Unsupervised quantitative methods to analyze student reasoning lines: Theoretical aspects and examples

Onofrio Rosario Battaglia,[1,*] Benedetto Di Paola,[2] and Claudio Fazio[1]

[1]*UOP_PERG (University of Palermo Physics Education Research Group)*
Dipartimento di Fisica e Chimica - Emilio Segrè, Università degli Studi di Palermo, Palermo 90128, Italia
[2]*GRIM (Mathematics Education Research Group)*
Dipartimento di Matematica e Informatica, Università degli Studi di Palermo, Palermo 90123, Italia

(Received 31 August 2018; published 3 July 2019)

[This paper is part of the Focused Collection on Quantitative Methods in PER: A Critical Examination.] A relevant aim of research in education is to find and study the reasoning lines that students deploy when dealing with problematic situations. This can be done through an analysis of the answers students give to a questionnaire. In this paper, we discuss some methodological aspects involved in the quantitative analysis of a questionnaire by means of two different clustering methods, a hierarchical one and a nonhierarchical one. We start from the coding procedures needed to obtain analyzable data from the questionnaire and from a definition of a correlation coefficient suitable for measuring student similarity in the case of binary coding of student answers. Then, criteria to choose the optimal number of clusters are discussed, and for the same purpose a new coefficient is introduced that measures the total amount of information we can obtain from a clustering solution. We show that each cluster can be characterized by its centroid that summarizes the answers most frequently given by the cluster students to the questionnaire. Finally, an example of the application of these procedures to a student sample is given, and a comparison between the two clustering methods is discussed.

DOI: 10.1103/PhysRevPhysEducRes.15.020112

I. INTRODUCTION

Many quantitative research studies in various fields of mathematics and physics discuss the use of student answers to questionnaires to obtain information about the reasoning lines[1] student deploy when dealing with problematic situations and to investigate students' conceptual understanding [6–9].

Other studies [10–12] identify student reasoning profiles with the aim of making explicit the possible different ways conceptual understanding is activated in students by analyzing the answers mostly frequently given by students to questions, and their relationships. These studies find groups of students that are homogeneous, or "similar," with respect to the ways in which they respond to the questionnaire. However, to clearly separate a sample of students into groups so that the elements of each group are similar to each other while being substantially different from elements in other groups can be a complex operation, especially for samples composed of many students. Cluster analysis (CLA) is one of the methodologies used for this purpose.

CLA techniques [13] are common in many fields of research, such as information technology, biology, medicine, archeology, econophysics, and market research [14–17]. These techniques allow the researcher to locate subsets, or clusters, within a set of objects of any nature, that have a tendency to be homogeneous "in some sense," without any prior knowledge of what forms those groups take (unsupervised classification [18–20]). The results of the analysis can reveal a high homogeneity within each group (intracluster) and high heterogeneity between groups (intercluster), in line with the chosen criteria.

CLA, introduced in psychology by Tyron in 1939 [21], saw its first systematic use by Sokal and Sneath [22] in 1963. Some studies using CLA methods can be found in the literature concerning research in education. They group and characterize students' responses by using open-ended questionnaires [10–11,23] or multiple-choice tests [12,23].

*onofriorosario.battaglia@unipa.it

[1]People's reasoning is often described as the "running" of the procedures present in their mental models [1–4]. Gilbert and Boulter [5] define *expressed models* as the external representations expressed by an individual through actions, speech, or writing. We understand "path, or line of reasoning" as the external representation of the mental models used by an individual when they try to describe, predict, or explain the physical world.

All these papers show that the use of cluster analysis leads to identifiable groups of students that make sense to researchers and are consistent with previous results obtained using more traditional methods. Particularly, Fazio *et al.* [10,11] and Pizzolato *et al.* [12] analyze students' responses to specially designed written questionnaires using researcher-generated categories of reasoning, based on the physics education research literature on student understanding of relevant physics content. Through cluster analysis methods, groups of students showing remarkable similarity in their reasoning categories are identified, and the consistency of their deployed mental models is validated by comparison with researcher-built ideal profiles of student behavior known from previous research. Springuel *et al.* [23] identify groups of responses in open-ended questions about two-dimensional kinematics by means of cluster analysis. These groups show striking similarity to response patterns previously reported in the literature and also provide additional information about unexpected differences between groups.

Ding and Beichner [24] study five commonly used approaches to analyzing multiple-choice test data (classic test theory, factor analysis, cluster analysis, item response theory, and model analysis) and show that cluster analysis is a good method for showing how student response patterns differ so as to classify students.

CLA can be carried out using various algorithms and techniques that differ significantly in their notion of what constitutes a cluster and how to effectively find them. However, the various techniques have seldom been closely explored and compared when applied on the same student sample, to reveal their mutual coherence and points of strength and weakness. Moreover, the criteria to find the best clustering solution among all possible ones as well as the choice of criteria of similarity between students, the choice of clustering algorithms, the individuation of the groups to be obtained and the evaluation of the solution found have been under-explored, especially in the educational field, and require further study.

For this reason, in this paper we start from an analysis of the data setup needed by CLA. Then, two methods commonly used in CLA are described and the variables and parameters involved are outlined and criticized. Section VIII deals with an example of the application of these methods to the analysis of data from the answers to an open-ended questionnaire administered to a sample of high school students, and discusses the significance of information that can be obtained by using the two different clustering methods. Finally, a comparison of the results is done in order to reveal and discuss their coherence.

II. CLASSIFICATION OF STUDENT ANSWERS AND DATA CODING

A cluster analysis of student answers to a questionnaire requires, as a first step, a classification of the answers. While in a closed-ended questionnaire the answers themselves can be considered as categories, the analysis of an open-ended questionnaire should start from the categorization of answers into a limited number of the "typical" ways students tackle each question. However, it is well known that there are inherent difficulties in the classification and coding of student responses. Hammer and Berland [25] point out that researchers "*should not treat coding results as data but rather as tabulations of claims about data and that it is important to discuss the rates and substance of disagreements among coders*" and proposes guidelines for the presentation of research that "quantifies" individual student answers. Among such guidelines, they focus on the need to make explicit that "*developing a coding scheme requires researchers to articulate definitions of categories well enough that others could interpret them and recognize them in the data*". Chi [26] describes the process of developing a coding scheme in the context of verbal data such as explanations, interviews, problem-solving protocols and retrospective reports. The method of verbal analyses is deeply discussed with the objective of formulating an understanding of the representation of the knowledge used in cognitive performances.

Following the approach previously described [25,26], the logical steps that should be used to process data coming from student answers to an open-ended questionnaire can be synthesized by the flow chart represented in Fig. 1.

The first and second steps (categorization and comparison by the k researchers involved in the study) involve the analysis of the records representing student answers

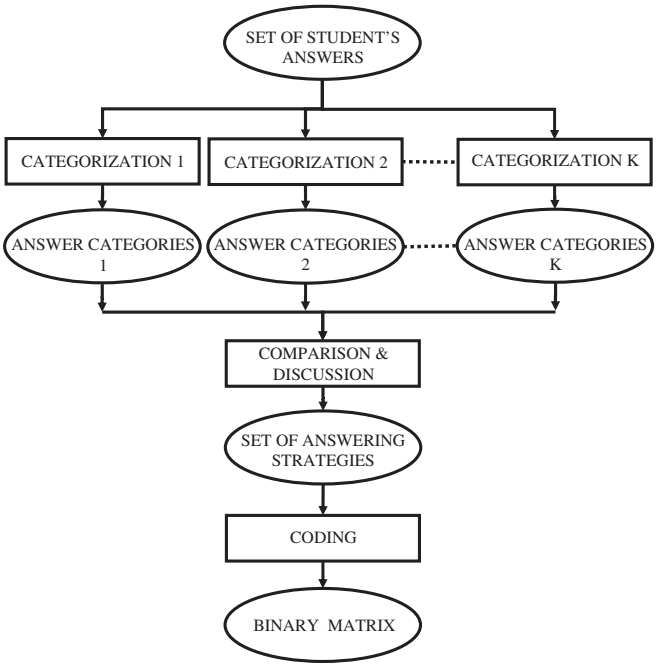

FIG. 1. Flow chart of the steps that can be followed by k researchers when processing data coming from student answers to an open-ended questionnaire.

(the data), in order to reveal patterns and trends, and to find common themes emerging from them. Through comparison and discussion among researchers, these themes are then developed and grouped in a number of categories whose definition takes into account as much as possible the words, the phrases, the wording used by students [26]. Such categories can be considered the typical answers of the N students to the questionnaire. The researchers involved in the study independently [27,28] read the students' answers in order to empirically identify the main characteristics of the different student records (the raw data). Each researcher constructs a coding scheme consisting in the identification of keywords, which characterize the student answers. During a first meeting, the selected keywords are compared and contrasted, and then grouped into "categories" (i.e., the typical answers given by students to the questions) based on epistemological and linguistic similarities.[2] These categories (e.g., see Appendix) are also re-analyzed through the researchers' interactions with the data, and take into account the existing educational research literature on the questionnaire topics.

At the end of this answer categorization phase, the whole set of answers given by students to the open-ended questionnaire is grouped into a limited number M of typical answers, which we call the "answering strategies." that the students deploy when tackling the questionnaire. M is obtained by combining all the answering strategies used by N students when answering each question.

The following phase is the same for both closed-ended and open-ended questionnaires, and involves the binary coding of student answers[3] according to the defined answering strategies, generating a binary matrix (as shown in Table I). So, through categorization (if needed) and coding, each student i can be identified by an array a_i composed of M components 1 or 0, where 1 means that the student used a given answering strategy or answer option to respond to a question and 0 means that they did not use it. Then, an $M \times N$ binary matrix (the "matrix of answering strategies"), modeled on the one shown in Table I, is built. Its columns show the N student arrays a_i and the rows represent the M components of each array, i.e., the M answering strategies or answer options.

For example, let us say that student S_1 used answering strategies AS_2, AS_3, and AS_5 to respond to the questionnaire questions. Therefore, column S_1 in Table I will contain the binary digit 1 in the three cells corresponding to these strategies, while all the other cells will be filled with 0.

TABLE I. Matrix of the data: the N students and the M answering strategies are denoted as $S_1, S_2, \ldots, S_N$., and as $AS_1, AS_2, \ldots, AS_M$, respectively.

Answering strategy	Student			
	S_1	S_2	…	S_N
AS_1	0	0	…	0
AS_2	1	0	…	1
AS_3	1	…	…	…
AS_4	0	…	…	…
AS_5	1	…	…	…
…	0	…	…	…
AS_M	0	1	…	0

It is worth noting that, independently of the open- or closed-ended nature of a questionnaire, actions like altering which questions are included in the analysis, or the weight of the question in the metric can, in principle, have a large impact on the clustering. Both the best partition of the sample (i.e., the optimal number of clusters found by the clustering procedure) and sizes, shapes, and population of clusters can be strongly influenced by simply adding or removing a question from the analysis. For example, let us consider a questionnaire in which different groups of questions are aimed at investigating different aspects of a given theme [e.g., the force concept in the well-known Force Concept Inventory (FCI) questionnaire]. In order to study only one of these aspects there are two options: (i) to do a cluster analysis of the whole questionnaire and then study only the answers related to the aspect one wants to investigate, or (ii) to do only a cluster analysis of the answers to the questions strictly related to that aspect. However, the results of the two clustering procedures will in principle be different, as we observed in another research [29].

III. CORRELATION COEFFICIENT FOR BINARY DATA AND SIMILARITY INDEX

The matrix in Table I contains all the information needed to describe the sample behavior according to the previously described categorization. However, it needs some elaboration to be used for CLA. Particularly, CLA requires the definition of new quantities that are used to build the grouping, such as "similarity" or "distance" indexes. These indexes are defined by starting from the $M \times N$ binary matrix discussed above.

In the literature [13,16,21], the similarity between two students i and j of the sample is often expressed by taking into account the distance d_{ij} between them (which actually expresses their "dissimilarity," in the sense that a higher value of distance involves a lower similarity).

[2]For example, in the analysis of the answers to six open-ended questions on the concept of models and modeling discussed in Sec. VIII as an application, students that defined models as simple phenomena or experiments or reproductions of an object on a small scale have been put on the same category since the three definitions have been intended as giving an ontological reality to models.

