Salles *et al. Large-scale Assess Educ* (2020) 8:7
https://doi.org/10.1186/s40536-020-00085-y

Large-scale Assessments in Education

RESEARCH Open Access

When didactics meet data science: process data analysis in large-scale mathematics assessment in France

Franck Salles*, Reinaldo Dos Santos and Saskia Keskpaik

*Correspondence: franck.salles@education.gouv.fr
Department of Evaluation (DEPP), Ministry of Education, 65 rue Dutot, Paris, France

Abstract

During this digital era, France, like many other countries, is undergoing a transition from paper-based assessments to digital assessments in education. There is a rising interest in technology-enhanced items which offer innovative ways to assess traditional competencies, as well as addressing problem solving skills, specifically in mathematics. The rich log data captured by these items allows insight into how students approach the problem and their process strategies. Educational data mining is an emerging discipline developing methods suited for exploring the unique and increasingly large-scale data that come from such settings. Data-driven methods can be helpful when trying to make sense of process data. However, studies have shown that didactically meaningful findings are most likely generated when data mining techniques are guided by theoretical principles on subjects' skills. In this study, theoretical didactical grounding has been essential for developing and describing interactive mathematical tasks as well as defining and identifying strategic behaviors from the log data. Interactive instruments from France's national large-scale assessment in mathematics have been pilot tested in May 2017. Feature engineering and classical machine learning analysis were then applied to the process data of one specific technology-enhanced item. Supervised learning was implemented to determine the model's predictive power of students' achievement and estimate the weight of the variables in the prediction. Unsupervised learning aimed at clustering the samples. The obtained clusters are interpreted by the mean values of the important features. Both the analytical model and the clusters enable us to identify among students two conceptual approaches that can be interpreted in theoretically meaningful ways. If there are limitations to relying on log data analysis in order to determine learning profiles, one of them is the fact that this information remains partial when it comes to describing the complete cognitive activity at play, the potential of technology-enriched problem solving situations in large-scale assessments is nevertheless obvious. The type of findings this study produced is actionable from teachers' perspective in order to address students' specific needs.

Keywords: Large-scale assessment, Mathematics, Machine learning, Data science, Theoretical framework, Technology, Didactics, Process data

Introduction

During this digital era, France, like many other countries, is undergoing a transition from paper-based assessments to digital assessments in order to measure student performance in education. New opportunities are emerging (cost reduction, innovative items, adaptive testing, real-time feedback into learning, etc.) which themselves give rise to new challenges (usability, security, equipment, digital divide, etc.).

There is a rising interest in France in technology-enhanced items which offer innovative ways to assess traditional competencies, as well as to address 21st century skills and to link assessment feedback to learning. The technology-enhanced items can be extremely valuable when measuring problem solving skills. Compared to traditional assessments, they not only provide scoring information–whether the response is correct or not–but allow rich data to be collected that enable the way that the students have arrived at their answers to be determined (Greiff et al. 2015).

These complex technology-enhanced items can be used to reflect how students interact in a given situation to analyze and solve a problem. The exercises and questions included in the interactive items engage the students on multiple levels, and capture not just their responses, but their thought process as well. The rich log data captured by these items, such as the time at which students start and stop their work, mouse movements, the use of different onscreen tools, idle time, and a screenshot of the last actions, allow insights to be gained into how students approach the problem, and to identify areas that might require additional focus.

Despite the potential gain in knowledge about student performance, the studies on log data from educational assessments remain relatively scarce (Greiff et al. 2015). One of the reasons for the scarcity of studies is the technicality that these analyses entail. Although attempts are made in order to standardize the logs and develop specific data analysis tools (Hao et al. 2016), logs are often messy, unstructured, and full of "noise"–all of which leads traditional data analysis tools and techniques to work less well with these data.

Process data, recorded as sequences of actions, can be likened to textual data and analyzed by making use of methodologies of natural language processing and text mining. Hao et al. (2015) transform process data into a string of characters, encoding each action name contained in the logs as a single character. The authors use the Levenshtein distance, defined as the minimum number of single-character edits needed to convert one character string to another, in order to compare how far the students' activities in game/scenario-based tasks are from the best performance. In the same vein, He and von Davier (2015), considering the similar structure between action sequences in process data and word sequences in natural language, make use of *N-grams*–a contiguous sequence of n items from a given sample of text–in order to discern action sequence patterns that are associated with success or failure in complex problem solving tasks.

Educational data mining is an emerging discipline that is concerned with developing methods specifically suited for exploring the unique and increasingly large-scale data that come from educational settings.[1] Qiao and Jiao (2018) showcase various data

[1] http://educationaldatamining.org/.

mining techniques in analyzing process data and demonstrate how both supervised as well as unsupervised learning methods can help revealing specific problem solving strategies and distinguish between different performance profiles.

Data-driven methods can be very helpful when trying to make sense of huge amounts of data and discover hidden patterns and relationships in these data. However, studies have shown that didactically meaningful findings are most likely yielded when data mining techniques are guided by theoretical principles allowing to describe subjects' skills (Gobert et al. 2013).

In the context of complex problem solving, studies have demonstrated how certain student behaviors yield better performances than others (Greiff et al. 2015, 2016). Theoretical grounding has been essential for defining and identifying these strategic behaviors from the log data and verifying their implementation among different samples of students.

CEDRE (Subject-related sample-based assessment cycle) is a sample-based large-scale assessment aiming to measure students' abilities in mathematics at the end of grade 9 every 5 years in France. Constructed and designed by the Department for Evaluation, Prospective and Performance (DEPP) at the French ministry of education, its framework is based on the national French curriculum in mathematics. First administered in 2008 and 2014, CEDRE was administered again in May 2019. This new cycle is computer-based for the first time. As trends must be secured so that the comparability with the previous cycles is guaranteed, a large part of the test instruments are similar to formerly paper-based items. However, the DEPP developed technology-enriched items, very different from more classical item formats, in order to profit fully from the potentialities of assessing mathematics with digital tools (Stacey and Wiliam 2013). Offering students the possibility to use digital tools during the assessment may outsource basic procedural work to such tools (Drijvers 2019). Therefore opportunities are given to students to engage more fully higher order skills such as problem solving, devising a strategy and carrying out mathematical thinking, which the CEDRE aims to capture.

