Good afternoon
My name is Elisabeth Delozanne from Paris, University Pierre et Marie Curie
My talk is about
Evaluating the performance of a diagnostic system in School Algebra
It is a part of the Pépite Project, a multidisciplinary project involving

• Computer Scientists (Naïma El-Kechaï, Dominique Prévôt, and I),
• Mathematical Learning Scientists (Brigitte Grugeon & Françoise Chenevotot)
• and Mathematics teachers from Sésamath, an important on-line community of math teachers in France with 14 500 registered teachers.

The project aims at building cognitive models to improve the teaching and learning of algebra in 9th and 10th grades.
First, I will introduce the Pépite project
Then The diagnostic system called PépiDiag and an evaluation of the system

I will illustrate with a simple exercise the problems we have faced and the solutions we have proposed
The Pépite project

- Objective
  - To design and implement a web-based application
  - To support math teachers
  - To manage the students’ cognitive diversity
- Domain
  - Algebra class
  - 9th-10th grade (15-16 years old)

The objective of the Pépite project is to design and implement a web-based application to support math teachers in managing the cognitive diversity of their students in school algebra classes.
A teaching scenario

Who?
- Mary is a Math teacher in 9th grade

Context
- To bring all her class to the same level

Steps
1. Students have the online diagnostic test
2. Pépite
   a) analyzes the students’ answers
   b) displays a cognitive profile of each student and of the whole class
   c) suggests 6 groups of students and a session of exercises adapted to their cognitive profiles
3. Mary can adjust the groups and the exercises
4. Students log in and do the exercises in their assigned session

Let’s see a teaching scenario
Mary is a math teacher in 9th grade
She wants to bring all her class to the same level
and organize some differentiated work
before teaching a new chapter « Algebraic calculation and equation »

She asks her students to connect to Labomep (the Sesamath platform) to have The Pépite diagnostic Test.

After the test

1. First, Pépite assesses the students’ answers: my talk today deals with this particular step, which is very important, because the whole system relies on the validity of this step.
2. Second, from this assessment, Pépite builds a cognitive profile for each student and for the whole class; Pépite displays these profiles to Mary
3. Third, Pépite suggests 6 groups of students and for each group a session of exercises adapted to the students’ cognitive profiles
For instance, in this small class (with fictitious names) Pépite displays 3 groups

- **Groupe A+**
  - Engage in algebraic thinking
  - Use algebraic calculation appropriately

- **Groupe B+**
  - Begin to use algebra to solve problems
  - Use sometimes mal-rules

- **Groupe C-**
  - Stuck in arithmetical thinking
  - Algebra makes no sense

For instance, in this small class (with fictitious names) Pépite displays 3 groups

- students in Group A+:
  - are already engaged in algebraic thinking and use algebraic calculation appropriately
- students in Group B+:
  - begin to use algebra but they sometimes use algebraic mal-rules
- students in Group C-:
  - keep stuck with arithmetical thinking and Algebra makes no sense for them

From this screen displaying the whole class profile, Mary can inspect each student’s cognitive profile, for instance Sens René in group B+
In this slide, you can see how Pépite displays the student’s profile in group B+

This student is level 2 on the three components of algebraic competence (what we call the student’s stereotype)

His personal features enlighten his strong points and weak points.

This is the basis for specifying differentiated sessions

<table>
<thead>
<tr>
<th>Level</th>
<th>Personal features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algebraic calculation</td>
<td>Success rate on asked questions: 4/12</td>
</tr>
<tr>
<td></td>
<td>Strong points</td>
</tr>
<tr>
<td></td>
<td>Interpreting algebraic expressions</td>
</tr>
<tr>
<td></td>
<td>Success rate on asked questions: 11/23</td>
</tr>
<tr>
<td></td>
<td>Partial mastery of algebraic rules</td>
</tr>
<tr>
<td></td>
<td>Some good interpretations of algebraic expressions</td>
</tr>
<tr>
<td></td>
<td>Weak points</td>
</tr>
<tr>
<td></td>
<td>Low mastery of algebraic calculation</td>
</tr>
<tr>
<td>Usage of algebra</td>
<td>Success rate on asked questions: 5/9</td>
</tr>
<tr>
<td></td>
<td>Strong points</td>
</tr>
<tr>
<td></td>
<td>Mathematical modeling</td>
</tr>
<tr>
<td></td>
<td>Good mastery of algebra usage on some problems</td>
</tr>
<tr>
<td></td>
<td>Weak points</td>
</tr>
<tr>
<td></td>
<td>Justification by school authority</td>
</tr>
<tr>
<td>Algebraic translation</td>
<td>Success rate on asked questions: 12/24</td>
</tr>
<tr>
<td></td>
<td>Strong points</td>
</tr>
<tr>
<td></td>
<td>Translating situations to algebra</td>
</tr>
<tr>
<td></td>
<td>Some good translations of mathematical relations</td>
</tr>
<tr>
<td></td>
<td>Weak points</td>
</tr>
</tbody>
</table>