[3]For simplicity here we refer to the use of a two-level coding, where 1 means that a given answering strategy or answer option was used, and 0 means that strategy was not used.

A distance index can be defined by starting from Pearson's correlation coefficient. This allows the researcher to study the correlation between students i and j if the related variables describing them are numeric. If these variables are non-numeric (as in our case, where we are dealing with arrays a_i and a_j containing a binary symbolic coding of the answers of students i and j, respectively), we must use a modified form of Pearson's correlation coefficient $R_{\text{bin}}(a_i, a_j)$, similar to that defined by Tumminello *et al.* [30]. We define it as[4]

$$R_{\text{bin}}(a_i, a_j) = \frac{C(a_i, a_j) - p(a_i)p(a_j)/M}{\sqrt{p(a_i)p(a_j)\left(\frac{M-p(a_i)}{M}\right)\left(\frac{M-p(a_j)}{M}\right)}}, \quad (1)$$

where $p(a_i)$, $p(a_j)$ are the numbers of 1's in the arrays a_i and a_j, M is the total number the answering strategies, and $C(a_i, a_j)$ is obtained by counting how many times the symbol 1 is present in the same position in the arrays a_i, and a_j. $[p(a_i)p(a_j)]/M$ is the expected value[5] of $C(a_i, a_j)$.

By following Eq. (1) it is possible to find for each student i the $N-1$ correlation coefficients $R_{\text{bin}}(a_i, a_j)$ between them and the other students (and the correlation coefficient with themselves, which is clearly 1). All these correlation coefficients can be placed in an $N \times N$ matrix that contains the information we need to consider the mutual relationships between our students.

The similarity between students i and j can be defined by choosing a metric to calculate the distance d_{ij}. Such a choice is often complex and depends on many factors. If we want two students, represented by arrays a_i and a_j and negatively correlated, to be more dissimilar than two uncorrelated students, a possible definition of the distance between a_i and a_j, making use of the modified correlation coefficient $R_{\text{bin}}(a_i, a_j)$ is

$$d_{ij} = \sqrt{2[1 - R_{\text{bin}}(a_i, a_j)]}. \quad (2)$$

This function defines a Euclidean metric [32] that is required for further calculations. A distance d_{ij} between two students equal to zero means that they are completely similar [$R_{\text{bin}}(a_i, a_j) = 1$], while a distance $d_{ij} = 2$ shows that the students are completely dissimilar [$R_{\text{bin}}(a_i, a_j) = -1$]. When the correlation between two students is 0, their distance is $d_{ij} = \sqrt{2}$.

By following Eq. (2) it is then possible to build a new $N \times N$ matrix, D (the distance matrix), containing all the mutual distances between the students. The main diagonal of D is composed of 0's (the distance between a student and themself is zero). Moreover, D is symmetrical with respect to the main diagonal.

IV. K-MEANS ALGORITHM

Nonhierarchical clustering (NH-CLA) methods partition the data space into a number of nonoverlapping subsets (clusters) containing data similar to each other according to given criteria. Among the currently used NH-CLA algorithms, we will consider the k-means one [15], proposed by MacQueen in 1967 [33], as it is well known, easy to implement in computer code, and computationally efficient.

The starting point is the choice of the number q clusters one wants to populate and of an equal number of "seed points." The data (students) are then grouped on the basis of the minimum distance between them and the seed points. Starting from an initial classification, students are iteratively attributed from one cluster to another one, until no further improvement can be made. The students belonging to a given cluster are used to find a new point representing the average position of their spatial distribution. This is done for each cluster Cl_k ($k = 1, 2, \ldots, q$), and the resulting points are called the cluster *centroids* C_k. This process is repeated and ends when the new centroids coincide with the old ones.

It is worth noting that the data input of the k-means algorithm could be the $M \times N$ binary matrix. However, a formally correct application of this algorithm strictly requires the use of a Euclidean metric, that cannot be used for binary data [34]. For this reason, it is necessary to transform the initial binary data. For this purpose, a procedure well known in the specialized literature as *multidimensional scaling* [34] can be used. For each student i we know the N distances d_{ij} between such a student and all the students of the sample ($d_{ii} = 0$). The multidimensional scaling procedure is applied to each student and starts from these N distances. It consists in a linear transformation between two vector spaces (from an N-dimensional vector space to a two-dimensional one). In summary, this procedure allows us to associate to each

[4]Equation (1) is formally similar to the similarity index used in Refs. [10–12]. However, our equation is a version of Pearson's correlation coefficient adapted to the case of non-numerical variables while the other is an index, defined by Lerman [31], that defines the similarity between two elements in a probabilistic form and can be directly used to partition a data sample.

[5]If we assume that each component of the array can be 1 with equal probability, the probability that two arrays have a 1 in common in the same component is given by 1 over the number of possible answers. So, the greater the number of possible answers, the smaller the probability will be. Moreover, if we consider more questions simultaneously, the number of possible answers in common between two students increases with the number of questions.

M-dimensional binary vector a two-dimensional vector composed of real numbers. The k-means algorithm is, therefore, applied on two-dimensional data.[6]

We remark here that many types of metrics are known in the literature, and the choice of a specific one to use with a given type of clustering algorithm depends strongly on both the type of data and the type of clustering algorithm itself. Different metrics may affect the size and members of a cluster as they imply the use of different approaches to find the distance between the data objects, which is the most important step of creation of clusters [35]. So, the k-means algorithm can be applied independently of the dimension of the space in which the data are represented, but an appropriate metrics is to be chosen wisely and according to the dataset.

Euclidean metrics are those most commonly used by nonhierarchical clustering methods, as the k-means one, as we said, cannot be directly used on binary data. The use of different, more appropriate metrics, such as, for instance, the correlation, Jaccard, or Hamming [36] ones, with a given clustering algorithm can lead to results that can be significantly different from each other even if the same algorithm is used. For example, metrics such as the Hamming or Jaccard ones, that can only be used with binary data, give inconsistent results when used with the k-means algorithm, like the inclusion of members with very different characteristics in the same cluster [37].[7]

We can conclude that using a k-means-like algorithm with a non-Euclidean metric (i.e., choosing the seed point in the M-dimensional space) likely introduces inconsistencies in the clustering results. This does not happen with the correlation metric, but unfortunately it does with Hamming or Jaccard ones.

Finally, the k-means results can be plotted in a two-dimensional Cartesian space (see Fig. 1), similar to a Voronoi diagram [38], containing points that represent the students of the sample placed in the plane according to their mutual distances. A cluster centroid can, therefore, be considered, from a geometrical viewpoint, as the average position of all the points (the students) in the cluster. The x and y axes of the Cartesian plot simply report the values needed to place the points according to their mutual distance.

We want now to show that the centroid possesses another relevant property. We start by defining a "virtual student" for each of the q cluster centroids. Since each student is characterized by an array a_i composed by the 0 and 1 values for each of the M answering strategies, the array for the virtual student $\bar{a}_k$ should also contain M entries with 0's for strategies that are not used by "him or her" and 1 for strategies that are used. It is possible to demonstrate that $\bar{a}_k$ contains 1 values exactly in correspondence to the answering strategies most frequently used by students belonging to Cl_k ($k = 1, 2, .., q$). In fact, since a centroid is defined as the geometric point that minimizes the sum of the distances between it and all the cluster elements, by minimizing this sum, the correlation coefficients between the cluster elements and the virtual student are maximized, and this happens when each virtual student has the greatest number of common strategies with all the students that are part of its cluster. This is a remarkable feature of the centroid that makes it able to characterize the cluster also from a pedagogical point of view.

Therefore, in order to find the array that describes the centroid, we can simply search for the answering strategies most frequently used by the students in cluster Cl_k. Another way to demonstrate that the centroid array is actually composed as previously described is to start from the coordinates of the centroid in the two-dimensional Cartesian space reporting the results of the k-means algorithm. We repeat the k-means procedure in reverse, by using the iterative method described as follows. For each cluster Cl_k we define a random array $\bar{a}'_k$ (composed of values 1 and 0, randomly distributed) and we calculate the following value:

$$\eta = \sum_i |d_{ik} - d'_{ik}|, \tag{3}$$

where d'_{ik} is the distance between the random array and the student i (belonging to the same cluster Cl_k) and d_{ik} is the distance between the centroid and the same student. The iterative procedure permutes the values of the random array $\bar{a}'_k$, minimizes the η value, and finds the closest array representation[8] $\bar{a}_k$ of the real centroid of C_k. The final results confirm that $\bar{a}_k$ is made up of the

[6] In order to verify that this procedure does not lead to wrong results, a check is to be made, e.g., applying a k-means-like algorithm directly to the initial binary matrix, defined in the M-dimensional space (using a metrics appropriate to the data, like the correlation one). We did so on the data that are discussed in Sec. VIII and we obtained exactly the same clusters found applying the k-means to two-dimensional data, showing that no errors are introduced in our procedure.

[7] For example, in the analysis of data discussed in Sec. VIII, when metrics such as Hamming or Jaccard were used in the application of the k means on M-dimensional data, the results obtained differed significantly from those obtained by using the k-means algorithm on two-dimensional data, i.e., after the multidimensional scaling process. In fact, some of the clusters found with Hamming or Jaccard metrics have been found to be inconsistent, because a detailed analysis performed *a posteriori* closely analyzing the responses of each student showed that students giving clearly different types of answers were included in the same cluster by the k-means-like algorithm performed directly on the binary matrix.

[8] As usual in a procedure to minimize an objective function (in our case, η), the result may not be unique. In order to be sure to obtain an absolute minimum of η, the procedure can be repeated several times, each time changing the initial conditions, i.e., array $\bar{a}'_k$.

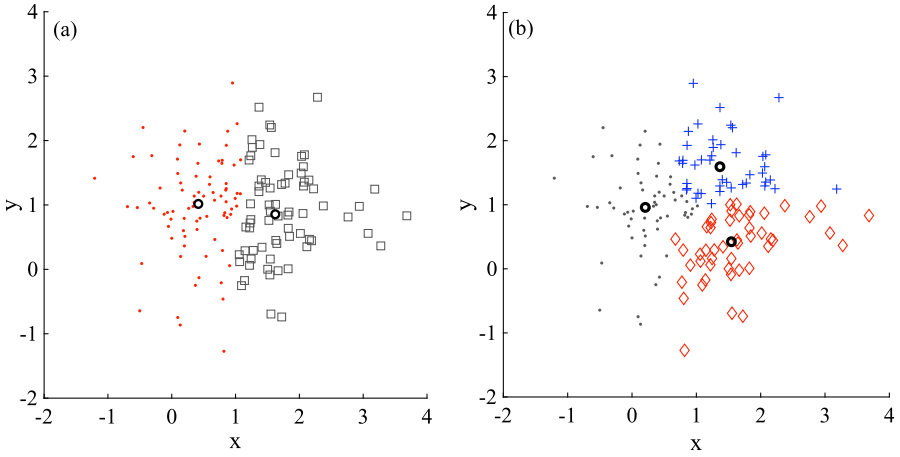

FIG. 2. A set of 150 hypothetical data partitioned into two (a) and three (b) clusters. The mean values of the silhouette function are 0.47 and 0.45, respectively. The x and y axes simply report the values needed to place the points according to their mutual distance.

answering strategies most frequently used by the students in cluster Cl_k.

The k-means algorithm has some points of weakness, and here we will describe how it is possible to overcome them. The first involves the *a priori* choice of the initial positions of the centroids. This is usually resolved [36,39] by repeating the clustering procedure for several values of the initial conditions and selecting those that lead to the minimum values of the distances between each centroid and the cluster elements. Furthermore, at the beginning of the procedure, it is necessary to arbitrarily define the number q of clusters. A method widely used to select this number q, initially used to start the calculations, as the one that best fits the sample element distribution is the calculation of the so-called *silhouette function*, S [40,41].

A. The silhouette function

In order to choose the number q of clusters to be initially used to perform the calculations, the silhouette function, S [40,41] is defined. This function allows us to decide if the partition of our sample into q clusters is adequate.[9]

For each selected number of clusters q and for each sample student i assigned to a cluster k, with $k = 1, 2, \ldots, q$, the value of the silhouette function $S_i(q)$ is calculated as

$$S_i(q) = \frac{\min_{p,p\neq k}[\sum_{l=1}^{N-n_k} d_{il}/(N-n_k)] - \sum_{j=1}^{n_k} d_{ij}/n_k}{\max\{\sum_{j=1}^{n_k} d_{ij}/n_k, \min_{p,p\neq k}[\sum_{l=1}^{N-n_k} d_{il}/(N-n_k)]\}}, \tag{4}$$

where the first term of the numerator is the average distance of the ith student in cluster k to the lth student placed in a different cluster p ($p = 1, \ldots, q$), minimized over clusters. The second term is the average distance between the ith student and another student j placed in the same cluster k.