Problem solving strategies in complex mathematical tasks are targeted by the CEDRE framework since the first cycle (MEN, DEPP 2017). Nevertheless, the way to capture student's strategies and processes in the paper-based assessment cycles used to be based on students' explanation of their answers and written arguments or sketches in response boxes. Communicating a mathematical answer being a mathematical competency in itself (OECD 2013), it does not directly indicate the actual strategy used by students to solve the task but only the way they were able to express it. Logging and analyzing process data can potentially lead to drawing a complete picture of how digital tools and interactions have been activated during the problem solving process. A theoretical framework was designed for the purpose of describing in detail the mathematical tasks in such items, with hypotheses about potential processes or strategic thinking to which these items give rise, and eventually identifying variables of interest. In support of this process, we convened several structuring concepts from research in mathematics didactics and more generally from educational research (notion of "conceptions", task analysis, content analysis, semiotic analysis, assessment, etc.). This framework is based on findings in mainly French didactic research regarding mathematic teaching activities and technology-enriched environments. Its main and seminal references are the Theory

of Didactical Situations (Brousseau 2006), the Activity Theory applied to mathematic education (Robert 1998) and the Instrumental Approach (Rabardel 2002). According to this framework, CEDRE's item designers developed new interactive items and identified which data need to be logged for future process analysis.

Research questions

Based on this preliminary work, this study aims to answer two main research questions:

- To what extent can process data analysis provide information about students' mathematics performance in large-scale assessments and explain achievement?
- To what extent can process data be used to categorize students' mathematical strategic behaviors and procedures, allowing didactical interpretation and profiling?

Theoretical framework

Determining what type of mathematical knowledge and skills are involved in items is a preliminary necessity for assessment task analysis. Beyond listing them, we have to identify and describe the way they must or could be operated and what operation's adaptions are necessary to resolve the tasks, given the underlying mathematical conceptions at stake. This level of analysis can be a first step towards determining conceptions involved, choices students have to make, the number of steps required, types of errors, etc. Based on seminal work of Robert and Douady, Roditi and Salles (2015) first applied a didactical framework to PISA 2012 mathematical assessment task analysis: the so called mathematical knowledge operation levels (MKOL). On one hand, MKOLs allow to distinguish between the object and tool characters of mathematical knowledge (Douady 1991). Some assessment questions focus on mathematical content where students must demonstrate an understanding of the concept without having to implement it, which some authors refer to as a *conceptual understanding* (Kilpatrick et al. 2001, p. 115–135); these questions address the object character of this content. Other questions assess the tool nature of knowledge; the student must then put mathematical knowledge into operation in order to solve a problem in the context indicated. On the other hand, MKOLs take into account the variety of ways mathematical knowledge can be implemented to solve an item. The model identifies levels of knowledge implementation ranging from direct operation and operation adaptations to introduction of intermediate steps (Robert 1998). However, this first framework was initially devised to tackle paper-based assessment instruments. In a technology-enhanced environment, specifically when digital tools are available to students, "*technology can impact the way students operate and reason when working on tasks, for example while using CAS to solve equations or while exploring a dynamic construction with a geometry package to develop a conjecture that may be proved*" (Drijvers et al. 2016, p. 12). Therefore the initial didactical framework was extended to additionally take into account this impact in which we distinguish tool's utilizations and instrumentations as well as student/machine interactions. In summary, the didactical framework implemented in this study was structured around three main questions: How does mathematical knowledge need to be adapted in order to resolve an

interactive mathematical task? Which tools' utilization is necessary to solve such problems? How can student/machine interactions influence the mathematical activity?

The way students may use a digital tool at the service of solving a primary task can be described in reference to the instrumental approach (Rabardel 2002). This approach distinguishes a tool from an instrument. The tool gets the status of an instrument when it is used as a mean to solve the problem. In this case, its use is decomposed into "*cognitive schemes containing conceptual understanding and techniques for using a tool for a specific type of task*" (Doorman et al. 2012, p. 1248). The analysis will then focus on identifying and describing utilization schemes potentially involved in the task. Rabardel distinguishes two types of utilization schemes: "*usage schemes, related to "secondary tasks" (…) and instrument-mediated action schemes, (…) related to "primary tasks" (…) [which] incorporate usage schemes as constituents.*" (Rabardel 2002, p. 83). In order to illustrate these two levels of schemes, Rabardel gives the example of an experienced driver overtaking a vehicle: "*An instrument-mediated action scheme underlies the invariable aspects of such an overtaking situation. This scheme incorporates as components usage schemes subordinate to its general organization, such as those necessary to manage a change of gears or a change of trajectory*" (Rabardel 2002, p. 83).

Interactions and feedbacks are important features of problem-solving situations, especially in an assessment context. Being immerged in a digital environment leads to specific kinds of human/machine interactions. Most of them are intentional and planned by developers when designing the assessment environment, others are not but all somehow carry information students can grasp to proceed in the problem-solving process. Laborde (2018) distinguishes two types of feedback with concern to technologically enriched situations: one issued from task-specific digital tools, another being the "teacher's voice". This last type of feedback is meant to help students catch, adapt and retain information given in the environment. In a summative assessment context, this type of feedback should be very limited as it could interfere with the objective of measuring students' ability. Nonetheless, it can be considered paramount in a formative approach. Even if a summative assessment platform such as the one used for the DEPP's assessments can be considered as a non-didactical environment where the teacher's voice is supposed to be absent, we can still consider this environment as potentially providing student/machine interactions. In his theory of didactical situations, Brousseau (2006) separates three levels of feedback depending on the nature of the environment's mathematical reaction: the feedback can either reflect on actions, formulations or validations. This last model has been used to help describe item/student interactions in DEPP's interactive items.

