distinguishable (with an average silhouette coefficient between 0.1 and 0.2). But what is particularly important for our goal is that the nature of questions asked by students and their profile characteristics remain the same, which means that we can characterize students from their questions (predict student profiles) using analyses from previous years. This confirms that our clusters can be used for predictive purposes. This should allow us to provide teachers with additional information about the questions they receive (even at the beginning of the year when not much data about the students asking them is available), based on the similarities to questions asked by students in previous years; for instance, by letting teachers know that a particular set of questions corresponds to those asked in the previous year by students who were regularly attending class and eventually succeeding.

8. Conclusions

We have developed a coding scheme for questions asked by students in a blended learning environment and reported the similarities between the existing taxonomies and our own. Moreover, we have shown how the automatic annotation of questions allowed us to identify student profiles, using exclusively the proportion of type of questions asked and their evolution over time. Two extreme clusters (lower than average students with popular questions, and higher than average students with unpopular questions) always appeared, with sometimes an intermediary cluster (higher than average students with popular questions). We also characterized students who did not ask questions (low attendance, less successful, and less voting) and compared them to students who asked questions (high attendance, more successful, and more voting). In addition, we have presented an approach that will allow us to predict online student profiles from their questions using a predictive model derived from clustering performed on data from the year before. This last result opens the possibility of helping teachers by providing them with more information about the students asking questions online, thanks to similarities to students who asked the same types of questions the year before.

A limit of this work lies in the fact that automatic annotation depends on weights that are empirically chosen and manually attributed to keywords. It is likely that with larger and more diverse datasets of questions, the automatic approach would outperform the manual approach proposed here, which could have an impact on the quality of clusters. The fact that we used a single source of data (even if we considered several courses across different years) could also raise questions regarding the applicability of our work in different contexts. Another limit lies in the quality of the clusters: it is likely that some students who are far from the centroid of their cluster could be associated just as well with another cluster. This distance could be quantified and translated in terms of reliability, to be able to inform the instructor that when they are told that the student asking a particular question has characteristics consistent with cluster A, this information is very reliable or not. This reliability metric would also depend on the number of questions asked by a student so far.

Overall, this work not only proves the relevance of the proposed coding scheme, but also validates the generality of this approach to identify student level online during the year, and to check whether they improve or not, based on the type of questions they ask in new courses. Moreover, from the teachers' perspective, automatic annotation could be helpful in organizing the questions they answer during their weekly Q&A session. In particular, it could provide them with additional information besides the mere votes in order to make sure they answer questions that can benefit different categories of students, not only the popular questions but also the precise (less popular) questions asked by the "good" students. Another interesting extension of this research would be to study the dynamics of questions in following the evolution of student questions in each course, but also across the courses. This could help students track the evolution of their questioning behaviour, and help teachers detect students struggling or getting better based on the type of questions they ask.

For the research community, we believe our coding scheme should be fairly easily reusable. First, it is language-independent: only the annotation relies on a set of keywords that would have to be translated for automation in another language. Second, it is domain-independent, as it does not rely on information about medical studies. Third, we believe it to be easy to use by a non-trained annotator. We had the sample of 200 question segments from the evaluation step annotated by a novice untrained annotator not involved in the creation of the coding scheme who was only provided with Table 3, and we obtained kappa values between novice and expert annotators of 0.59–0.67 on Dim1, 0.38 on Dim2, and 0.33–0.39 on Dim3. It would be particularly interesting to apply this coding scheme to annotate questions (manually or automatically) in other contexts (such as MOOCs) to evaluate its generalizability for different applications, and check if correlations found between student characteristics and the type of questions they ask remain the same. We also expect to see that the relevance of some dimensions may be context-dependent, which would allow us to move towards a more general coding scheme.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work has been supported by the HUBBLE project (ANR-14-CE24-0015).

Acknowledgements

The authors would like to thank Pierre Gilois for providing data and supporting this work from the beginning, and Océane Bouhineau-Chevigny for preparing data to facilitate the replication of the analyses on different datasets.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development, 60*(3), 383–398. http://doi.org/10.1007/s11423-012-9235-8

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596. http://doi.org/0.1162/coli.07-034-R2

Bergsma, W. (2013). A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society*, *42*(3), 323–328. http/doi.org/10.1016/j.jkss.2012.10.002

Bloom, B. S., Engelhart, M. B., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives. *The classification of educational goals. Handbook I: Cognitive domain*. New York: Longmans Green.

Bouchet, F. (2009). Characterization of conversational activities in a corpus of assistance requests. *Proceedings of the 14th Student Session of the European Summer School for Logic, Language, and Information* (ESSLLI 2018), 6–17 August 2018, Sofia, Bulgaria (pp. 40–50).