$S_i(q)$ gives a measure of how similar student i is to the other students in its own cluster, when compared to students in other clusters. It ranges from -1 to $+1$: a value near $+1$ indicates that student i is well matched to its own cluster and poorly matched to neighboring clusters. If most students have a high silhouette value, then the clustering solution is appropriate. If many students have a low or negative silhouette value, then the clustering solution could have either too many or too few clusters (i.e., the chosen number q of clusters should be modified).

Subsequently, the values $S_i(q)$ can be averaged over each cluster k finding the values $\langle S(q)\rangle_k$, and on the whole sample finding the total average silhouette value $\langle S(q)\rangle$ for the chosen clustering solution. Large values of $\langle S(q)\rangle_k$ mean that (on average) cluster k elements are tightly arranged in the cluster and/or are clearly distinct with respect to elements of the other clusters [40,41]. Similarly, large values of $\langle S(q)\rangle$ relate to the existence of well-defined cluster solutions [40,41]. It is, therefore, possible to perform several repetitions of the cluster calculations (with different values of q) and to choose the number of clusters q that give the maximum value of $\langle S(q)\rangle$. It has been shown [42] that for values of $\langle S(q)\rangle < 0.50$, reasonable cluster structures cannot be identified in the data. If $0.51 < \langle S(q)\rangle < 0.70$, the data set can be reasonably partitioned into clusters, and values of $\langle S(q)\rangle$ greater than 0.70 show a strong cluster structure of the data. Figure 2 shows a partition of a hypothetical data set made up of 150 elements into two [Fig. 2)] and into three [Fig. 2)] clusters. It is easy to see that in both cases, a partition into clusters is not easily found, and this is confirmed by the low values of $\langle S(q)\rangle$ in each of the two partition attempts.

Figure 3 shows an example of the spatial distribution of the results of a k-means analysis on another hypothetical set

[9]By "adequate partition" here we mean a partition in which clusters are clearly distinct from each other and compact. We will better address this point in the following.

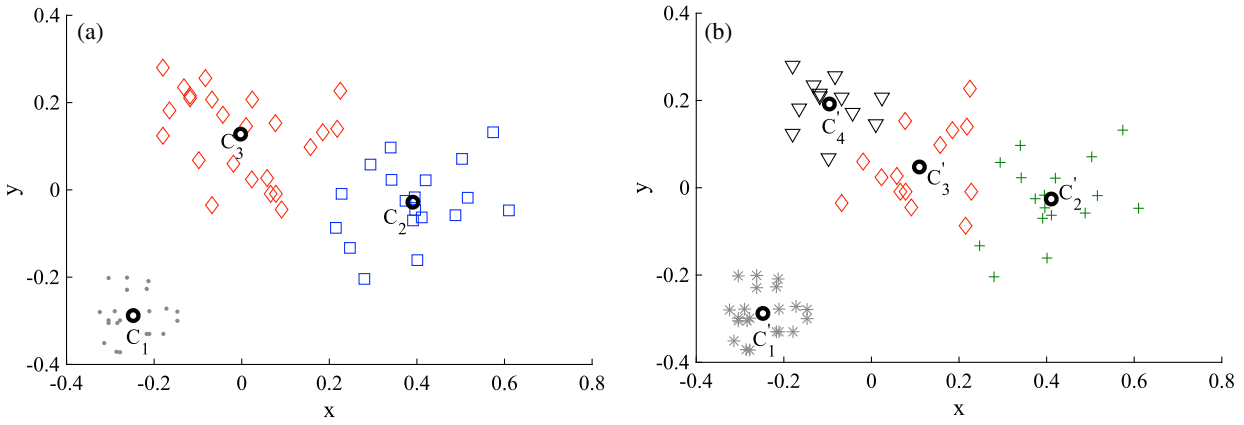

FIG. 3. Clustering of $N = 64$ hypothetical data using the k-means method for (a) $q = 3$ clusters and (b) $q = 4$ clusters.

of data, represented in a two-dimensional Cartesian space.[10] First three clusters [$q = 3$ in Fig. 3)], and then four [$q = 4$ in Fig. 3)] have been chosen to start the calculations. The average silhouette values [$\langle S(3)\rangle > \langle S(4)\rangle$, as shown in Table II] indicate that in the three-cluster solution the clusters are more defined, i.e., they are more compact and distinct from each other than in the four-cluster case.

It is interesting to study how well a centroid geometrically characterizes its cluster. Two parameters affect this: both the cluster density and the number of its elements.[11] For this purpose, we propose a coefficient r_k defined as the centroid *reliability*. It is calculated as follows:

$$r_k = \frac{\langle S(q)\rangle_k}{1 - \langle S(q)\rangle_k}\frac{1}{n_k}, \tag{5}$$

where n_k is the number of students contained in cluster Cl_k and $\langle S(q)\rangle_k$ is the average value of the S function on the same cluster that, as we pointed out, gives information on the cluster density.[12] High values of r_k indicate that the centroid characterizes the cluster well. This also means that the characteristics of all cluster elements are not differentiated very strongly from each other and with respect to those of the centroid.

[10]Other examples, based on real data, can be found in the literature. See, for example, the recent works of Di Paola *et al.* [43] and Battaglia *et al.* [44,45].

[11]For example, two clusters with similar density but different student numbers (i.e., with different variability of student properties) are differently characterized by their centroids: the more populated cluster being less well characterized by its centroid than the other one.

[12]The term $1 - \langle S(q)\rangle_k$ in Eq. (4) is needed to differently weight $\langle S(q)\rangle_k$ and n_k because when $\langle S(q)\rangle_k \to 1$, the r_k value should be independent of the value of n_k.

V. AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

In the hierarchical clustering algorithm (H-CLA), each student is initially considered as a separate cluster. Then the two "closest" students are linked as a cluster and this process is continued (in a stepwise manner) to join one student with another, a student with a cluster, or a cluster with another cluster, until all the students are combined into one single cluster as one moves up the hierarchy (*agglomerative hierarchical clustering*) [13].

The results of hierarchical clustering are graphically displayed as a tree, referred to as the *hierarchical tree* or *dendrogram*. The term closest is identified by a specific rule in each of the *linkage algorithms* [13] used in H-CLA. Hence, in different linkage algorithms the corresponding distance matrix after each merger is differently computed.

A. Linkage algorithms

The choice of a linkage algorithms is one of the most relevant aspects of H-CLA, because different algorithms can generate different dendrograms and, so, different results.

Among the many linkage algorithms described in the literature, the following have been taken into account in education studies: *single, complete, average, and weighted average*. Each algorithm differs in how it measures the distance between two clusters r and s by means of the definition of a new metric (an "ultrametric"), and, consequently, influences the interpretation of the word closest. *Single linkage*, also called *nearest neighbor linkage*, links r and s by using the smallest distance between the students in r and those in s. *Complete linkage*, also called *farthest neighbor linkage*, uses the largest distance between the students in r and the ones in s. *Average linkage* uses the average distance between the students in the two clusters. *Weighted average linkage* uses a recursive definition for the distance between two clusters. If cluster r was created by

TABLE II. Silhouette values for clusters depicted in Fig. 3. The confidence intervals are reported according to a significance level (CI) of 95%.[11]

Number of clusters(q)	Silhouette average value $\langle S(q)\rangle$ (CI)	Silhouette average value for cluster $\langle S(q)\rangle_k$, $k = 1, \ldots, q$ (CI)			
3	0.795 (0.780–0.805)	k			
		1	2		3
		0.953 (0.951–0.956))	0.79 (0.78–0.81)		0.66 (0.63–0.68)
4	0.729 (0.711–0.744)	k			
		1	2	3	4
		0.953 (0.951–0.956)	0.67 (0.64–0.69)	0.77 (0.74–0.79)	0.44 (0.40–0.47)

combining clusters p and q, the distance between r and another cluster s is defined as the average of the distance between p and s and the distance between q and s.

Several conditions can determine the choice of a specific linkage algorithm. For instance, when the source data are in binary form (as in our case), the single and complete linkage algorithms do not give a smooth progression of the distances [20]. For this reason, when the source data are in binary form, the viable linkage algorithms actually reduce to the average or weighted average ones.

In the specialized literature, it is easy to find numeric indexes driving the choice of a specific linkage algorithm, such as the "*cophenetic correlation coefficient*" [48,49].

The cophenetic correlation coefficient, c_{coph}, gives a measure of the concordance between two matrices: D, the matrix of the distances and Δ, the matrix of the ultrametric distances. It is defined as

$$c_{\text{coph}} = \frac{\sum_{i<j}(d_{ij} - \langle D\rangle)(\delta_{ij} - \langle\Delta\rangle)}{\sqrt{\sum_{i<y}(d_{ij} - \langle D\rangle)^2 \sum_{i<j}(\delta_{ij} - \langle\Delta\rangle)^2}}, \quad (6)$$

where d_{ij} is the distance between elements i and j in D, δ_{ij} is the ultrametric distance between elements i and j in Δ; i.e., the height of the link at which the two elements i and j are first joined together, and $\langle D\rangle$ and $\langle\Delta\rangle$ are the average of D and Δ, respectively.

The higher the c_{coph} values, the more the matrix Δ is actually representative of matrix D and, consequently, the more the ultrametric distances δ_{ij} are representative of distances d_{ij}.

The c_{coph} value is based on the correlation (similar to the Pearson one [50]) between the original distances[13] in D and the ultrametric distances given by the linkage type (contained in a new matrix, Δ) and it evaluates how much the latter are actually representative of the former. More precisely, the cophenetic coefficient is a measure of how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points.

Although useful in helping to choose the optimal linkage (in the sense of a linkage that does not distort the distance matrix excessively), the cophenetic coefficient is not free of critical issues.

In fact, as a Pearson-like correlation coefficient, it tries to quantify the "goodness" of a possible linear relationship between D and Δ under the hypothesis that these two matrices are statistically independent. However, this hypothesis is not generally verified, and in many cases the relationship between D and Δ may not be monotonic. Moreover, even in the case of a linear relationship between the corresponding values of the two matrices (and therefore a high value of the cophenetic coefficient), the difference between these may not be small.

These critical issues can be overcome by the methodology proposed by Merigot *et al.* [51]. They discuss a method based on measuring the distance between the two matrices D and Δ. In this way the goodness of the linkage to be used is quantified in terms of a distance. However, the metric proposed by the authors is, in many cases, not effective because it returns the same distance values for different types of linkages, thus failing to discriminate between them. So, we here propose the following definition of distance between two corresponding elements of D and Δ:

$$\sqrt{\sum_i \sum_j (d_{ij} - \delta_{ij})^2}, \quad (7)$$

which is inspired by the well-known Frobenius norm [52] and is a matrix 2-norm. The smaller the value of this expression, the smaller the distance between the two matrices will be, minimizing the distortion introduced by the type of linkage. As we will show later, the values obtained are better differentiated with respect to the values obtained through the cophenetic coefficient, and therefore the results obtained by using Eq. (7) seem to be useful to select the optimal linkage.

Finally, we note here that the use of a matrix norm does not need any hypothesis on the relationship between the

[13]It is worth noting here that hierarchical clustering methods are much less influenced by the type of metric chosen than non-hierarchical clustering methods.

distance and the ultrametric distance. So, this method can be more efficient and flexible than the one based on cophenetic correlation, because it can be indifferently used for each linkage algorithm.

B. Stopping criteria

Reading a dendrogram and finding clusters in it can be a rather arbitrary process. There is not a widely accepted criterion that can be applied to determine the distance values to be chosen for identifying the clusters. Different criteria, named *stopping criteria*, aimed at finding the optimal[14] clustering solution are discussed in the literature (see, for example, Springuel [50]). Many of these cannot be applied to non-numeric data, as is our case, and no one criterion is recognized as the ultimate way to decide what the best clustering solution to a specific situation should be. Here we discuss some of these criteria, and we will jointly use them to find a solution that is based on an agreement among them. The first one involves the calculation of the *inconsistency coefficient* (I_k) [53]. The second is known as the *variation ratio criterion* (VRC) [54]. The third one is proposed here by us and will be called the *cluster differentiation coefficient* (CDC).