Task analysis

For the purpose of this study, a specific item Fig. 1 has been chosen among CEDRE's new interactive items piloted in May 2017. In this task, two nonlinear functions model two different tree growths. Both are given in linked numerical and graphical representations (Stacey and Wiliam 2013). Students act in the numerical representation (a table of values), entering the age of the trees in months. A calculation tool returns the corresponding tree heights and a graphing tool spots the points in the graph. Both actions are realized when one presses the "Calculate and graph" button. By default, values for

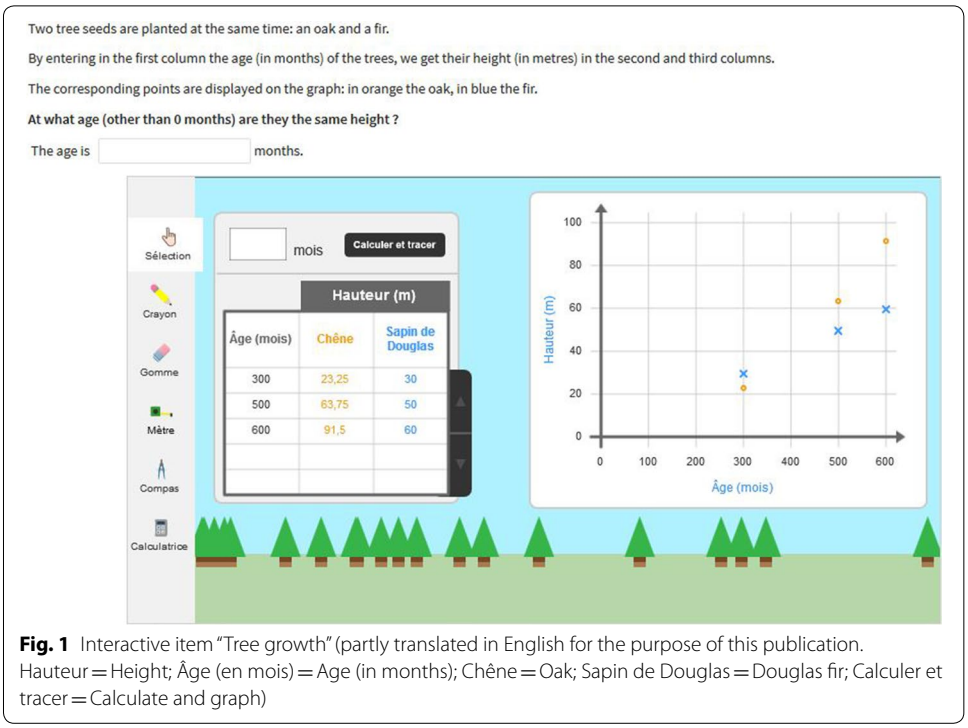

Fig. 1 Interactive item "Tree growth" (partly translated in English for the purpose of this publication. Hauteur = Height; Âge (en mois) = Age (in months); Chêne = Oak; Sapin de Douglas = Douglas fir; Calculer et tracer = Calculate and graph)

300, 500 and 600 months are given. Additional tools can be used: a pencil, an eraser (not allowing to erase given information), a ruler, a compass, a non-scientific calculator. The question can be translated as: "At what age (other than 0 months) do both trees have the same height?"

From a grade 9 students' point of view, this task requires conceptual understanding of functions and their representations (table of values, graph). Functions conception has two different characters as Sfard (1991) and Doorman et al. (2012) showed: "*In lower secondary grades, functions mainly have an operational character and are seen as an input–output 'machine' that process input values into output values. In higher grades, functions have a more structural character with various properties (Sfard 1991). They become mathematical objects that are represented in different ways, are ordered into different types according to their properties, and are submitted to higher-order processes such as differentiation and integration. We argue that the transition from functions as calculation operations to functions as objects is fundamental for conceptual understanding in this domain.*" (Doorman et al. 2012, p. 1243). Students can adapt the problem by adding intermediate information using the calculation and graphing tool. On one hand, they can then opt for a "trial and error" method. This would consist in entering a number of months, comparing the results returned either in the numerical or graphical representation, deciding to enter another number of months until the solution (390) is found. Alternating tries around the target value or aiming at it from below or above could improve the trial and error process. These students show essentially good understanding of the concept of functions in their operational character. This method could imply a relatively large number of tries. On the other hand, students understanding that both functions are increasing, notably from

studying them in the graphical representation, can quickly aim at the target number of months. The pencil can for example be used to draw lines and introduce a continuous representation of the functions. The inversion of tree heights between 300 and 500 months can also be noticed. These students understand functions as objects with properties, in their structural character.

The following digital tools are at students' disposal within the item:

- A keyboard (with or without number pad) and mouse.
- A "calculation and graph" tool, specific to the item.
- A pencil (common to any item on the platform). Usage: click the starting point, move the mouse to trace, click to stop writing.
- An eraser only allowing erasing pencil traces or measurement tool traces. Usage: clicking erases all pencil traces together.
- A compass
- A calculator

The "calculation and graph" tool does not require complex usage schemes. Two usage schemes are identified for this tool: enter a number of months within the domain [0; 600] via an input box and a popup number pad, and understand that the tool returns unique heights (outputs) for both trees (numerical and graphical representations) when the button "calculate and graph" is clicked. No tutorial or tool training is proposed to students. Usage schemes are close to relatively usual tools such as a currency converter. Nevertheless, we can imagine some students feeling the need to appropriate the tool by first using it for testing purposes, for example entering extreme values or values not directly connected to the primary task. Using this tool is compulsory to succeed the item. In reference to the instrumental approach, we describe next how the tool can be instrumented in this situation, once assumed that students will build an instrument from the "calculation and graph" tool, within the item environment, in order to solve the task (Trouche 2003). Instrumented action schemes are organized around the core elements that follow:

1. Knowing the difference between input and output in a contextual use of a function as a model.
2. Understanding that the tool returns unique heights (outputs) for both trees (numerical and graphical representations) when choosing a number of month (inputs).
3. Entering a number of months within the domain [0; 600].
4. Comparing outputs either in the numerical or graphical representation. Validating by linking to the real life situation.
5. Deciding on the next number of month to enter considering the comparison to the previous one.
6. Iterating the process.

This type of instrumentation is linked to an operational approach of the concept of functions.

As mentioned earlier, the pencil can also be used in order to get a continuous model on the domain or part of it. Students can link points together using it. This is an

intermediate step towards the primary task. The following core elements participate to instrumented action schemes using the pencil as well as the "calculate and graph" tool and are characteristic of a structural understanding of the concept of a function.

1. Understanding that the growth phenomenon is a continuous one. Hence, the functions modelling it are continuous.
2. Assuming both functions will be strictly increasing.
3. Using the pencil to link consecutive points together.
4. Decide on the next number of months to enter considering line intersection.
5. Going back to numerical values to aim at accuracy.