Cao, M., Tang, Y., & Hu, X. (2017). An analysis of students' questions in MOOCs forums. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (EDM2017), 25–28 June 2017, Wuhan, China (pp. 412–413). International Educational Data Mining Society.

Chin, C., & Brown, D. E. (2000). Learning deeply in science: An analysis and reintegration of deep approaches in two case studies of grade 8 students. *Research in Science Education, 30*(2), 173–197. http://doi.org/10.1007/BF02461627

Chin, C., & Brown, D. E. (2002). Student-generated questions: A meaningful aspect of learning in science. *International Journal of Science Education, 24*(5), 521–549. http://doi.org/10.1080/09500690110095249

Chin, C., & Kayalvizhi, G. (2002). Posing problems for open investigations: What questions do pupils ask? *Research in Science & Technological Education, 20*(2), 269–287. http://doi.org/10.1080/0263514022000030499

Chin, C., & Osborne, J. (2008). Students' questions: A potential resource for teaching and learning science. *Studies in Science Education, 44*(1), 1–39. http://doi.org/10.1080/03057260701828101

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Colbert, J. T., Olson, J. K., & Clough, M. P. (2007). Using the web to encourage student-generated questions in large-format introductory biology classes. *CBE-Life Sciences Education, 6*(1), 42–48.

Cook, C., Olney, A. M., Kelly, S., & D'Mello, S. K. (2018). An open vocabulary approach for detecting authentic questions in classroom discourse. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th International Conference on Educational Data Mining* (EDM2018), 16–20 July 2018, Buffalo, New York, USA (pp. 116–126). International Educational Data Mining Society.

Elgort, I., Lundqvist, K., McDonald, J., & Moskal, A. C. M. (2018). Analysis of student discussion posts in a MOOC: Proof of concept. *Companion Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 5–9 March 2018, Sydney, NSW, Australia. SoLAR. Retrieved from https://solaresearch.org/core/companion-proceedings-of-the-8th-international-learning-analytics-knowledge-conference-lak18/

Etkina, E., & Harper, K. A. (2002). Weekly reports: Student reflections on learning. *Journal of College Science Teaching, 31*(7), 476.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–8. http://doi.org/10.1037/a0024338

Graesser, A. C., Ozuru, Y., & Sullins, J. (2009). What is a good question? In M. G. McKeown & L. Kucan (Eds.), *Threads of coherence in research on the development of reading ability* (pp. 112–141). New York: Guilford.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*(1), 104–137. http://doi.org/10.2307/1163269

Graesser, A. C., Person, N. K., & Huber, J. D. (1992). Mechanisms that generate questions. In T. Lauer, E. Peacock, & A. C. Graesser (Eds.), *Questions and information systems* (pp. 167–187). Hillsdale, NJ: Erlbaum. http://doi.org/10.4324/9780203763148

Harrak, F., Bouchet, F., & Luengo, V. (2017). Identifying relationships between students' questions type and their behavior. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (EDM2017), 25–28 June 2017, Wuhan, China (pp. 402–403). International Educational Data Mining Society.

Harrak, F., Bouchet, F., Luengo, V., & Gillois, P. (2018). Profiling students from their questions in a blended learning environment. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 5–9 March 2018, Sydney, NSW, Australia (pp. 102–110). New York: ACM. http://doi.org/10.1145/3170358.3170389

Harper, K. A., Etkina, E., & Lin, Y. (2003). Encouraging and analyzing student questions in a large physics course: Meaningful patterns for instructors. *Journal of Research in Science Teaching, 40*(8), 776–791. http://doi.org/10.1002/tea.10111

Ishola, O. M., & McCalla, G. (2017). Predicting prospective peer helpers to provide just-in-time help to users in question and answer forums. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (EDM2017), 25–28 June 2017, Wuhan, China (pp. 238–243). International Educational Data Mining Society.

Kim, J., Shaw, E., & Ravi, S. (2011). Mining student discussions for profiling participation and scaffolding learning. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of Educational Data Mining* (pp. 299–310). Boca Raton, FL: CRC Press. http://doi.org/10.1201/b10274-24

Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics, 32*(4), 485–525. http://doi.org/10.1162/coli.2006.32.4.485

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice Hall.

Kolb, D. (1985). *LSI learning style inventory: Self-scoring inventory and interpretation booklet*. Boston, MA: McBer & Co.