One way to find the largest number of clusters that can be considered distinct from each other in a cluster tree is to compare the height of each link with the heights of neighboring links below it in the tree. A link that is approximately the same height as the links below it indicates that there are no distinct divisions between the objects joined at this level of the hierarchy. These links are said to exhibit a high level of consistency because the distance between the objects being joined is approximately the same as the distances between the objects they contain. On the other hand, a link whose height differs noticeably from the height of the links below it indicates that the objects joined at this level in the cluster tree are much farther apart from each other than their components were when they were joined. This link is said to be inconsistent with the links below it.

The relative consistency of each link in a hierarchical cluster tree can be quantified through I_k. This coefficient compares the height of each link in a cluster tree made of N elements with the heights of neighboring links above it in the tree. The calculations of inconsistency coefficients are performed on the matrix of the ultrametric distances Δ generated by the chosen linkage method.

We consider two clusters, s and t, whose distance value is reported in matrix Δ and that converge in a new link k (with $k = 1, 2, \ldots, N-1$). If we indicate by $\delta(k)$ the height in the dendrogram of such a link, its inconsistency coefficient is calculated as follows:

$$I_k = \frac{\delta(k) - \langle \delta(k) \rangle_n}{\sigma_n(\delta(k))}, \tag{8}$$

where $\delta(k)$ is the height of the link k, $\langle \delta(k) \rangle_n$ is the mean of the heights of n links below the link k (usually $n = 3$ links are taken into account), and $\sigma_n(\delta(k))$ is the standard deviation of the heights of these n links.

Equation (8) shows that a link whose height differs noticeably from the height of the n links below it indicates that the objects joined at this level in the cluster tree are much farther apart from each other than their n components. Such a link has a high value of I_k. On the contrary, if the link k is approximately the same height as the links below it, no distinct divisions between the objects joined at this level of the hierarchy can be identified. Such a link has a low value of I_k.

The higher the value of this coefficient, the less consistent is the link connecting the students. A link that joins distinct clusters has a high inconsistency coefficient; a link that joins clusters that should be indistinct has a low inconsistency coefficient.

In the specialized literature [48], the I_k value of a given link is considered by also taking into account the ultrametric distance of the link, in order to avoid a too low or too high fragmentation[15] of the sample clusters.

Once I_k has been used to find the highest number t of clusters that can be identified, we must consider that they could still be merged together in different ways or "configurations" of q clusters $(q \leq t)$ to obtain the optimal clustering solution, which in our case can mean to obtain the maximum information on the sample we are studying.

Cowgill *et al.* [55] obtain the best clustering solution to a given problem by using the VRC (without specifying what this "best" solution actually means). This coefficient relates the best clustering solution to two factors: high cluster separation and/or high cluster compactness. For a given configuration of N elements in q clusters, this value is defined as

$$\mathrm{VRC} = \frac{\mathrm{BGSS}}{q-1} \Big/ \frac{\mathrm{WGSS}}{N-q}. \tag{9}$$

Here *within group squared sum* (WGSS) represents the sum of squares of the distances between the elements belonging to a same cluster. A low WGSS value is related to the cluster elements being tightly packed, i.e., to the cluster compactness. *Between group squared sum* (BGSS), on the other hand, defines the sum of squares of the distances between elements of a given cluster group and the

[14]The meaning of the "optimal" clustering solution will be clarified below.

[15]A "too low" fragmentation is here to be intended as a situation where one or two big clusters are produced, which do not allow us to effectively describe the sample behavior. A "too high" fragmentation means that many small clusters, containing only a few students, are produced.

external ones. A high value of BGSS corresponds to clearly distinct clusters.

According to Calinski and Harabasz [54], the larger the VRC value, the better the clustering solution. On the other hand, it is easy to see that when the number of clusters q increases, the WGSS markedly decreases, making the VRC an increasing function of q independent of the cluster separation. As a consequence, VRC alone may not be sufficient to choose the best clustering solution for the problem analyzed [56].

Generally speaking, we are interested in obtaining information about the sample we are studying that depends on both the cluster separation and their number.[16] In fact, when we obtain many clusters, we can characterize the sample by means of many typical student behaviors, one for each cluster. However, this leads to a characterization that is significant for the researcher in physics education only if the clusters are also clearly distinct from each other. For this reason, in order to quantitatively study information about the sample, we take into account two parameters: the number of clusters and their distinctness,[17] and define the product between them. Taking into account our data, the cluster distinctness tends to increase as the number of clusters increases, reaching a maximum value and then starting to decrease again, at least within a certain range of clusters.[18] Therefore, the product between the number of clusters and the cluster distinctness will have a nonmonotonic behavior with respect to the number of clusters. Its maximum value will give us the maximum information about the sample. It is worth noting here that for our purposes it is not sufficient to consider the number of clusters corresponding to the maximum distinctness, as we are interested in finding as many as possible, still clearly distinct, clusters and not just the maximum of distinctness. For example, a solution with three clearly distinct clusters is more significant than a two-cluster solution with greater distinctness as the former allows us to find a higher number of typical behaviors in the sample.

Following the above discussion, we define the CDC as follows:

$$\mathrm{CDC} = \frac{4q}{N^2 l \binom{q}{2}} \sum_{i=1\ldots q} \sum_{j=1\ldots q} n_i n_j \Theta_{ij} \qquad (10)$$

where n_i and n_j are the number of elements in clusters i and j, respectively, Θ is the "distinctness" of clusters i and j, defined as the number of components of cluster i and j centroids[19] that are different from each other, l is the total number of centroid components, and $\binom{q}{2}$ is the number of combinations of q elements taken two at a time. The factor

$$\frac{4}{N^2 l \binom{q}{2}}$$

is needed to normalize the CDC value with respect to the total number of clusters.

According to this definition, the higher the CDC value for a given cluster configuration, the greater the amount of information we obtain from it.

Finally, we note that the maximum of cluster distinctness we obtain is a relative maximum. Nothing permits us to think that in another range of values, for example, for a much higher number of clusters, distinction may start to increase again. However, too many small clusters could make the sample analysis less interesting for the researcher, as many "microbehaviors" related to the various clusters are found and must be explained.

In conclusion, we suggest that the optimal clustering solution for our problem may be the one for which both VRC and CDC have their maximum values.

It is worth noting that both VRC and CDC criteria can also be used when we want to obtain the best clustering solution by using the k-means algorithm, as they are independent of the specific clustering method used.

VI. A BRIEF COMPARISON OF THE TWO CLUSTERING METHODS

A strong point of the k-means method is certainly the greater ease of computer implementation and of reading the results with respect to the H-CLA method. In fact, the two-dimensional graph immediately allows the reader to have an overall view of the partition of the sample under examination. Reading dendrograms is much more difficult and less intuitive. However, dendrograms can provide more detailed information, but above all a greater robustness of the variable used to measure the similarity between the cluster elements. In particular, in the case of k means, we are obliged to use a distance that is Euclidean, while in the case of dendrograms there is no need to use Euclidean metrics, and similarity and/or dissimilarity can be used.

The only constraint in hierarchical clustering consists in having a function that is monotonic. From this point of view, the dendrogram appears to be more flexible and less tied to the particular methodology used to estimate the similarity between elements. Furthermore, in the case of dendrograms it is not necessary to know *a priori* the

[16]Many well distinct clusters give better information about the sample than a few clusters, not well distinct from each other.

[17]The distinctness of two clusters i and j can be measured as the number of components of cluster i and j centroids that are different from each other.

[18]We studied the cluster distinctness as a function of the cluster number ranging from two to ten clusters.

[19]For centroid, we here understand an array whose components are the answers most frequently given to the questions by the cluster students as in the NH-CLA case.

number of clusters into which the dataset is to be partitioned. On the other hand, the determination of the optimal cluster number, as we have shown, is more complex than in the k-means method.

VII. CHARACTERIZATION OF CLUSTERS

Once the appropriate partitioning of data has been found (for both nonhierarchical and hierarchical clustering), the educational researcher is interested in characterizing each cluster, in order to make sense of what the partition means in pedagogical terms. A possible way to do this is to study the answering strategies most frequently used by the students in each cluster. According to Springuel *et al.* [23], only strategies that are used by at least 75% of the cluster students can be considered cluster "prominent characteristics" and are to be used to characterize the cluster. Another possibility is to consider, for each cluster, the centroid array and say that it "well characterizes" its cluster if a high percentage (let us say at least 60%) of the components of each cluster student array are equal to the components of the centroid array. When this percentage is lower, we cannot say that the cluster is well characterized by its centroid, but we can still talk about "emerging behaviors" in the cluster. In other words, when the percentage of each cluster student's answers in common with those of the cluster centroid is lower than 60%, it may still be possible to identify in the cluster specific behaviors, but the centroid cannot be considered as truly representative of the whole cluster.

In the following sections we will present an application of the two CLA procedures we described above to the analysis of real data from the administration of an open-ended questionnaire to a group of high school students. We aim to show how CLA procedures can be useful to characterize student groups according to the typical ways they tackle the problems and situations proposed in the questionnaire, without any prior researcher knowledge of what form those groups would take. We will also briefly discuss the meaning of the characterization in terms of the answers given by the students to the questionnaire, However, we will not elaborate the pedagogical implications of the characterization results, as our main aim is to show the raw results that the two CLA procedures provide to the researcher, comparing them and searching for mutual coherence.

VIII. AN EXAMPLE APPLICATION

Our sample consists of 117 Italian students (aged 18–19) attending the last year of their 5-year secondary school course. They have completed a questionnaire made up of six open-ended questions on the concept of models and modeling. The questionnaire is a revised and shortened version of the one we used in previous research [10] with university students, modified in order to adapt it to a younger sample. It is summarized as follows:

1. Models are widely used in the sciences, but what is, in your opinion, a model in physics?
2. What is a mathematical model?
3. Are models human creations or do they already exist in nature?
4. What are the main characteristics of a model?
5. Can any natural phenomena be described or explained by a model? Explain your answer.
6. Can a natural phenomenon always be expressed by mathematical formulas? Explain your answer.

A list of 43 typical students' answering strategies has been prepared according to the coding procedure described in Sec. II.[20] So, we analyze a binary matrix (modeled on the basis of the one in Table I) composed of 43 rows and 117 columns. All the clustering calculations are made using custom software, written in the C language, for the k-means algorithm as well as for the agglomerative hierarchical clustering. The graphical representations of clusters in both cases are produced using the well-known software MATLAB [57].

A. Nonhierarchical clustering analysis (*NH*-CLA)

In order to define the number q of clusters that best partitions our sample, the mean value of the S function $\langle S(q)\rangle$ is calculated for different numbers of clusters, from 2 to 8 (see Fig. 4).[21] The figure shows that the best partition of our sample should be achieved by choosing the three-cluster solution, as the $\langle S(q)\rangle$ value has its maximum for $q = 3$. Particularly, $\langle S(3)\rangle$ is 0.69, with CI $= (0.65, 0.73)$. This indicates that a reasonable cluster structure is found with $q = 3$. However, we note that $\langle S(4)\rangle = 0.63$, with CI $= (0.58, 0.67)$. So, $\langle S(3)\rangle$ and $\langle S(4)\rangle$ 95% confidence intervals intersect, and the four-cluster solution cannot be easily discarded.[22] Therefore, a comparison between the two solutions is to be done.

Figure 5 shows the representation of the 3-cluster partition in a two-dimensional graph. The three clusters show a partition of our sample into groups made up of different numbers of students (see Table III).

The three clusters $Cl_k (k = 1, \ldots, 3)$ are identified by their related centroids, C_k. As we have seen, they are the three points in the graph whose arrays $\bar{a}_k$ contain the

[20] In the following, 1A, 1B, …, 1E represent the five identified answering strategies used by students to tackle question 1, 2A, 2B, …, 2H are the eight answering strategies for question 2, and so on. The complete list is reported in the Appendix.

[21] As discussed in Sec. IV. A, for each value of q the clustering procedure was repeated for several values of the initial conditions. In each case, we selected the cluster solution that leads to the minimum values of the distances between each centroid and the cluster elements.