Of course one might operate composite instrumented action schemes, mixing the use of both principal tools. For example, one can use the pencil to sketch continuous graphs (potentially after trying 100 and or 200 months to get a more complete view of the graphs shapes), or rely on points' colours and choose 400 months in the first tries and then use a trial and error strategy to aim precisely at the target with the "calculate and graph" tool.

Interactions at stake within the item are principally addressing the two different representations of the functions: numerical and graphical. When students use the "calculation and graph" tool, the feedback is given in both representations as new numbers in the table of values and two points on the graph. Students are consequently relieved from having to convert one representation into another. The colours of numbers and points related to the same function match so students can more easily interpret the feedback. This important interaction participates in students' reflections towards formulating the problem in both representations and then comparing results (for example using colors' inversions in the graphical display) to either conclude or decide on other attempts to make. Besides it can also participate in invalidating attempts that are outside the domain or very far from target.

Data and methods

CEDRE's interactive items, the "Tree growth" item among them, have been used in a pilot test in May 2017 with a sample of 3000 grade 9 students per item. Students' digital traces have been recorded in log data files. As log data files contain a very large amount of data and in order to aim at interpretable results as well as to avoid noisy signals, variables of interest were defined, as a result of the a priori didactical analysis. They could potentially lead to building a model able to explain success or failure in the task considering either the operational or structural character of functions used by students. Problem solving strategies based on a procedural understanding of functions imply using the "calculation and graph" tool more often, potentially through a dichotomous strategy, hence testing many numbers of months and spending more time on the item. Features such as the month list length, the number of alternating within this list, the time spent on the item could then participate bringing the light on such strategies in the data. Symmetrically, a strategy based on a structural conception could lead to optimizing both the number of tries and the time spent, aiming at the target interval from the very first tests, perhaps using the assistance of the pencil tool. Features such as the standard deviation

of the values in the month list and the distance between first inputs and the target show specific interest. Accordingly, the main features used in the analytical models and their nature are the following:

- Month list length:
  - integer variable
  - present in the log data
- First input between 200 and 600:
  - boolean
  - constructed through feature engineering
- Number of alternating within the month list:
  - integer variable
  - constructed through feature engineering
- Time spent on the item, in seconds:
  - continuous variable
  - present in the log data
- Distance between the first input and the target value:
  - integer variable
  - constructed through feature engineering
- Distance between the second input and the target:
  - integer variable
  - constructed through feature engineering
- Distance between the last input and the target:
  - integer variable
  - constructed through feature engineering
- Standard deviation of the values in the month list:
  - continuous variable
  - constructed through feature engineering
- Target value is in the month list:
  - boolean

- present in the log data

- Pencil use:

  - boolean
  - present in the log data

Classical machine learning analysis was then applied to the data. Supervised learning was implemented to determine the predictive power of students' achievement of the model and estimate the weight of the variables in the prediction. Unsupervised learning aimed at clustering the sample. Mean values of the most important variables were then calculated for each of the clusters issued from unsupervised learning.

The type of statistical analysis used, including a power calculation when appropriate, is presented in the following part of the paper. All of the analysis has been done using Python 3.0 and specifically the scikit-learn library.

Supervised learning

The choice of the algorithms to use in order to build the model is partly determined by the task we are trying to achieve. In our case, we are looking for a supervised classification, as the objective is to predict a label (the correct boolean) from the other features. Another criteria is the explain ability of our modeling. We are not trying in this study to build a predictor in itself, we rather aim at determining which features are the most predictive of the score of a student. Therefore, we excluded the use of neural networks. Finally, the efficiency of the model remains the final criteria. As to be able to compare the different algorithms, we chose the area under the ROC curve as our fit, because it is an indicator that can be calculated for any kind of algorithm.

Random forests

Random forests (Breiman 2001) are an ensemble learning method that works by building a multitude of decision trees during learning, before returning the class mode (classification) or the average forecast (regression) of individual trees. The purpose of a decision tree is to create a model that predicts the value of a target variable from several input variables (Hopcroft et al. 1983). A decision tree or classification tree is a tree in which each internal node (non-leaf) is marked with an input characteristic. Arcs from a node labelled with an input characteristic are labelled with each of the possible values of the target or output characteristic, or the arc leads to a subordinate decision node on a different input characteristic. Each leaf of the tree is labelled with a class or probability distribution on the classes, which means the data set has been classified by the tree either in a specific class or in a particular probability distribution.

Decision trees are nevertheless known for their many disadvantages. The first is the tendency to overfit the training set. The second is its non-deterministic aspect: the order of use of the functionalities generates a completely different tree structure.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging (Breiman 1996), to tree learners. Given a training set $X = x1,\ldots, xn$ with responses $Y = y1,\ldots, yn$, bagging repeatedly (B times) selects a

random sample with replacement of the training set and fits trees to these samples. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x', or by taking the majority vote in the case of classification trees.

The number of samples/trees, B, is a free parameter. An optimal number of trees B can be found using cross-validation. However, the question of the order of the features remains unsolved. That is why random forests differ slightly from the general bagging: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called "feature bagging" or "random subspace method" (Barandiaran 1998).

The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated.

The number of trees for the random forest (200) has been determined through cross-validation. The other hyperparameters (number of leaves, maximal depth) have been tuned automatically.

Area under the ROC curve

The receiver operating characteristic (ROC), also known as the performance characteristic or sensitivity/specificity curve, is a measure of the performance of a binary classifier.

Graphically, the ROC measurement is often represented as a curve that gives the rate of true positives (fraction of positives that are actually detected) versus the rate of false positives (fraction of negatives that are incorrectly detected).

They are often used in statistics to show progress using a binary classifier when the discrimination threshold varies. Sensitivity is given by the fraction of Positives classified as Positive, and antispecificity (1 minus specificity) by the fraction of Negatives classified as Positive. The antispecificity is plotted on the x-axis and the sensitivity on the y-axis to form the ROC diagram. Each S value will provide a point on the ROC curve, which ranges from (0, 0) to (1, 1).

- At (0, 0) the classifier always declares 'negative': there are no false positives, but also no true positives. The proportions of true and false negatives depend on the underlying population.
- At (1, 1) the classifier always declares 'positive': there are no true negatives, but also no false negatives. The proportions of true and false positives depend on the underlying population.
- A random classifier will draw a line from (0, 0) to (1, 1).
- At (0, 1) the classifier has no false positives or false negatives, and is therefore perfectly accurate, never getting it wrong.
- At (1, 0) the classifier has no true negatives or true positives, and is therefore perfectly inaccurate, always being wrong. Simply invert its prediction to make it a perfectly accurate classifier.