Li, H., Duan, Y., Clewley, D. N., Morgan, B., Graesser, A. C., Shaffer, D. W., & Saucerman, J. (2014). Question asking during collaborative problem solving in an online game environment. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (ITS 2014), 5–9 June 2014, Honolulu, HI, USA (pp. 617–618). New York: Springer. http://doi.org/10.1007/978-3-319-07221-0_80

Liu, Q., Peng, W., Zhang, F., Hu, R., Li, Y., & Yan, W. (2016). The effectiveness of blended learning in health professions: Systematic review and meta-analysis. *Journal of Medical Internet Research, 18*(1). http://doi.org/10.2196/jmir.4807

Marbach-Ad, G., & Sokolove, P. G. (2000a). Can undergraduate biology students learn to ask higher level questions? *Journal of Research in Science Teaching, 37*(8), 854–870. http://doi.org/10.1002/1098-2736(200010)37:8<854::AID-TEA6>3.0.CO;2-5

Marbach-Ad, G., & Sokolove, P. G. (2000b). Good science begins with good questions. *Journal of College Science Teaching, 30*(3), 192.

Otero, J., & Graesser, A. C. (2001). PREG: Elements of a model of question asking. *Cognition and Instruction, 19*(2), 143–175. . Good science begins with good questions. *Journal of College Science Teaching, 30*(3), 192. http://doi.org/10.1207/S1532690XCI1902_01

Scardamalia, M., & Bereiter, C. (1992). Text-based and knowledge-based questioning by children. *Cognition and Instruction, 9*(3), 177–199. http://doi.org/10.1207/S1532690XCI1902_0110.1207/s1532690xci0903_1

Pedrosa de Jesus, H., Almeida, P. C., & Watts, M. (2004). Questioning styles and students' learning: Four case studies. *Educational Psychology, 24*(4), 531–548. http://doi.org/10.1207/S1532690XCI1902_0110.1080/0144341042000228889

Pedrosa de Jesus, H. P., Teixeira-Dias, J. J., & Watts, M. (2003). Questions of chemistry. *International Journal of Science Education, 25*(8), 1015–1034. http://doi.org/10.1080/09500690305022

Pizzini, E. L., & Shepardson, D. P. (1991). Student questioning in the presence of the teacher during problem solving in science. *School Science and Mathematics, 91*(8), 348–352. http://doi.org/10.1111/j.1949-8594.1991.tb12118.x

Sindhgatta, R., Marvaniya, S., Dhamecha, T. I., & Sengupta, B. (2017). Inferring frequently asked questions from student question answering forums. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (EDM2017), 25–28 June 2017, Wuhan, China (pp. 256–261). International Educational Data Mining Society.

Supraja, S., Hartman, K., Tatinati, S., & Khong, A. W. (2017). Toward the automatic labeling of course questions for ensuring their alignment with learning outcomes. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (EDM2017), 25–28 June 2017, Wuhan, China (pp. 56–63). International Educational Data Mining Society.

Watts, M., Gould, G., & Alsop, S. (1997). Questions of understanding: categorising pupils' questions in science. *School Science Review, 79*(286), 57–63.

White, R. T., & Gunstone, R. F. (1992). *Probing understanding*. London: Falmer Press.

APPENDIX

Table 12: List of Regular Expressions Based on Keywords Associated to Each Dimension (in French)

List of regular expressions	Dim1	Dim2	Dim3	Dim4
[Rr].[-_]*expliqu.*|refai.*|reven.*|redi.*|re-?pr.cis.*|rR]appele.*|repren.* | re-?d.fini.*	1	0	0	0
expliqu.*|explic.*|pr.cis.*|d.velopp.*|d.taill.*|clarifi.*|[Qq]uoi|.u\'entend.*|[Qq]u(\')?el(le)?s?|[Ll]es?quels?|r.sum.|[qQ]uand|[Cc]ombien|(D\')?[Oo]ù|ne_comprends?_pas|donn.*|[Qq]ue_représent.*	2	0	0	0
d.fini.*|signifi*|.u\'appell?e.*|[Qq]ue_veut|vaut|.u\'est*	2	0	1	0
..urquoi|[rR]aison	2	0	3	0
(l\')?erreur|comment_est-ce_possible|Quelle_possibilité|cependant|[Pp]ourtant|contradiction|alors_qu\'ils?|alors_que	3	0	0	1
[Cc]omment|cmt|d.marche|m.canisme	2	0	2	0
[Oo][uU]$|[pP]eu.*[-_].*|[Dd]oit[-_].*|dit_on|S\'agit*|.st[-_]|N\'est.*|c\'est_bien	3	0	0	0
correspond(ent)?-.*|concerne(nt)?-.*|s?ont-.*	3	0	5	0
concours|coeur|apprendre|retenir|toujours_donn*|._savoir|[fF]aut[-_]il|conna.t-on|exam.*	3	0	0	3
.xemple	0	1	0	0
sch.ma|repr.sent.*|configur.*|[Ff]igure.*|image*|l.gend.*	0	2	0	0
correct.*|corrig.*|r.ponse|raisonnement|(l\')?.xercice|.uestion	0	3	0	0
ce_qu\'est|ce_que_sont	0	0	1	0
ser.*|[Rr].l.*|(l\')?utilit.*|fonction.*	0	0	4	0
diff.ren.*|lien.*|(l\')?.quivalen.*|m.me_chose|m.me_.*|relat.*|simil.*|limite_.*|.ntre_.*|synonym.*	0	0	5	0