[22] We should also consider that the confidence intervals of clustering solutions with $q = 7$ and 8 also intersect with the $q = 3$ solution. However, an excessive fragmentation of the clustering solution would not be very useful because we would have many small clusters each with only a few students.

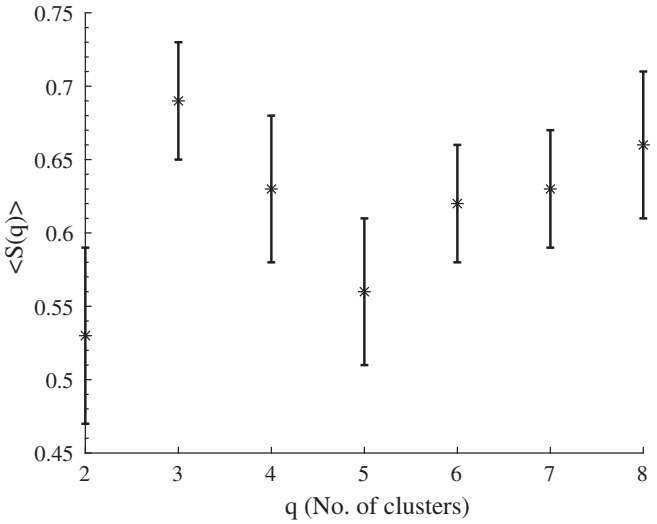

FIG. 4. Average silhouette values and related 95% CI for different cluster partitions of our sample. The highest values are obtained for partitions in $q = 3$ clusters $\langle S(3)\rangle = 0.69$, $\mathrm{CI} = (0.65, 0.73)$.

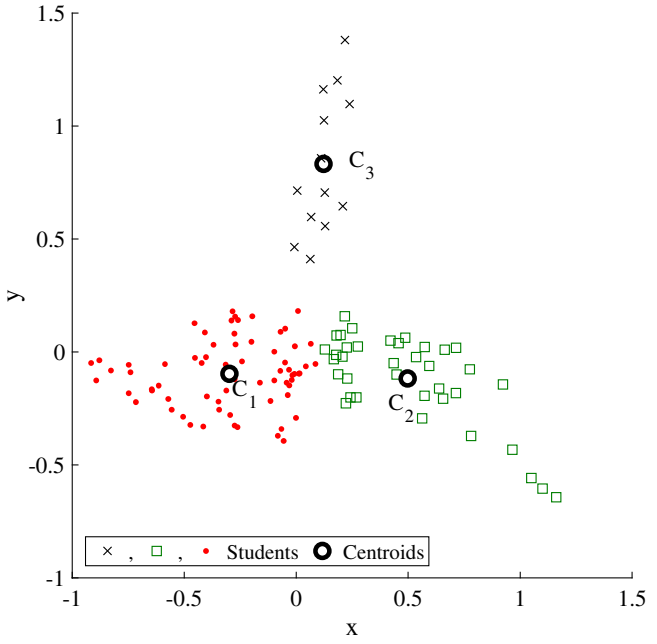

FIG. 5. K-means graph of the 3-cluster partition of our data. Each point in this Cartesian plane represents a student. Points labeled C_1, C_2, and C_3 are the centroids. The x and y axes simply report the values needed to place the points according to their mutual distance.

answering strategies most frequently applied by students in the related clusters (see Table III). The codes used refer to the answering strategies for the six questions, as discussed here. The table also shows the number of students in each cluster, the mean values of the *S function* $\langle S(3)\rangle_k$ $(k = 1, .., 3)$ for the three clusters, and the reliability index r_k of their centroids.

$\langle S(3)\rangle_k$ values indicate that cluster Cl_3 is more compact than the others and more distinct from them, and Cl_2 is the most spread out. Furthermore, the values of r_k show that centroid C_3 best characterizes its cluster, whereas C_1 is the centroid that the least characterizes its cluster.

Figure 6 shows the representation of the four-cluster partition in a two-dimensional graph. Again, the four clusters $Cl_k (k = 1, \ldots, 4)$ can be characterized by their related centroids, C_k. Table IV summarizes all the relevant information as in the previous case.

An inspection of Tables III and IV reveals that the four-cluster solution, while not being completely distinct from the three-cluster one in terms of silhouette value, gives average S values on clusters, $\langle S(4)\rangle_k$ $(\mathrm{k} = 1, \ldots, 4)$ generally lower than in the three-cluster one. Moreover, cluster C_3' has an average S value lower than C_3, despite being exactly the same as it, as can be easily verified from the two clustering solution values. Finally, the CDC values for the two clustering solutions $[\mathrm{CDC}(3) = 1,\ \mathrm{CDC}(4) = 0.8]$ help us to conclude that the three-cluster solution is to be considered the better one from a methodological point view.

1. Discussion of the three-cluster solution results

The interpretation of *ClA* results mainly involves the identification of the typical features characterizing students' answers belonging to the same cluster as well as differences and similarities in answering strategies of students belonging to different clusters. The results reported in the figures and tables shown are clearly global, as they are related to the answering strategies most frequently deployed by the students when tackling the questionnaire. For this reason, we have compared each student array with the centroid array of the cluster in which the student is placed, finding that for clusters Cl_2 and Cl_3 there is at least a 67% accordance. On the other hand, the student arrays of cluster Cl_1 highlight a 50% accordance with the components of the cluster centroid array.

These results, which are coherent with the r_k values reported in Table III, allow us to conclude that, as discussed

TABLE III. An overview of results obtained by the *NH*-CLA method: 3-cluster solution.

Cluster centroid	C_1	C_2	C_3
$\bar{a}_k$ (Most frequently given answers)	1C, 2B, 3B, 4F, 5E, 6G	1A, 2C, 3B-C-D, 4A, 5A, 6B	1E, 2H, 3F, 4H, 5G, 6H
Number of students	67	37	13
$\langle S(3)\rangle_k$	0.72, $\mathrm{CI} = (0.67, 0.75)$	0.62, $\mathrm{CI} = (0.54, 0.69)$	0.78, $\mathrm{CI} = (0.65, 0.84)$
r_k	0.038	0.044	0.27

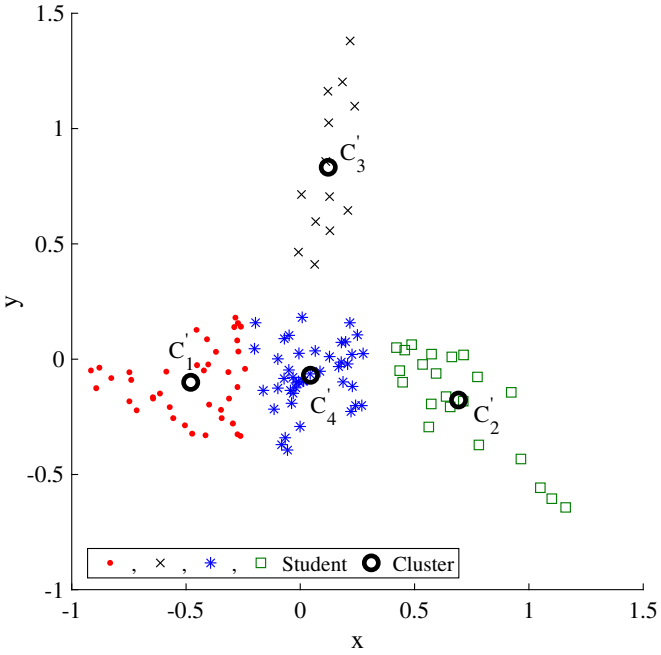

FIG. 6. *K*-means graph of the 4-cluster partition of our data. Each point in this Cartesian plane represents a student. Points labeled C_1, C_2, C_3, and C_4 are the centroids. The x and y axes simply report the values needed to place the points according to their mutual distance.

in Sec. VI, the student profiles sketched by the answers in Cl_2 and Cl_3 centroid arrays sufficiently characterize the behavior of the students in those specific centroids. The centroid array of cluster Cl_1, on the other hand, can only give us information about the "emerging behavior" of students in that cluster.

We will now describe the answers most frequently given by the students in each cluster. We will then comment on the student profiles that can be inferred from these answers, without going into in-depth pedagogical considerations about their meaning, as this is not the main aim of this paper.

Cluster 1

Students in cluster 1 seem to understand the idea of model in mathematical terms (1C and 2B), and to see a model as something real and inspired by bigger models that really exist in nature (3B). They think that it is important that a model clearly highlights the physical variables found relevant for a description of the real phenomenon and again evidence the importance of mathematics in the need to find the relationships between these variables (4F). For Cl_1 students, not all phenomena have been explained, yet, but they are sure that they will all be, in the future (5E). Finally, they say that mathematics is the language the human brain uses to quantitatively describe or explain a real situation (6G).

Cluster 2

Students in cluster 2 say that a model is something physically constructed to be a copy of a real object (1A) and that a mathematical model is a quantitative, basic reproduction of a phenomenon that retains its main aspects (2C). They think that models really exist in nature, and are used to reflect real situations, or to understand objects (3B-3C). Some of them, instead, think that models are abstract constructions and are described by formulas (3D). They are convinced that a model must represent and describe all the features of the object it represents (4A) and that there can exist phenomena that are not described or explained by models, as it is not always possible to describe a phenomenon by means of physical quantities (5A). Finally, they say that a real phenomenon cannot always be expressed by mathematical formulas, as mathematics is an abstract construction (6B).

Cluster 3

Students in cluster 3 say that a model represents a phenomenon and its functioning at different levels (1E) and that a mathematical model represents a phenomenon and can be used for descriptive or predictive aims (2H). They think that models are human creations and help us to understand the world (3F) and that a model should mainly be expressed by using mathematics and/or accepted by the scientific community (4H). All natural phenomena can be described or explained by a model, depending on the scientists' skills (5G), and a mathematical formula can always be used to describe a phenomenon as it is a way to express relationships between variables (6H)

As it is easy to see, students in cluster 3 are the ones that best understand the idea of a model, both in physical and in mathematical terms. Students in cluster 1 show an understanding of the model concept and functions more related to its mathematical aspects. Students in cluster 2 appear to be the ones with the lowest understanding of the model concept and functions, as they often identify models with

TABLE IV. An overview of results obtained by the *NH*-CLA method: 4-cluster solution.

Cluster centroid	C'_1	C'_2	C'_3	C'_4
$\bar{a}_k$ (Most frequently given answers)	1C, 2D-E, 3D-E, 4F, 5E-F, 6G	1A, 2C, 3A-B-C, 4A, 5D-C, 6B	1D, 2H, 3F, 4H, 5G, 6H	1C, 2B-C, 3B, 4E, 5A, 6B-C-D
Number of students	39	22	13	43
$\langle S(4)\rangle_k$	0.54, CI = (0.44, 0.62)	0.55, CI = (0.41, 0.65)	0.68, CI = (0.44, 0.79)	0.73, CI = (0.68, 0.77)
r_k	0.030	0.056	0.16	0.064

TABLE V. Cophenetic correlation coefficient values and 2-norm distance values for different linkage methods.

Type of linkage	C_{coph} value	2-norm distance value
Single	0.76	5603
Complete	0.69	3528
Average	0.83	1793
Weighted average	0.81	1889

real objects and do not think that mathematics can describe real phenomena as it is an abstract construction.

B. Hierarchical clustering analysis (H-CLA)

In order to apply the agglomerative hierarchical clustering to our data, we first have to choose the kind of linkage to use in order to minimize the distortion due to ultrametric distance. We calculate the *cophenetic correlation coefficient* and the *2-normdistance* for the linkages, as reported in Table V.

We choose to use the *average* linkage since the highest values for the cophenetic coefficient and the smallest value for the 2-norm are obtained in this case. It is worth noting that the larger variability of the 2-norm criterion values allows us to choose the best linkage with more confidence. Figure 7 shows the resulting dendrogram of the nested cluster structure.

In Fig. 7, the vertical axis represents the ultrametric distance between two clusters when they are joined; the horizontal axis is divided in 117 ticks, each representing a student. Furthermore, vertical lines represent students or groups of students and horizontal lines represent the joining of two clusters. Vertical lines are always placed in the center of the group of students in a cluster, and horizontal lines are placed at the height that corresponds to the distance between the two clusters that they join.