When using normalized units, the area under the curve is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative').

The area under the ROC curve (ROC AUC) is a common method used for model comparison (Bradley 1997).

Unsupervised learning

Unsupervised learning is a machine learning technique used to detect patterns in a data set, with no prior information about this data. It is mainly used as a clustering technique, to group or segment the data, in order to identify communalities.

Clustering algorithms can be classified into several families. The main ones are the density-based or centroid-based algorithms. In this study, we chose to use the most widely used algorithm of each of those families. (Wierzchoń and Kłopotek 2018).

DBScan

DBSCAN (density-based spatial clustering of applications with noise) (Ester et al. 1996) is a density-based data partitioning algorithm to the extent that it relies on the estimated density of clusters to perform partitioning. The DBSCAN algorithm uses 2 parameters: the distance ϵ and the minimum number of MinPts points that must be within a radius ϵ for these points to be considered as a cluster. The input parameters are therefore an estimate of the point density of the clusters. The basic idea of the algorithm is then, for a given point, to retrieve its ϵ-neighbourhood and to check that it contains MinPts points or more. This point is then considered as part of a cluster. We go then through the ϵ-neighbourhood step by step to find all the points in the cluster. The DBSCAN algorithm can be abstracted into two steps. First, find the points in the ε-neighbourhood of every point, and identify the core points with more than minPts neighbours. Second, find the connected components of core points on the neighbour graph, ignoring all non-core points. Assign each non-core point to a nearby cluster if the cluster is a ε-neighbour, otherwise assign it to noise.

Advantages of the DBSCAN algorithms are numerous: DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-means; DBSCAN can find arbitrarily shaped clusters; due to the MinPts parameter, the single-link effect (different clusters being connected by a thin line of points) is reduced; DBSCAN has a notion of noise, and is robust to outliers; parameters MinPts and ε can be set by a domain expert, if the data is well understood. Drawbacks, however can also be listed: DBSCAN is not entirely deterministic, border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data are processed; it is not comfortable with data sets with large differences in densities, because the MinPts-ε combination cannot then be chosen appropriately for all clusters; if the data and scale are not well understood, choosing a meaningful distance threshold ε can be difficult.

The determination of the two parameters ($\varepsilon = 0{,}55$ and $\text{MinPts} = 40$) has been done in order to minimize the number of outliers.

K-means

Partitioning into k-means is a data partitioning method and a combinatorial optimization problem (MacQueen 1967). Given points and an integer k, the problem is to divide the points into k groups, often called clusters, in order to minimize a certain function. We consider the distance from a point to the average of the points of its cluster (called the centroid); the function to minimize is the sum of the squares of these distances.

The initialization is a determining factor in the quality of the results (local minimum). There are two common initialization methods: Forgy's algorithm (Lloyd 1982) and Random Partitioning. Forgy's algorithm assigns the k points of the initial averages to k randomly selected input data. Random partitioning randomly assigns a cluster to each data point and then (pre-first) calculates the initial mean points.

Given an initial set of k means randomly initialized, Lloyd-Forgy's algorithm proceeds by alternating between two steps. First the assignment step is designed to assign each observation to the cluster whose mean has the least distance. This is intuitively the "nearest" mean. Mathematically, the assignment step deals with partitioning the observations according to the Voronoi diagram generated by the means. Then the update step deals with calculating the new means (centroids) of the observations in the new clusters. The algorithm has converged when the assignments no longer change, although it does not guarantee to find the optimum. The k-means method is a fast and simple method for clustering. It is systematically convergent, and offers an easy visualization. On the other hand, the number of clusters k is an input parameter: an inappropriate choice of k may yield poor results. That is why, when performing k-means, it is important to determine the number of clusters in the data set through cross-validation.

Moreover, convergence to a local minimum may produce counterintuitive results. The k-means++ algorithm (Arthur and Vassilvitskii 2006) tackles this by specifying a procedure to initialize the cluster centers before proceeding with the standard k-means optimization iterations.

Kmeans++

The idea behind this method is that the more dispersed the initial k cluster centers are, the better: the first cluster center is uniformly randomly selected from the data points, and then each subsequent cluster center is selected from the remaining data points with a probability proportional to its distance squared from the nearest existing cluster center.

This seeding method significantly improves the final error of the k-means. Although the initial selection in the algorithm takes longer, the k-means part converges very quickly after this seeding and the algorithm actually reduces the computation time.

Moreover, the k-means++ algorithm guarantees an approximation ratio O (log k), with k the number of clusters used. This is a significant improvement over the standard k-means, which can generate clusters that are arbitrarily worse than the optimum (local minima).