More information can be accessed by clicking on the respective cells in the table.
Then, Mary wants to organize sessions of online exercises adapted for each group.

Pépite suggests to Mary, for each group, a session of online exercises

Mary can drag and drop students or exercises or she can create new groups to adapt the session for special needs
At the end, students log in and do the exercises of their assigned session
Implementing a Cognitive modeling system is known to be a difficult and time consuming task especially when it diagnoses open answers, multistep reasoning or when multiple reasoning processes are equivalent.

Our approach relies on a three stage process:

The first step is called a local diagnosis: each student’s answer is assessed on several dimensions. In this step, the diagnostic system provides a set of codes that characterize types of anticipated answers.

The second step collects similar codes on different exercises to build the student’s cognitive profile expressed by success rates, strong points and weak points (what we call the student’s personal features).

The third step locates the student on a scale of competence (what we call the student’s stereotype).

The whole process relies on the quality of the first step: assessment of each student’s answer.

Now I will present, with a simple example, how the PépiDiag diagnostic system deals with this first step.
In this example, the student is asked to find the area of the blue rectangle.

Here, you can see a very common answer from students.

This student calculated the area of each sub rectangle (that’s correct) but he translated a squared as: 2 times a.
The PépiDiag diagnostic System assesses this answer on 5 dimensions

- The answer is not valid (coded V3)
- Letters are used with incorrect rules (coded L3)
- Translation from geometry to algebra is incorrect (coded T3)
- The student used an incorrect rule  $a*a$ as $2* a$ (coded EA 41) (ié forty one)
- The student used algebra to prove but with an incorrect rule (coded J3)

In the following step, the system uses these codes to build the student’s cognitive profile by collecting the same codes along different exercises

But to inform the teacher, those codes are not very useful ; so there is a label to describe the type of answer for the teacher :

Recognition of sub-figures but turning a product into a sum
The PépiDiag diagnostic system works with two kinds of data

First, when a student solves the exercises of a test, the answers are stored in a database.
Second, each exercise is stored in a database with an XML file called the coding prescription file.

When the teacher asks for her students’ profiles, PépiDiag loads 2 XML files
  • One with the student’s answers
  • and one with each exercise coding prescription file
  • As a result, PépiDiag stores an XML file with an assessment of each answer and with a student’s profile into the student’s database
This is a partial view of the coding prescription file for our example

It describes every type of anticipated answers for a question

A type is characterized by
1. A label
2. A set of codes
3. And a list of different patterns of anticipated answers for this type

PépiDiag matches the student’s answer with one of the patterns in the prescription coding file. To do that, it is necessary to use Computer Algebra System to deal with equivalence of expressions (here with commutativity)

This file was first specified with mathematical learning scientists and then has been improved with data collected in usage because completeness is not a realistic goal with open questions
To sum up:

In the first step of the diagnostic process, PépiDiag assesses the student’s answers by matching the answer with a pattern of anticipated answers. The result of this step (called local diagnosis), is for each question a set of codes.

In the second step, PépiDiag collects the similar codes in the different exercises of the test to build the student’s cognitive profile.

This profile is made of success rates, strong points and weak points and a level on different components of the competence (as we saw on slide 6).
Let’s see now how we evaluated the automatic assessment of students’ answers
We stated 3 criteria:

1. No correct answer is to be badly coded
2. Better no code than a wrong one
3. Minimal number of answers left unanalyzed

To validate the system we compared the system assessment with the assessment of 3 human experts: 2 researchers and a teacher.