By describing the cluster tree from the top down, as if clusters are splitting apart, we can see that all the students come together into a single cluster, located at the top of

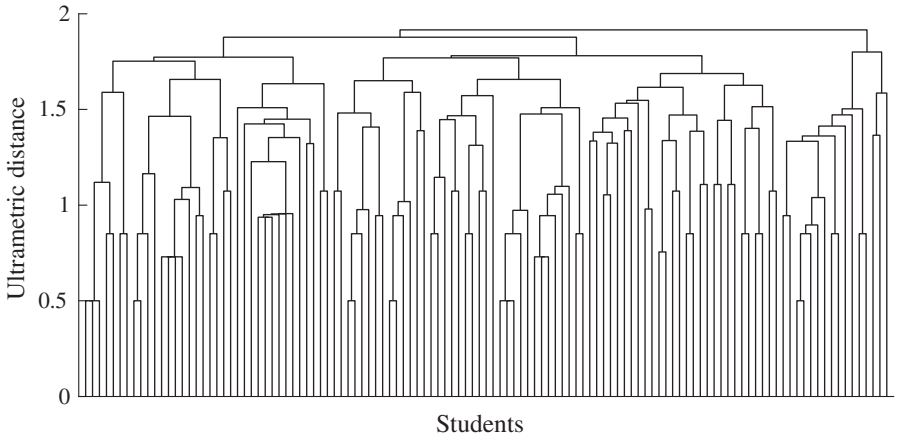

FIG. 7. Dendrogram of our data. Horizontal and vertical axes represent students and ultrametric distances, respectively.

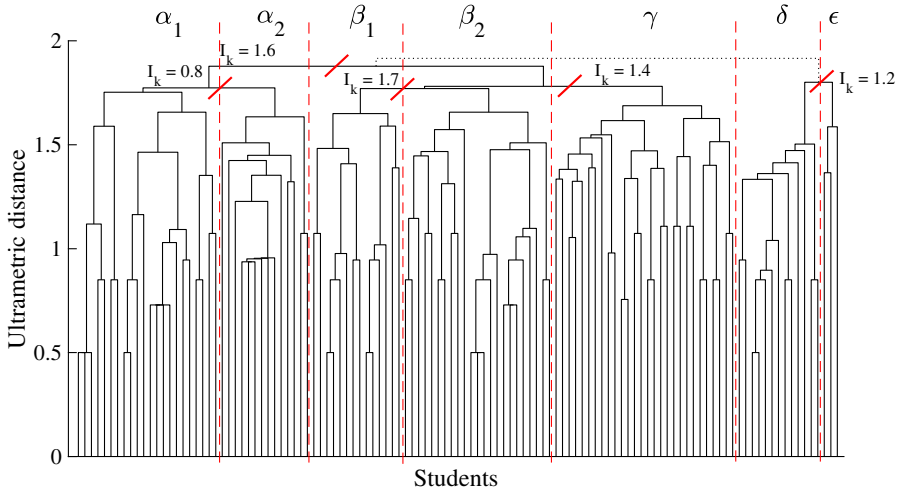

FIG. 8. Dendrogram of our data. Horizontal and vertical axes represent students and ultrametric distances, respectively. The inconsistency coefficients are reported and seven clusters (α_1, α_2, β_1, β_2, γ, δ, ε) are taken into account.

TABLE VI. Possible cluster configurations obtained by merging together the seven clusters found by means of I_k evaluation. $\alpha = \alpha_1 \cup \alpha_2; \beta = \beta_1 \cup \beta_2$.

	Configurations			
No. of clusters	A	B	C	D
2	$\alpha \cup \beta \cup \gamma;\ \delta \cup \varepsilon$	⋯		
3	$\alpha;\ \beta \cup \gamma;\ \delta \cup \varepsilon$	$\alpha \cup \beta \cup \gamma;\ \delta;\ \varepsilon$		
4	$\alpha;\ \beta;\ \gamma;\ \delta \cup \varepsilon$	$\alpha;\ \beta \cup \gamma;\ \delta;\ \varepsilon$	$\alpha_1;\ \alpha_2;\ \beta \cup \gamma;\ \delta \cup \varepsilon$	
5	$\alpha;\ \beta_1;\ \beta_2;\ \gamma;\ \delta \cup \varepsilon$	$\alpha;\ \beta;\ \gamma;\ \delta;\ \varepsilon$	$\alpha_1;\ \alpha_2;\ \beta;\ \gamma;\ \delta \cup \varepsilon$	$\alpha_1;\ \alpha_2;\ \beta \cup \gamma;\ \delta;\ \varepsilon$
6	$\alpha;\ \beta_1;\ \beta_2;\ \gamma;\ \delta;\ \varepsilon$	$\alpha_1;\ \alpha_2;\ \beta;\ \gamma;\ \delta;\ \varepsilon$	$\alpha_1;\ \alpha_2;\ \beta_1;\ \beta_2;\ \gamma;\ \delta \cup \varepsilon$	
7	$\alpha_1;\ \alpha_2;\ \beta_1;\ \beta_2\gamma;\ \delta;\ \varepsilon$	⋯		

the figure. The problem is to identify the "best" clustering solution.

We start by using the inconsistency coefficient, I_k (see Sec. V B) in order to see which links could be neglected because they are inconsistent. It is interesting to note that the links we neglect due to their high I_k values should be limited to a range of ultrametric distances that does not produce too high a fragmentation of the clustering solution [52].

Figure 8 shows that by neglecting all the cut off (in red) links (that have high I_k), we can identify in the dendrogram not more than seven clusters:α_1, α_2, β_1, β_2, γ, δ, ε. By neglecting the links at smaller distances, the clustering solution would be too fragmented, complicating the search for significant information to the researcher.

However, in order to best extract information from the dendrogram, we must identify, by means of the VRC and CDC criteria, the most significant cluster configuration. It could be the one with seven clusters, or another composed of a smaller number of clusters.

Table VI reports the possible cluster configurations that can be obtained by merging together the seven clusters found by means of I_k evaluation.

Figure 9 shows the VRC values for the different cluster configurations. The maximum value is obtained for $q = 6$, with configuration C. However, configuration 4A has a VRC value not so different from 6-C, so it could also be considered in order to decide which is the best cluster configuration for the dendrogram of our data.

Figure 10 shows the CDC values for the different cluster configurations. The maximum value is obtained for configuration 3A, and configuration 4A is next highest. Note that the CDC value for configuration 6C is markedly lower than these two. As in Sec. V B we suggested choosing an H-CLA clustering solution supported by both VRC and CDC criteria, we can conclude that the best one is 4-A.

Figure 11 shows the 4-A clustering solution. It can be studied and characterized by analyzing the most frequent answers to each of the six questions in the questionnaire and giving them a meaning. Table VII provides significant information concerning this cluster configuration.

1. Discussion of the four-cluster solution results

As we have already done in the section regarding NH-CLA, we will now describe the answers most frequently given by the students in the four clusters α, β, γ, and

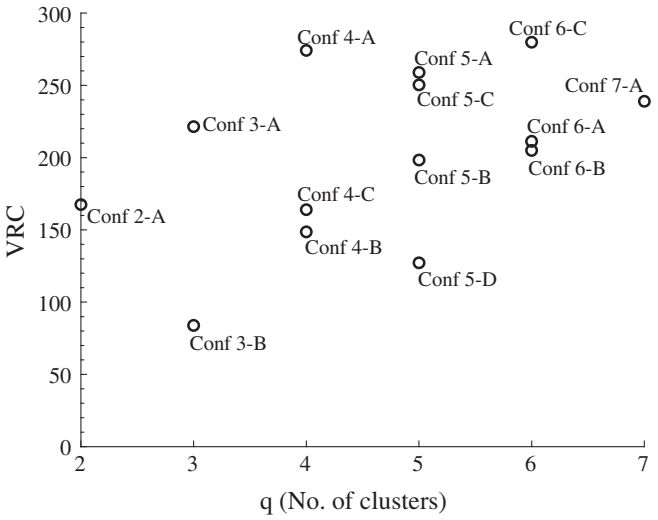

FIG. 9. VRC values for the different cluster configurations obtained by means of H-CLA methods.

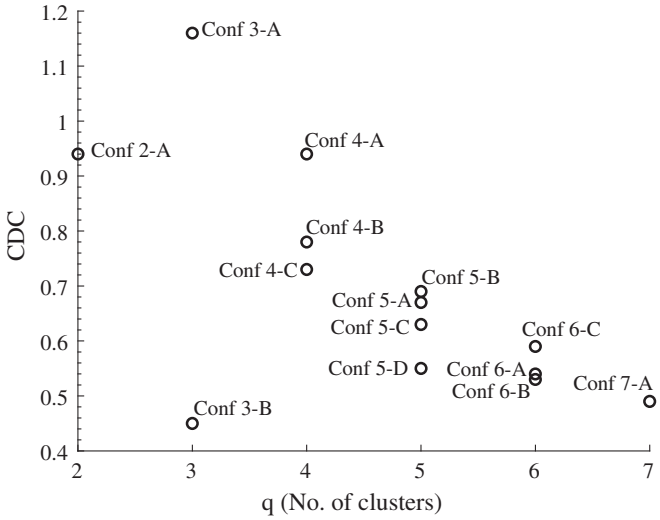

FIG. 10. CDC values for the different cluster configurations obtained by means of H-CLA methods.

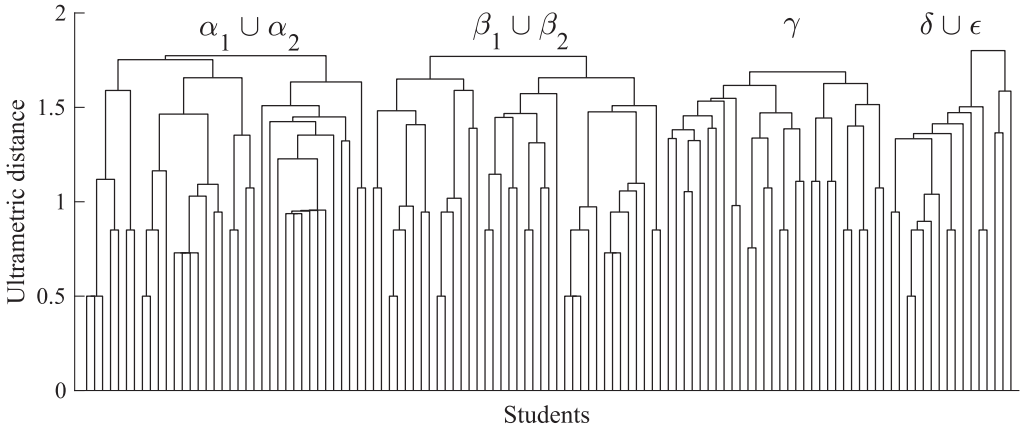

FIG. 11. Dendrogram plot of our sample in which four clusters ($\alpha_1 \cup \alpha_2$, $\beta_1 \cup \beta_2$, γ, $\delta \cup \varepsilon$) are clearly highlighted.

$\delta \cup \varepsilon$. Then, we will briefly comment on the student profiles that can be inferred from these answers

Cluster α

Students in cluster α see a model as something physically constructed to be a copy of a real object (1A) and that a mathematical model is a quantitative reproduction of a phenomenon that retains its main aspects (2C). They think that models are described by formulas (3D) and that a model must represent and describe all the features of the object it represents (4A). Moreover, they think that there can exist phenomena that are not described or explained by models, and some that cannot be described by means of physical quantities (5A). Finally, they say that a natural phenomenon cannot always be expressed by mathematical formulas, as mathematics is an abstract construction (6B).

Cluster β

Students in cluster β see a model as simply a formula to be used to describe a physical phenomenon (1C) and a mathematical model as a symbolic representation of a phenomenon (2D). For them a model is a real object that is inspired by preexisting real natural models (3B), and it must be able to account for the features of reality that are of practical interest (4E). Moreover, they seem to confuse nature with the descriptions or explanations physics gives of it (5B) and think that a natural phenomenon can always be described in mathematical language because a mathematical demonstration is always possible (6E). In summary answers to the questions by students in β highlight an approach to the idea of model that is strictly mathematics inspired and linked to the idea of model as a real object.