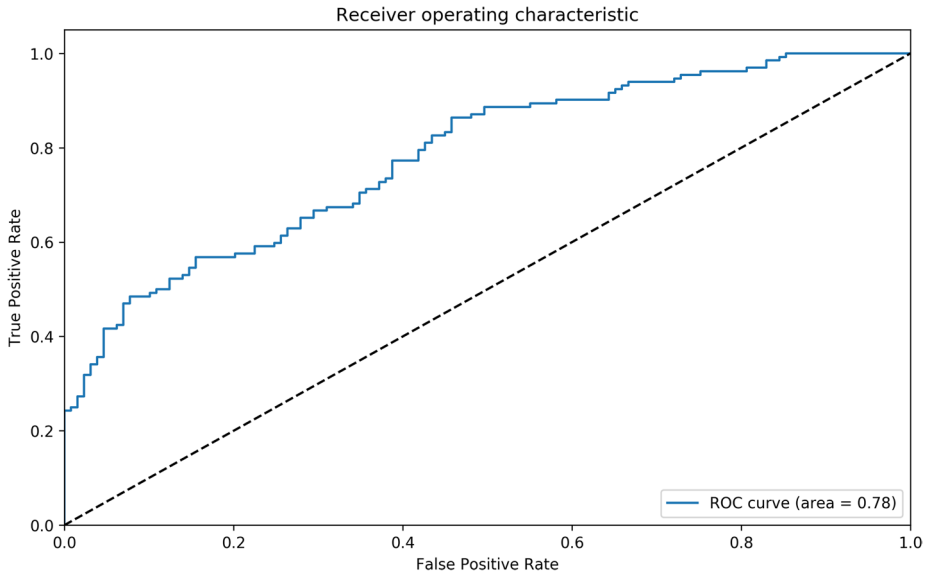

Fig. 2 Receiver operating characteristic

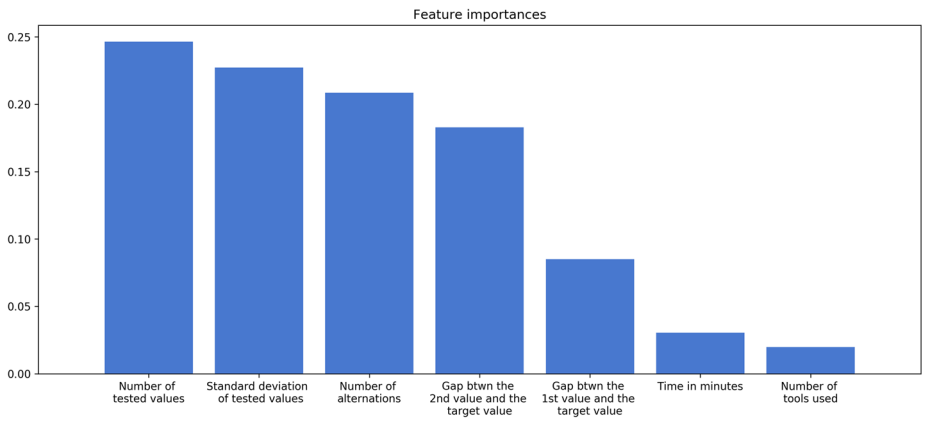

Fig. 3 Feature importance

Results

Supervised learning: detect the significant features

Our first objective is to identify the variables that best explain the student's success. To do this, we will form a model based on the collected data and some secondary data constructed from it. These secondary elements come from the didactic analysis carried out a priori.

After comparing several methods, we chose the random forests. For this method, the area under the ROC curve reaches 0.78 (Fig. 2). This is more than satisfactory for a binary classifier.

The analysis of the variables of importance in this model allows us to select the parameters that are decisive in the student's success in completing the item (Fig. 3).

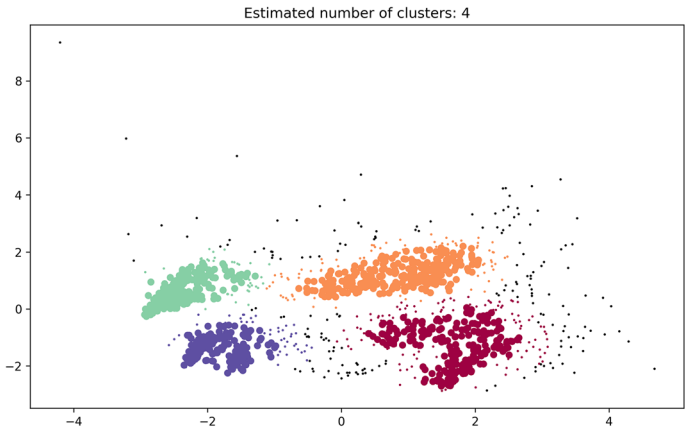

Fig. 4 Result of DBSCAN

The first important characteristic is the number of values tested in the tree data table. The second is the variance of the values tested, i.e. the extent to which these values are concentrated around the target value. In the same spirit, we also find the difference between the first input value and the target value, and also (to avoid the noise of the first value), the difference between the second input value and the target value. In addition, we also find the number of alternations around the target value, a characteristic that expresses the choice of a dichotomous search procedure. Finally, but to a lesser extent, we find the time spent resolving the exercise and the number of tools used. The fact that time is less important for classification than other features, if counterintuitive, is consistent with previous research (Qiao and Jiao 2018).

Clustering using DBSCAN

In parallel with this modeling, we have sought to segment our population to be able to describe the different strategies developed by the students.

To do this, we first chose the DBSCAN algorithm to avoid the difficulty of determining a priori the number of clusters. The Fig. 4 shows the result of this grouping in the space created by the first two dimensions generated by a PCA on the data.

The PCA results are not of importance here. They are simply used in order to allow a graphical projection of the clusters on a two-dimensional space.

The limit of the DBSCAN is the difficulty of processing clusters with very different densities. The DBSCAN avoids this constraint by treating as outliers all values that would cause cluster densities to diverge too widely.

As it is for this dataset, the success rate of the item is 47, 3%, and the 4 clusters have approximately the same size. This well balanced result might not be very surprising considering density based algorithm behavior with average difficult items. Let us suppose that we were working on an item with a more drastic success rate, very easy or very difficult. In that case, the profiles we are looking for would have been of very different densities, with one or two large clusters and other small or sparse ones. The DBSCAN might have failed in detecting all the clusters, and would have labeled the small ones as outliers. Or worse, the algorithm would have merged them into the bigger ones.

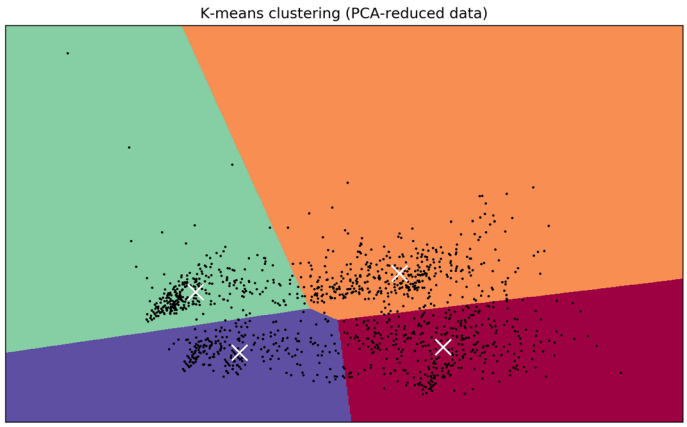

Fig. 5 K-means clustering (PCA-reduced data)