Table 13: Centroids for the 44 Features Associated to Each Cluster of Each Course

Features	BCH				HBDD				BCE			ANT		
	A	B	C	D	A	B	C	D	A	B	D	A	B	D
prop_d1_1	**.80**	.12	*.10*	.16	*.03*	.09	**.50**	.09	.13	*.08*	**.19**	**.21**	*.08*	.16
prop_d1_2	*.15*	.34	**.83**	.23	**.94**	.29	.42	*.24*	**.75**	*.26*	.46	**.71**	*.27*	.38
prop_d1_3	*.02*	.54	.05	**.60**	*.03*	.62	.04	**.66**	*.09*	**.66**	.32	*.08*	**.65**	.46
prop_d2_1	*.06*	**.14**	.09	.07	.02	.04	*.00*	**.06**	**.04**	.03	*.00*	**.10**	**.10**	*.03*
prop_d2_2	*.04*	.22	.16	**.35**	*.07*	.28	.26	**.37**	*.04*	**.10**	**.10**	*.17*	**.36**	.27
prop_d2_3	*.02*	**.23**	**.23**	.12	.02	.02	*.00*	**.04**	.08	**.09**	*.04*	*.00*	.01	**.03**
prop_d3_1	**.23**	*.09*	.14	.13	**.36**	.17	*.11*	.15	**.21**	.11	*.10*	**.12**	*.10*	.11
prop_d3_2	.16	.21	**.28**	*.10*	*.06*	**.11**	.10	*.06*	**.24**	.09	*.04*	.17	**.20**	*.13*
prop_d3_3	*.04*	.20	**.22**	.10	.07	*.05*	**.12**	.09	*.11*	.14	**.23**	.19	*.06*	**.27**
prop_d3_4	*.02*	.04	.04	**.08**	**.10**	*.06*	.09	.07	*.10*	.11	**.18**	.08	**.10**	*.05*
prop_d3_5	*.15*	.20	.16	**.23**	.16	**.35**	*.14*	.23	*.12*	**.16**	*.12*	**.25**	**.25**	*.14*
prop_d4_1	.07	*.06*	.12	**.19**	*.05*	.06	.06	**.14**	.08	*.03*	**.15**	.12	**.14**	*.08*
prop_d4_2	*.01*	**.84**	.03	.73	.05	**.91**	*.02*	.78	*.04*	**.90**	.36	*.07*	**.80**	.64
prop_d4_3	*.03*	**.10**	.09	.08	*.03*	*.03*	**.08**	**.08**	**.12**	*.06*	.07	**.08**	.07	*.05*
P_d1_1_S1	**.74**	.11	.09	*.07*	.03	.08	**.40**	*.02*	**.13**	.05	*.01*	**.20**	.06	*.00*
P_d1_1_S2	.06	*.01*	*.01*	**.09**	*.00*	.01	**.10**	.07	*.00*	.03	**.18**	*.00*	.02	**.16**
P_d1_2_S1	.12	.32	**.75**	*.09*	**.93**	.27	.11	*.08*	**.73**	.21	*.03*	**.69**	.24	*.04*
P_d1_2_S2	.04	*.02*	.08	**.14**	*.02*	*.02*	**.31**	.16	*.01*	.05	**.43**	*.02*	.03	**.34**
P_d1_3_S1	*.01*	**.53**	.04	.07	*.02*	**.60**	.03	.16	.09	**.60**	*.03*	*.04*	**.60**	.06
P_d1_3_S2	*.01*	*.01*	*.01*	**.54**	*.00*	.02	.01	**.50**	*.00*	.06	**.29**	*.05*	*.05*	**.40**
P_d2_1_S1	.06	**.14**	.09	*.05*	.02	**.03**	*.00*	*.00*	**.04**	.01	*.00*	**.10**	**.10**	*.03*
P_d2_1_S2*	*.00*	*.00*	*.00*	**.02**	*.00*	*.00*	*.00*	**.06**	*.00*	**.02**	*.00*	*.00*	*.00*	*.00*
P_d2_2_S1	*.04*	**.22**	.14	.12	*.07*	**.28**	.19	.08	.04	**.10**	*.01*	.07	**.34**	*.00*
P_d2_2_S2	*.00*	*.00*	.01	**.23**	*.00*	*.00*	.06	**.29**	*.00*	*.00*	**.09**	.10	*.01*	**.27**
P_d2_3_S1	*.02*	**.21**	**.21**	.06	.02	.02	*.00*	**.04**	**.08**	.07	*.00*	*.00*	*.00*	*.00*