Cluster γ

Students in cluster γ see a model as simply a formula, to be used to describe a physical phenomenon (1C) and a mathematical model as a formula aimed at solving problems (2E) or as a way to construct argumentations or demonstrate hypotheses (2G). According to them, models are abstract constructions, described by mathematical formulas (3D) or used to build theories of the world (3E), and they must highlight the physical quantities that are useful for the description of the phenomenon and its mathematical relationships (4F). They say that not all phenomena have been explained, yet, but they will be in the future (5E), and think that a natural phenomenon can always be expressed by mathematical formulas because mathematics is the language we use to this aim (6G). Students in γ see models as abstract constructions that, by using mathematics, can describe natural phenomena, now or in the future.

Cluster $\delta \cup \varepsilon$

Students in cluster $\delta \cup \varepsilon$ say that a model represents a phenomenon and its functioning at different levels (1E) and that a mathematical model represents a phenomenon and can be used for descriptive or predictive aims (2H). They think that models are human creations and help us to understand the world (3F) and that a model should mainly be expressed by using mathematics and/or accepted by the scientific community (4H). All natural phenomena can be described or explained by a model, depending on the scientists' skills (5G), and a mathematical formula can always be used to describe a phenomenon as it is a way to express relationships between variables (6H)

TABLE VII. An overview of results obtained by the H-CLA method: 4-cluster solution.

Cluster	$\alpha = \alpha_1 \cup \alpha_2$	$\beta = \beta_1 \cup \beta_2$	γ	$\delta \cup \varepsilon$
Most frequently given answers	1A, 2C, 3D, 4A, 5A, 6B	1C, 2D, 3B, 4E, 5B, 6E	1C, 2E-G, 3D-E, 4F, 5E, 6G	1E, 2H, 3F, 4H, 5G, 6H
Number of students	36	37	28	16

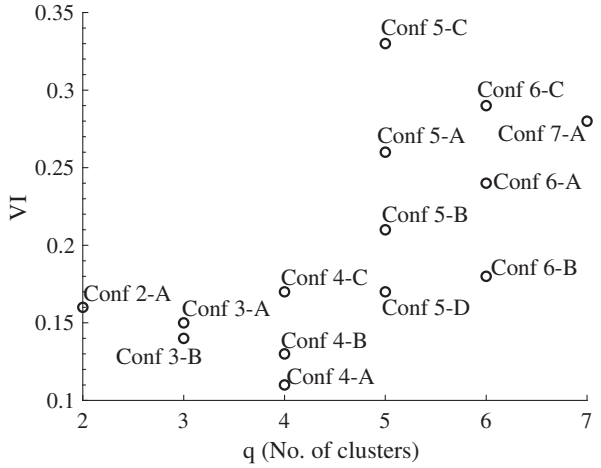

FIG. 12. VI values calculated between the 3-cluster results of the k-means method and each of the clustering results of the H-CLA method, as summarized in Table VI. The x axis reports the number of clusters in the various H-CLA results.

Some more considerations about students in clusters β and γ, which appear to highlight different levels of understanding with respect to the model concept, will be made at the end of the following section.

IX. A COMPARISON BETWEEN *NH-CLA* AND *H-CLA* RESULTS

We now discuss how the results we obtained by using hierarchical and nonhierarchical analysis methods can be compared in order to study their differences and possible coherence aspects, trying to see if one of the two methods allows the researcher to obtain a finer grain detail of the student's ways of answering the proposed questions. As Meila *et al.* [58] point out: "*Just as one cannot define a best clustering method out of context, one cannot define a criterion for comparing clusters that fits every problem.*" Many coefficients have been identified to compare two partitions of the same data set obtained with different methods, but the majority of them are not applicable to our data as they are in binary form. However, a criterion called *variation of information* (VI) can be applied in our case. It measures the difference in information shared between two particular partitions of data and the total information content of the two partitions. In this sense, the smaller the distance between the two clustering solutions, the more these are coherent with each other, and vice versa. VI values can be normalized to the 0–1 range: a value equal to 0 indicates very similar clustering results, and a value equal to 1 corresponds to very different ones. Meila *et al.* [58] supply all the details for VI calculation as well as examples of its application.

We calculated the value of VI to compare the 3-cluster results of the k-means method with the 2-cluster, 3-cluster, 4-cluster, 5-cluster, and 6-cluster results of the H-CLA method (summarized in Table VI) and obtained the values shown in Fig. 12.

We conclude that the best agreement can be found between the 3-cluster solution of the k-means method and the 4A clustering results of the H-CLA method, as the VI value between these solutions is the lowest. We note that this result somehow supports our previous decision to consider solution 4A as the best H-CLA one, considering the higher value of VI obtained for solution 6-C (that, we recall, gave the higher VRC value, as shown in Table VI and Fig. 9).

Therefore, although the two partitions of our student sample obtained by applying the NH-CLA and H-CLA methods are different, they are similar in terms of information conveyed. It is worth noting that the characterization via the dendrogram should also allow us to obtain more detail. In fact, as is detailed below, students in NH-Cl cluster Cl_1, which turns out to be very extensive with a large number of students and a low value of r_k, are basically redistributed into three clusters of the H-CLA solution.

By individually looking at each student placed in the clusters found by means of the two methods (see Table VIII), we can see that in the H-CLA 4-cluster configuration, we have a redistribution of the students originally assigned by NH-CLA to the three clusters shown in Fig. 5 and Table III. Particularly, students in clusters Cl_1 and Cl_2 are mainly assigned by H-CLA to clusters α, β, and γ. Students assigned by NH-CLA to cluster Cl_3, on the other hand, stay in a single cluster, i.e., $\delta \cup \varepsilon$. Going into more detail, we also note that cluster α contains five students assigned by NH-CLA to cluster Cl_1; cluster β also contains four students assigned by NH-CLA to cluster Cl_2; and cluster $\delta \cup \varepsilon$ contains two students placed in Cl_2 and one placed in Cl_1. These students are 12 in all and are labeled in Fig. 13, from where it is clear that they mainly stay in the borderlands of Cl_1 and Cl_2.

Finally, some observations can be made about the improvement in the information that can be gained by considering the results obtained by means of H-CLA. In fact, as described at the end of Sec. VIII, the main difference between the results of NH-CLA and H-CLA is that the big NH-CLA cluster Cl_1 is split by H-CLA into two smaller ones, β and γ. Cluster β mainly contains students that understand the model idea and functions in

TABLE VIII. Redistribution of students placed in NH-CLA clusters into H-CLA clusters.

Cluster	$\alpha = \alpha_1 \cup \alpha_2$	$\beta = \beta_1 \cup \beta_2$	γ	$\delta \cup \varepsilon$
Students in cluster obtained by the k-means method	$(31)\mathrm{Cl}_2 + (5)\mathrm{Cl}_1$	$(33)\mathrm{Cl}_1 + (4)\mathrm{Cl}_2$	$(28)\mathrm{Cl}_1$	$(13)\mathrm{Cl}_3 + (2)\mathrm{Cl}_2 + (1)\mathrm{Cl}_1$

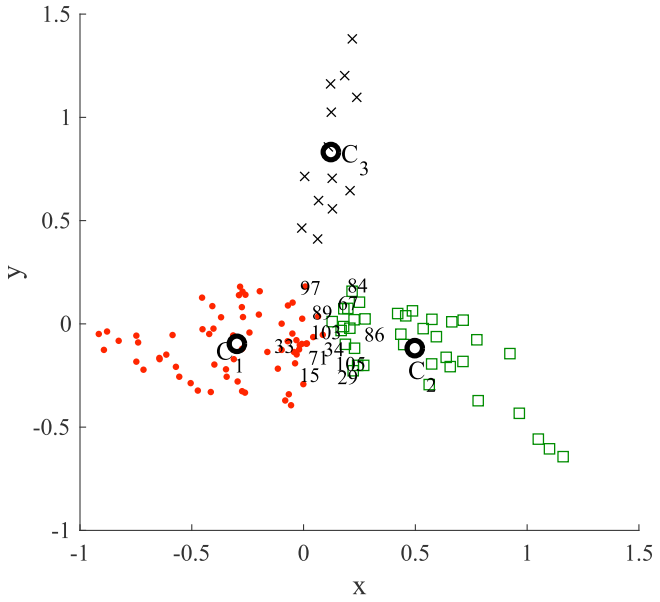

FIG. 13. K-means graph of the 3-cluster partition of our data. The numerically labeled points represent 12 students that in the 4-cluster H-CLA configuration 4A are placed into clusters α, β, and $\delta \cup \varepsilon$. The x and y axes simply report the values needed to place the points according to their mutual distance.

mathematical terms (but also think that a model is a concrete, real object) and think that physics is nature itself. Students in cluster γ, on the other hand, seem to have a more advanced understanding of the model idea and functions, as they see a model as an abstract construction that can explain natural phenomena. However, students in γ still see a model in mathematical terms, thinking that it is a formula used to highlight the relationships between variables measured in experiments.

X. CONCLUSIONS

The use of cluster analysis techniques is common in many fields of research such as, for example, information technology, biology, medicine, archeology, econophysics, and market research. In the field of education, only a limited number of examples of the application of *CLA* are available, and many aspects of the use of the various available techniques have been little studied to reveal their strength and weakness points.

In this paper we described a nonhierarchical clustering method, k-means, which allows the educational researcher to separate students into groups that can be recognized and characterized by common traits in their answers to a questionnaire. It is also possible to represent these groups in a two-dimensional Cartesian graph containing points that represent the students of the sample on the basis of their mutual distances, related to the mutual correlation among students answering the questionnaire. Each of the clusters found by the analysis can be characterized by a point, the "centroid," represented by the answers most frequently given by the students comprising the cluster.

Following this, we described a different method of analysis, based on hierarchical clustering. This method allows the researcher to visualize the clustering results in a graphical tree, called a "dendrogram," which shows the links between pairs and/or groups of students on the basis of their mutual ultrametric distances. Each cluster can be characterized on the basis of the answers most frequently given by the students in it. Again, functions and parameters useful to evaluating the reliability of the results have been discussed.

An application of these two methods to the analysis of the answers to a real questionnaire has been given, in order to clearly show the choices that the researcher must make and what parameters they must use in order to obtain the best partitions of the whole student group and to check the reliability of the result. In order to study the coherence of the results obtained by using hierarchical and nonhierarchical clustering methods, we compare the results to each other. We found that two of the three clusters found by NH-CLA are also present in H-CLA, yet the other is further split, and can thus be better characterized, by means of H-CLA. In fact, the NH-CLA cluster Cl_1, which is the one worst characterized by its centroid (see the r_k value in Table III), is split by H-CLA into two smaller clusters. The answering strategies most frequently used by the students in these two new clusters allow us to isolate student behaviors that were not completely highlighted in the analysis of the student profile emerging from the NH-CLA cluster Cl_1.

We can conclude that the NH-CLA method we discussed here allows the researcher to easily obtain and visualize in a 2D graph a global view of student behavior with respect to the answers to a questionnaire and to obtain a first characterization of student behavior in terms of their most frequently used answering strategies. The H-CLA method, on the other hand, although producing a graph that is not as easy to read as the one produced with the other method (dendrogram vs Voronoi diagram), allows the researcher to obtain results coherent with the NH-CLA ones and that may offer a finer grain detail of student behavior.

APPENDIX: QUESTIONNAIRE AND RELATED ANSWERING STRATEGIES USED BY STUDENTS FOR EACH QUESTION

1. Models are widely used in science and mathematics, but what, in your opinion, is a model in physics?

1A. It is a constructed copy of a real object that we use to study it.

1B. It is a scale or life-size copy of a real object, aimed at helping us to interact with it and/or describe it.

1C. It is a formula that we use to describe a physical phenomenon.