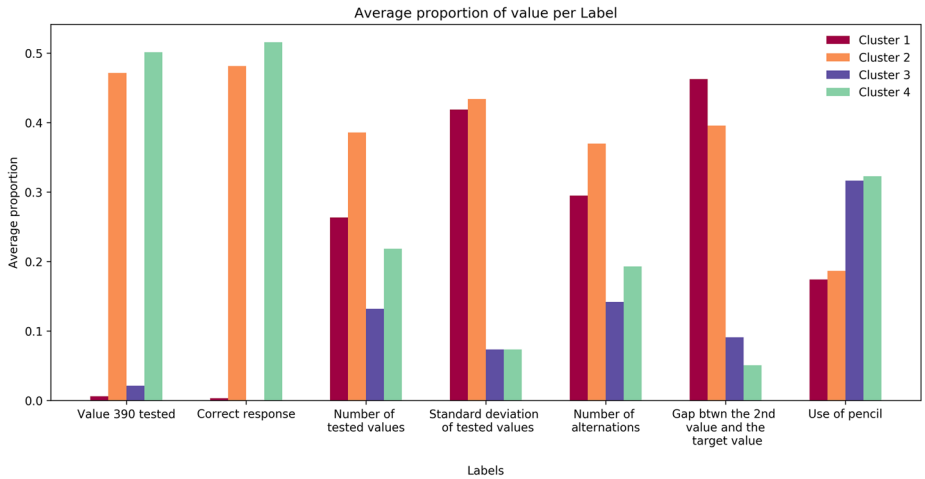

Fig. 6 Average proportion of value per Label

Using a k-means method to assess the clustering

This drawback of the DBSCAN is of no concern for the centroid-based algorithms, such as the k-means. By trying to minimize the distance between each point and the centroid of its cluster, this method allows for clusters with different shapes, sizes and densities.

By choosing $k = 4$, we are observing the stability of the cluster distribution. The density is not a determining factor for the k-means clustering and all observations are to be assigned to a cluster. Therefore, if the cluster distribution we obtain is comparable to the one from DBSCAN, it tells us that a 4-cluster partition with varying densities would not be a better partition.

We will build 4 clusters in the same space constituted by the first two dimensions of the PCA, and compare it to the clustering offered by the DBSCAN.

The distribution of observations is very close to the one resulting from the DBSCAN (Fig. 5). We will therefore use the segmentation from the DBSCAN to try to identify each one of the clusters.

Characterization of the 4 categories of students

To do this, we will calculate the reduced average proportion in the cluster for each important characteristic (Fig. 6).

We will also calculate the reduced average proportion for three additional characteristics. First, those who express success or failure (is the value correct? Has the correct value been entered in the spreadsheet?), and second, characteristics raised by the didactic analysis (has the graphic tool "Pencil" been used by the student?).

The comparison between the clusters of these small average proportions shows very clearly that two axes of division can be distinguished.

First, the model identifies successful and unsuccessful students. As expected, the discriminating variables between these two groups are the presence of the "correct answer" and whether the target value was entered in the "calculation and graph" tool.

Second, the model identifies two approaches to the problem. Students in clusters 3 and 4 used the pencil a lot, recorded few values in the "calculation and graph" tool, and these values were highly concentrated around the target value. In contrast, students in clusters 1 and 2 used less the pencil and entered many values. The values were quite scattered and from a certain distance from the target value. These students were also more often alternating between lower and higher values.

Discussion

We can argue that direct analysis of student process data on this interactive item is a powerful tool to determine not only the student's success in completing the item, but also the strategy he or she uses to solve it. This log analysis confirms also the strategies discerned by the didactical analysis.

Clustering analysis distinguishes 4 clusters corresponding to 4 different students' profiles. Each cluster's size represents 25% of the responding students. Two profiles (green and orange on the graph) achieved the task. The other two (blue and red) correspond to students who failed. Apart from the obvious variable "value 390 tested", no other variable allows to discriminate between success and failure. This result is disappointing in the sense that the model cannot explain students' achievement on the item, which is one of our research questions. However, variables of interest in the a priori analysis allow describing profiles along a dimension other than achievement. Two clusters (orange and red on the Fig. 2) show a large number of inputs, a large number of alternating inputs, a first input far from target and a large distribution of inputs. These characteristics allow us to interpret that students from these groups preferred a "trial and error" solving strategy, approaching the underlying concept of function in its operational aspect. Half of them achieved the task successfully, the other half did not. The other two groups share a different and opposite profile description according to the same variables: a small number of inputs, a small number of alternating inputs, a first input close from target, a narrow distribution of inputs. Moreover this second category of students used the pencil more often, altogether identifying strategies related to the structural aspect of functions. Like the first two groups, the structural approach led to failing the task for half of the students favoring it.

Hence, if the didactical and analytical models used in this study could not help us explain grade 9 students' achievement to this item, they could nevertheless help us

identify two solution strategies well known in didactics literature. From a curriculum designer perspective, this kind of result on a national level could be very valuable. It could help decision making based on the evidence of the manifestation of both structural and operational understanding among grade 9 students. Of course, the analysis should be replicated and extended to a full set of items potentially aiming at discriminating between the two conceptions within the variation and relationships domain or even in other domains. Furthermore, as it is strongly based on research findings in didactics, such result can be fruitfully disseminated towards subject specialists such as policy makers and stake-holders in charge of teacher training, contributing to address better the use of large-scale assessment findings at the classroom level.

However, there are limitations to relying on log data analysis in order to determine learning profiles. One of them lies in the fact that, even if carrying a lot of information, this information remains partial when it comes to describing the complete cognitive activity at play when solving the item. Quite a large amount of "idle time" is usually recorded in the logs. Log data alone is unable to help us understand what students did during this time. Moreover, there is a strong chance that students are not that "idle" when no activity is logged within the assessment system: a lot can happen within a classroom, even during a standardized test administration. The use of a scratch paper or a personal calculator, as well as interactions between students or with the test administrator, are sometimes critical within the problem-solving process. Taking these external factors into account, in addition to the log data, would contribute to addressing fully the question of the interactive items' validity in terms of capturing learning strategies. Complementary research, focusing on user experience, would then consist of collecting and analyzing data from actual test administration observations, possibly enriched with eye-tracking technology or "think aloud" recordings and case studies.