First, experts worked separately on 3 hundred and sixty answers for each exercise. Then they discussed together and corrected some minor errors. Let's see the results of this comparison in our example.
PépiDiag analyzes more than eighty percent of the student’s answers. We observed that this was more than a regular teacher would do, but less than an experienced researcher actually did.
Correct answers

<table>
<thead>
<tr>
<th>Correct answers</th>
<th>Correct automatic coding</th>
<th>Unanalyzed answers by PépiDiag</th>
<th>Unanalyzed answers By experts</th>
</tr>
</thead>
<tbody>
<tr>
<td>149/266</td>
<td>136/149</td>
<td>13/149</td>
<td>0</td>
</tr>
<tr>
<td>(56%)</td>
<td>(91%)</td>
<td>(9 %)</td>
<td></td>
</tr>
</tbody>
</table>

- No correct answer is analyzed incorrectly
- Unanalyzed answers
  - 5/13 : Using X instead of *
  - 8/13 : Mixing algebra and natural language :
    - « (a+b) times (3+a) »
- Solutions
  - To prevent students from typing letters other than those relating to the exercise statement

No correct answer is analyzed incorrect
So our first criteria is actually satisfied

PépiDiag could not analyze some correct answers because five students used letters instead of operators and eight mixed natural language and algebraic writing

In the new test we now deliver on line, the interface prevents students from typing letters others than those in the statement
So we have solved this kind of problem
There are only two incorrect answers badly assessed because the Computer Algebra System cannot distinguish between 2 different expressions that are mathematically equivalent but that are built by two different student’s reasoning processes.

In the first, the student forgot the parenthesis; in the second, he gathered the figure items.

The experts also failed to analyze 21% of the incorrect answers, because completeness is not a realistic goal in open answers as we already notice.

The difference between PépiDiag and experts, have the same explanation than for correct answers: the students used letters instead of operators and mixed natural language with algebraic expression.

Let’s note that it is more difficult for human to diagnose precisely incorrect answers, and so experts made more slips before correcting them in the debriefing discussion.

So we observed that PépiDiag was more effective than one human alone but less than the three experts together.
Conclusion

- Our approach to assess students’ open answers
  - Human experts
    - typify anticipated patterns of correct and incorrect solutions
    - code each type on several dimensions
  - PépiDiag, using a CAS,
    - matches the student’s solution with a pattern
- PépiDiag is reliable
  - 100 % : closed answers
  - 80 % : answers with one algebraic expression
  - 70 % : answers with a multi-step reasoning [ITS 2008]
- PépiDiag saves teachers’ time and tedious effort

Our approach to assess student’s answers relies
  • on a human expertise to typify anticipated correct and incorrect solutions
  • And on coding each student’s answer by matching it with a anticipated pattern of solutions using a Computer Algebra System

Let’s note that this first human expertise can improve with usage : we can easily add new patterns in the XML coding prescription file when we meet new answers that human experts can interpret.

The different studies comparing PépiDiag codes to experts assessment have shown that PépiDiag was reliable even with open questions.

Although it can not diagnose some open answers, it is better than a single teacher on a big set of answers.
Overall it provides a tool to save time and very tedious work for a human.
Our current work, on the diagnostic part, deals the validation of the building of the student’s cognitive profile from the answers analysis.

At the moment, we have implemented a heuristic method based on a mathematical learning science expertise.

We are currently comparing the results with other methods such as Item Response Theory or Neural Networks that are mathematically grounded.

Now that we consider we have a reliable diagnostic tool, the problem is to design and to validate the effectiveness of different learning paths adapted to each profile.

This work is done by a PhD student, Julia Pilet, in collaboration with math teachers from Sésamath.

She conducted experiments with four teachers using Pépite at two school levels and at different times in the school year.

She is studying the student’s evolution during the school year.

An other work is done by Naïma El-Kechaï, a post-doctoral researcher, to define the meta data and index the whole data-base of Sésamath (2 thousand exercises). Her system displays sessions on Sésamath platform adapted to the student’s profile as we see in Mary scenario.

Thank you.