1D. It is an algorithm that we build to describe a specific real situation or physical phenomenon.
1E. It is a representation of a phenomenon, which accounts more or less accurately for its functioning.
2. And what is a mathematical model?
2A. It is a reproduction of an object by means of a geometrical shape.
2B. It is a way to mathematically describe objects.
2C. It is a quantitative, but basic reproduction of a phenomenon.
2D. It is a symbolic representation of a situation or phenomenon.
2E. It is a mathematical formula aimed at solving problems.
2F. It is a quantitative representation of a system, whose basic elements (variables, sources and contexts) are connected by relationships (a set of rules).
2G. It is away to express an argumentation or to demonstrate a hypothesis.
2H. It is a mathematical representation of a phenomenon that can be used to describe it or predict its evolution.
3. Are models human creations or do they already exist in nature?
3A. Models really exist in nature and are real life situations that we use to describe other, more complex, ones.
3B. Models are real objects and are inspired by preexisting real natural models that summarize other real situations.
3C. Models really exist in nature and are used to understand other objects, sometimes only imperfectly.
3D. Models are abstract constructions and are described by mathematical formulas.
3E. Many models are creations of the human mind and are used to build theories of the world.
3F. Models are created by our mind and they allow us to make sense of natural phenomena. They come from a continuous interaction with the world.
4. What are the main characteristics of a model?
4A. A model must describe all the features of the object it represents.
4B. A model must be able to account for all the features of the phenomenon it represents.
4C. The main model aims is to well describe real life situations.
4D. A model should be carefully built, so as to describe a phenomenon well and not be easily substituted by alternative ones.
4E. It should be able to account for the features of reality that are of practical interest.
4F. It must highlight the physical quantities that are useful for the description of the phenomenon, and their mathematical relationships.
4G. It can be qualitative, semi quantitative or quantitative.
4H. It should be expressed in mathematical language and/or accepted by the scientific community.
4I. It must be used to make predictions about the future behavior of physical systems
5. Can all natural phenomena be described or explained by a model? Explain your answer.
5A. No, some phenomena cannot be described or explained by models, as there are some that cannot be described by means of physical quantities.
5B. Yes. A natural phenomenon can always be described by a physical model, as physics is nature itself.
5C. Yes. It just depends on the scientist's ability to carefully reproduce the features of the phenomenon.
5D. Not always. Even the ablest scientist will not be able to reproduce particularly complex systems (like weather, or human behavior…).
5E. Not always. Some phenomena still have not been explained, but they will be in the future.
5F. Yes. A model is simply a way to describe a phenomenon.
5G. Yes, if the scientist is able to isolate all the variables that characterize the phenomenon.
6. Can a natural phenomenon always be expressed by mathematical formulas? Explain your answer.
6A. Yes, but only if it quantitatively describes the entire real situation.
6B. No, as mathematics is an abstract science.
6C. No, because reality is so complex that it cannot always be expressed by means of mathematics.
6D. No, because not all phenomena can be quantitatively described.
6E. Yes, because a mathematical demonstration can always be performed, so mathematics can always explain the quantities measured during an experiment.
6F. No, as a real phenomenon can have characteristics that cannot easily be expressed in mathematical language.
6G. Yes, because mathematics is the language we use to quantitatively describe or explain a real situation.
6H. Yes, as mathematical formulas are a way to express relationships between variables.

[1] P. N. Johnson-Laird, Mental models: Towards a cognitive science of language, *Inference, and Consciousness* (Cambridge University Press, Cambridge, England, 1983).

[2] P. N. Johnson-Laird, *How We Reason* (Oxford University Press, Oxford, UK, 2006).

[3] E. F. Redish, The implications of cognitive studies for teaching physics, Am. J. Phys. **62**, 796 (1994).

[4] I. M. Greca and M. A. Moreira, Mental models, conceptual models, and modeling, Int. J. Sci. Teach. **22**, 1 (2000).

[5] J. K. Gilbert and C. Boulter, Learning science through models and modelling, in *International Handbook of Science Education*, edited by B. J. Fraser and K. G. Tobin (Kluwer Academic Publisher, Dordrecht, Netherlands, 1998), p. 53.

[6] L. Bao and E. F. Redish, Model analysis: Representing and assessing the dynamics of student learning, Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).

[7] J. P. Mestre, Probing adults' conceptual understanding and transfer of learning via problem posing, J. Appl. Dev. Psychol. **23**, 9 (2002).

[8] A. Redfors and J. Ryder, University physics students' use of models in explanations of phenomena involving interaction between metals and electromagnetic radiation, Int. J. Sci. Educ. **23**, 1283 (2001).

[9] M. Borrego, E. P. Douglas, and C. T. Amelink, Quantitative, qualitative, and mixed research methods in engineering education, J. Engin. Educ. **98**, 53 (2009).

[10] C. Fazio, B. Di Paola, and I. Guastella, Prospective elementary teachers' perceptions of the processes of modeling: A case study, Phys. Rev. ST Phys. Educ. Res. **8**, 010110 (2012).

[11] C. Fazio, O. R. Battaglia, and B. Di Paola, Investigating the quality of mental models deployed by undergraduate engineering students in creating explanations: The case of thermally activated phenomena, Phys. Rev. ST Phys. Educ. Res. **9**, 020101 (2013).

[12] N. Pizzolato, C. Fazio, R. M. Sperandeo-Mineo, and D. Persano-Adorno, Open-inquiry driven overcoming of epistemological difficulties in engineering undergraduates: A case study in the context of thermal science, Phys. Rev. ST Phys. Educ. Res. **10**, 010107 (2014).

[13] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis* (John Wiley & Sons, Ltd, Chichester, UK, 2011).

[14] J. Ott, *Analysis of Human Genetic Linkage*, 3rd ed. (Johns Hopkins University Press, Baltimore, London, 1999).

[15] *Cluster Analysis in Neuropsychological Research: 13 Recent Applications*, edited by D. N. Allen and G. Goldstein (Springer Science+Business Media, New York, 2013).

[16] R. N. Mantegna, Hierarchical structure in financial markets, Eur. Phys. J. B **11**, 193 (1999).

[17] M. C. Cowgill and R. J. Harvey, A Genetic Algorithm Approach to Cluster Analysis, Comput. Math. Appl. **37**, 99 (1999).

[18] A. Coates and A. Y. Ng, Learning Feature Representations with K-Means, in *Neural Networks: Tricks of the Trade*, 2nd ed. edited by G. Montavon, G. B. Orr, and K. R. Muller (Springer LNCS 7700, Berlin Heidelberg, 2012), pp. 561–580.

[19] P. Dayan, Unsupervised Learning, in *The MIT Encyclopedia of the Cognitive Sciences Wilson*, edited by R. A. Wilson and F. Keil (MIT Press, London, 1999), pp. 1–7.

[20] R. Sathya and A. Abraham, Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification, Int. J. Adv. Res. Artificial Intell. **2**, 34 (2013).

[21] R. C. Tryon, *Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality* (Edwards Brothers, Ann Arbor, 1939).

[22] R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy* (W. H. Freeman & Co., San Francisco, London, 1963).

[23] R. P. Springuel, M. C. Wittmann, and J. R. Thompson, Applying clustering to statistical analysis of student reasoning about two-dimensional kinematics, Phys. Rev. ST Phys. Educ. Res. **3**, 020107 (2007).

[24] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

[25] D. Hammer and L. K. Berland, Confusing Claims for Data: A critique of Common Practices for Presenting Qualitative Research on Learning, J. Learn. Sci. **23**, 37 (2014).

[26] M. T. H. Chi, Quantifying Qualitative Analyses of Verbal Data: A Practical Guide, J. Learn. Sci. **6**, 271 (1997).

[27] M. Q. Patton, *Qualitative Research and Evaluation Methods*, 3rd ed. (Sage Publications, Thousand Oaks, CA, 2001).

[28] N. Denzin, *Sociological Methods: A Sourcebook*, 5th ed. (Routledge, New York, 2006) (Aldine Transaction).

[29] C. Fazio and O. R. Battaglia, Conceptual Understanding of Newtonian Mechanics Through Cluster Analysis of FCI Student Answers, Int. J. Sci. Math. Educ. (2018).

[30] M. Tumminello, S. Micciché, L. J. Dominguez, G. Lamura, M. G. Melchiorre, M. Barbagallo, and R. N. Mantegna, Happy aged people are all alike, while every unhappy aged person is unhappy in its own, PLoS One **6**, e23377 (2011).

[31] I. C. Lerman, R. Gras, and H. Rostam, Elaboration et evaluation d'un indice d'implication pour des données binaires I, Math. Sci. Hum. **74**, 5 (1981).

[32] J. C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika Trust **53**, 325 (1966).

[33] J. MacQueen, Some methods for classification and analysis of multivariate observations, in *Proceedings in the 5th Berkeley Symposium on Mathematical Statistics and Probability 1965/66*, edited by L. M. LeCam and J. Neyman (University of California Press, Berkeley, 1967), Vol. I, p. 281–297.

[34] R. Padraic Springuel, M. C. Wittmann, and J. R. Thompson, Erratum: Applying clustering to statistical analysis of student reasoning about two-dimensional kinematics [Phys Rev. ST Phys. Educ. Res. 3, 020107 (2007)], Phys Rev. ST Phys. Educ. Res. **3**, 020107 (2007).

[35] I. Borg and P. Groenen, *Modern Multidimensional Scaling* (Springer Verlag, New York, 1997).

[36] R. Loohach and K. Garg, Effect of Distance Functions on K-Means Clustering Algorithm, Int. J. Comput. Appl. **49**, 7 (2012).

[37] F. Leisch, A toolbox for K-centroids cluster analysis, Computational Statistics and Data Analysis **51**, 526 (2006).
[38] G. Voronoi, Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites, J. Reine Angew. Math. **133**, 97 (1908).
[39] J. Stewart, M. Miller, C. Audo, and G. Stewart, Using cluster analysis to identify patterns in students' responses to contextually different conceptual problems, Phys. Rev. ST Phys. Educ. Res. **8**, 020112 (2012).
[40] P. J. Rouseeuw. Silhouttes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. **20**, 53 (1987).
[41] P. Saxena, V. Singh, and S. Lehri, Evolving efficient clustering patterns in liver patient data through data mining techniques, Int. J. Comput. Appl. **66**, 23 (2013).
[42] A. Struyf, M. Hubert, and P. J. Rousseeuw, Clustering in an Object-Oriented Environment, J. Stat. Softw. **1**, 1 (1997).
[43] B. Di Paola, O. R. Battaglia, and C. Fazio, Non-Hierarchical Clustering to analyse an open-ended questionnaire on algebraic thinking, S. Afr. J. Educ. **36**, 1 (2016).
[44] O. R. Battaglia, B. Di Paola, and C. Fazio, K-means clustering to study how student reasoning lines can be modified by a learning activity based on Feynman's unifying approach, Eurasia J. Math. Sci. Technol. Educ. **13**, 2005 (2017).
[45] O. R. Battaglia, B. Di Paola, and C. Fazio, A quantitative analysis of educational data through the comparison between hierarchical and not-hierarchical clustering, Eurasia J. Math. Sci. Technol. Educ. **13**, 4491 (2017).
[46] T. J. DiCiccio and B. Efron, Bootstrap confidence intervals, Stat. Sci. **11**, 189 (1996).
[47] D. V. Inkley, *Bootstrap methods and their applications*, Cambridge Series in Statistical and Probabilistic mathematics (Cambridge University Press, Cambridge, England, 1997).
[48] R. R. Sokal and F. J. Rohlf, The Comparison of Dendrograms by Objective Methods, Int. Assoc. Plant Taxonomy **11**, 33 (1962).
[49] S. Saracli, N. Dogan, and I. Dogan, Comparison of hierarchical cluster analysis methods by cophenetic correlation, J. Inequalities Appl. **203**, 1 (2013).
[50] R. P. Springuel, Applying cluster analysis to physics education research data, Ph.D. thesis, the University of Maine Graduate School, Orono, Maine, United States, 2010, https://www.academia.edu.
[51] B. Merigot, J. Durbec, and J. Gaertner, On goodness of fit measure for dendrogram based analysis, Ecology **91**, 1850 (2010).
[52] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. (Johns Hopkins, Baltimore, MD, 1996).
[53] M. GhasemiGol, H. S. Yazdi, and R. Monsefi, A new Hierarchical Clustering Algorithm on Fuzzy Data (FHCA), Int. J. Comput. Electrical Engin. **2**, 134 (2010).
[54] T. Calinski and J. Harabasz, A dendrite method for cluster analysis, Commun. Stat., Theory Methods **3**, 1 (1974).
[55] C. Cowgill, R. J. Harvey, and L. T. Watson, A Genetic Algorithm Approach to Cluster Analysis, Comput. Math. Appl. **37**, 99 (1999).
[56] M. Manisera and M. Vezzoli, Finding number of groups using a penalized internal cluster quality index, Syrto Working Paper Series, Working paper No. 9 (2013).
[57] MATLAB version 8.6 Natick, Massachusetts: The MathWorks Inc., 2015, www.mathworks.com/products/matlab/.
[58] M. Meila, Comparing clusterings—an information based distance, J. Multivariate Anal. **98**, 873 (2007).