The potential of technology-enriched problem solving situations in large-scale assessments is obvious. The type of findings this study produced is actionable from teachers' perspective in order to address students' specific needs. The DEPP is currently designing and implementing a census based assessment in mathematics at the beginning of grade 10. Its main objective is to report individual profiles in mathematics, consisting of a score in various mathematical sub-domains. Being able to deliver additionally a qualitative profile based on students' strategic behavior solving technologically-enhanced problems would add value to the national feedback and help teachers to better support students in their learning. Assuming they benefit from consistent and didactic-based training, teachers would be able to differentiate teaching in the classroom according to the cognitive profile of each student provided by the national testing platform. A lot needs to be achieved before being able to devise this kind of assessment instrument. One obstacle, in particular, to the generalization of the present study is the fact that analysis depends on the studied item's characteristics. The feature engineering and construction of new variables, for instance, cannot be exactly replicated for another item, due to the fact that it depends on the specific tool available in the item. Therefore, this research is not easily reproducible on a different set of items or other situations. Developing an experimental methodology for each situation raises a lot of issues regarding large-scale assessment constraints, but this very promising first step shows certainly the need for further research and new partnerships. The DEPP is now exploring ways to industrialize

its methods in order to run automatic log data analysis. More specifically, it is engaging partnerships to address research objectives regarding the relationship between problem solving strategies and achievement to a mathematic test rather than on a single item. Among these objectives we are investigating whether lower-achieving/higher-achieving students consistently adopt one strategy over another, whether higher-performing students adapt strategies to the task and whether the choice of strategy is part of the competency.

Acknowledgements
We thank Victor Azria and Stéphane Germain of CapGemini France for supporting the statistical analysis. We thank the SCRIPT of the Ministry of National Education, Children and Youth of Luxembourg and Vretta Inc. for supporting the technology-enhanced items' development.

Authors' contributions
FS contributed to the design of the theoretical framework. SK and RDS contributed to the statistical analysis. All authors read and approved the final manuscript.

Funding
The DEPP is a department of the ministry of education, therefore publically funded.

Availability of data and materials
Availability of data and materials is contingent on specific agreement between the ministry of education in France and interested research institutions.

Competing interests
The authors declare that they have no competing interests.

Received: 10 January 2020 Accepted: 19 May 2020
Published online: 29 May 2020

References

Arthur D., Vassilvitskii S. (2006) k-means++: The advantages of careful seeding, *ilpubs.stanford.edu*
Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 1–22.
Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.
Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Brousseau, G. (2006). *Theory of didactical situations in mathematics: Didactique des Mathématiques, 1970–1990*. Berlin: Springer Science & Business Media.
Doorman, M., Drijvers, P., Gravemeijer, K., Boon, P., & Reed, H. (2012). Tool use and the development of the function concept: from repeated calculations to functional thinking. *International Journal of Science and Mathematics Education, 10*(6), 1243–1267. https://doi.org/10.1007/s10763-012-9329-0.
Douady, R. (1991). Tool, Object, Setting, Window : Elements for Analysing and Constructing Didactical Situations in Mathematics. In A. J. Bishop, S. Mellin-Olsen, & J. Van Dormolen (Éd.), *Mathematical Knowledge : Its Growth Through Teaching* (p. 107–130). Springer Netherlands. https://doi.org/10.1007/978-94-017-2195-0_6
Drijvers, P. (2019). Digital assessment of mathematics: opportunities, issues and criteria. *Mesure et Évaluation En Éducation, 41*(1), 41–66. https://doi.org/10.7202/1055896ar.
Drijvers, P., Ball, L., Barzel, B., Heid, M. K., Cao, Y., & Maschietto, M. (2016). *Uses of digital technology in lower secondary mathematics education*. Berlin: Springer.
Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd, 96,* 226–231.
Gobert, J. D., Pedro, M. S., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521–563. https://doi.org/10.1080/10508406.2013.837391.
Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: an analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61,* 36–46. https://doi.org/10.1016/j.chb.2016.02.095.
Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers and Education, 91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018.
Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: an edit distance approach. *Journal of Educational Data Mining, 7*(1), 33–50.
Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). Taming log files from game/simulation-based assessments: data models and data analysis tools: taming log files from game/simulation-based assessments. *ETS Research Report Series, 2016*(1), 1–17. https://doi.org/10.1002/ets2.12096.

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. *Quantitative psychology research* (pp. 173–190). Berlin: Springer.
Hopcroft, J. E., Ullman, J. D. (1983). Data structures and algorithms.
Kilpatrick, J., Swafford, J., Findell, B., National Research Council (U.S.), & Mathematics Learning Study Committee. (2001). Adding it up : Helping children learn mathematics. National Academy Press. http://site.ebrary.com/id/10038695
Laborde, C. (2018). Intégration des technologies de mathématiques dans l'enseignement. In *Guide de l'enseignant. Enseigner les mathématiques.* (Belin, pp. 336–366). https://publimath.irem.univ-mrs.fr/biblio/PGE18015.htm
Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137.
MacQueen, J., & others. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297. Oakland.
MEN, DEPP. (2017). *Cedre 2014: Mathématiques en fin de collège* (No. 209). Retrieved from Ministère de l'éducation nationale website: https://www.education.gouv.fr/cid122693/cedre-2014-mathematiques-en-fin-de-college.html
OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD.
Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: a didactic. *Frontiers in Psychology, 9,* 2231.
Rabardel, P. (2002). *People and technology: A cognitive approach to contemporary instruments.* (Université Paris 8).
Robert, A. (1998). Outils d'analyses des contenus mathématiques à enseigner au lycée et à l'université, Recherches en didactique des mathématiques, Vol 18 2 pp. 139–190. *Recherches En Didactique Des Mathématiques*, *Vol 18 2*, 139–190.
Roditi, E., Salles, F. (2015). Nouvelles analyses de l'enquête PISA 2012 en mathématiques, un autre regard sur les résultats. *Revue Éducation et formations*, (n° 86–87), 24.
Sfard, A. (1991). On the dual nature of mathematical conceptions: reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*(1), 1–36.
Stacey, K., & Wiliam, D. (2013). Technology and assessment in mathematics. In M. A. (Ken) Clements, A. J. Bishop, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Third International Handbook of Mathematics Education* (p. 721–751). https://doi.org/10.1007/978-1-4614-4684-2_23
Trouche, L. (2003). From artifact to instrument: mathematics teaching mediated by symbolic calculators. *Interacting with Computers, 15*(6), 783–800. https://doi.org/10.1016/j.intcom.2003.09.004.
Wierzchoń, S. T., & Kłopotek, M. A. (2018). Cluster analysis. In S. Wierzchoń & M. Kłopotek (Eds.), *Modern Algorithms of Cluster Analysis* (pp. 9–66). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-69308-8_2.

Publisher's Note