P_d2_3_S2*	*.00*	.03	.01	**.06**	*.00*	*.00*	*.00*	*.00*	*.00*	.02	**.04**	.00	.01	**.03**
P_d3_1_S1	**.16**	.08	.13	*.02*	**.36**	.17	*.03*	.06	**.21**	.09	*.04*	**.12**	.06	*.01*
P_d3_1_S2	.07	*.01*	*.01*	**.11**	*.01*	*.01*	.08	**.09**	*.00*	.02	**.06**	*.00*	.04	**.11**
P_d3_2_S1	.13	.20	**.26**	*.08*	.06	**.11**	*.00*	.05	**.24**	.07	*.00*	.16	**.19**	*.01*
P_d3_2_S2	**.03**	*.01*	.02	**.03**	*.00*	.01	**.10**	.02	*.00*	.02	**.03**	*.01*	*.01*	**.12**
P_d3_3_S1	*.04*	**.19**	.18	*.04*	.07	.04	**.08**	*.02*	**.10**	**.10**	*.01*	**.17**	.05	*.02*
P_d3_3_S2	*.00*	*.00*	.03	**.06**	*.00*	.01	.04	**.07**	*.01*	.04	**.22**	.02	*.01*	**.25**
P_d3_4_S1	.02	**.04**	.02	*.01*	**.09**	.06	.06	*.04*	**.10**	.09	*.00*	.08	**.10**	*.01*
P_d3_4_S2*	*.00*	*.00*	.01	**.07**	.01	*.00*	*.00*	**.03**	*.00*	.01	**.18**	*.00*	*.00*	**.04**
P_d3_5_S1	.09	**.19**	.16	*.06*	.16	**.34**	*.05*	.10	.12	**.13**	*.03*	.23	**.24**	*.03*
P_d3_5_S2	.06	.02	*.01*	**.17**	*.00*	*.00*	.09	**.13**	*.00*	.03	**.09**	.02	*.01*	**.11**
P_d4_1_S1	.04	.05	**.10**	*.03*	.05	**.06**	**.06**	*.03*	*.00*	**.07**	.01	*.01*	**.07**	*.01*
P_d4_1_S2	.03	*.01*	.02	**.16**	.01	*.00*	*.00*	**.11**	*.00*	.01	**.03**	*.00*	*.00*	**.02**
P_d4_2_S1	*.00*	**.82**	.01	.10	.04	**.87**	*.00*	.21	.04	**.81**	*.03*	*.02*	**.76**	.09
P_d4_2_S2	*.01*	.02	*.01*	**.62**	*.01*	.03	.02	**.58**	*.00*	.09	**.33**	.05	*.04*	**.54**
P_d4_3_S1	*.00*	**.10**	.07	.00	.03	.03	.03	*.01*	**.12**	.05	*.01*	**.08**	.05	*.00*
P_d4_3_S2*	.02	*.00*	.01	**.08**	*.00*	*.00*	.05	**.06**	*.00*	.01	**.06**	*.00*	.01	**.05**
PropQ_S1	.87	**.96**	.89	*.23*	**.98**	.95	.54	*.27*	**.99**	.86	*.07*	**.92**	.90	*.10*
PropQ_S2	.13	*.04*	.11	**.77**	*.02*	.05	.46	**.73**	*.01*	.14	**.93**	*.08*	.10	**.90**

Prop_dx-y: the proportion of questions asked overall with value y in dimension x.
P_dx-y_S1: the proportion of questions asked during the first half of the course with value y in dimension x.
PropQ_S1: the proportion of questions asked during the first half of the course for all the dimension.
* Non-significant feature (not discriminant in either of the four clusterings).

Figure 11: Comparison of 2012 and 2013 clusters in term of median (centre of the bar), 1st quartile (bottom of the bar), and 3rd quartile (top of the bar) of dependent variables and mean and standard deviation of RetStu and SucStu for each cluster of HBDD, BCE, and ANT.