Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory

Paul Tschisgale,[1] Peter Wulff,[2] and Marcus Kubsch[3]

[1]*Department of Physics Education, Leibniz Institute for Science and Mathematics Education,*
Olshausenstraße 62, 24118 Kiel, Germany
[2]*Department of Physics and Physics Education Research, Heidelberg University of Education,*
Im Neuenheimer Feld 561, 69120 Heidelberg, Germany
[3]*Department of Physics—Physics Education Research, Freie Universität Berlin,*
Arnimallee 14, 14195 Berlin, Germany

(Received 30 January 2023; accepted 11 August 2023; published 1 September 2023)

[This paper is part of the Focused Collection on Qualitative Methods in PER: A Critical Examination.] Qualitative research methods have provided key insights in physics education research (PER) by drawing on non-numerical data such as text or video data. While different methods towards qualitative research exist, they share two essential steps: recognizing patterns in the data and interpreting these patterns. Although these methods have led to the development of rigorous theory, there are challenges: As such methods require a series of judgments by the analyst, they are difficult to validate and reproduce. Further, they are hard to scale so that they are unavailable to the analysis of large-scale data. In this way, important phenomena may remain inaccessible to qualitative analysis. Reacting to these challenges and leveraging the potential of emerging methods of artificial intelligence (AI) such as machine learning and natural language processing, sociologist Nelson has proposed the concept of computational grounded theory (CGT). CGT proceeds in a process of three consecutive steps: In the first step, one leverages the power of computational techniques, especially natural language processing and unsupervised machine learning techniques, for pattern detection in large datasets—those of a size and scope that may prohibit human-driven analysis from the outset. In the second step, one relies on the integrative and interpretative capabilities of human researchers to add quality and depth to the quantity and breadth of the first step. In the last step, one again uses computational techniques to test the extent to which the detected and refined patterns from the first and second step hold throughout the whole dataset under investigation. Interestingly, CGT does not aim at simply automating parts of the qualitative process by using AI, but rather aims at integrating AI into the human analyst's workflow within a qualitative analysis. This leads to an analytical system that can do something that is quantitatively and qualitatively different from what a human or machine can do alone. In this way, CGT aims at addressing questions about validity, reproducibility, and scalability in qualitative research while preserving the theoretical sensitivity and unique inferencing capabilities of the human analyst. In this paper, we provide a primer on CGT, present how it can be used to investigate the physics problem-solving approaches of $N = 417$ students based on textual data, and discuss CGT's potentials and challenges in PER. In consequence, this paper can provide critical input to the discussion of how emerging AI technologies can provide new avenues in qualitative PER.

DOI: 10.1103/PhysRevPhysEducRes.19.020123

I. INTRODUCTION

Qualitative research methods have provided key insights in physics education research (PER). The rich descriptions of phenomena such as students' (mis-)conceptions [1], teachers' epistemic cognition [2], expert-novice differences in physics problem solving [3], or the lived experiences of women in physics [4,5], already providing valuable insights in their own right, are fundamental to the development of substantive theory. In this way, qualitative research often enables later quantitative work such as the development of concept inventories and other test instruments [6].

Qualitative research methods can provide these insights as they draw on non-numerical data such as text or video data. While different approaches towards qualitative research exist [7,8], they share two essential steps: recognizing patterns in the data and interpreting these patterns. Both steps require an analyst with expert knowledge and hermeneutic skills, especially in applications in which the patterns and interpretations emerge inductively from the data, i.e., researchers code and interpret data without *a priori* existing codes, such as in grounded theory. In PER, inductive qualitative methods are commonly used, e.g., in the context of investigating students' (mis-)conceptions or reasoning [1,9,10]. While these methods have led to the development of rigorous theory, e.g., knowledge in pieces [10], there are challenges. As these methods require a series of judgments by the analyst, they are difficult to validate and reproduce [11]. Further, they are hard to scale so that they are unavailable to the analysis of large-scale data, e.g., chats in online learning environments. In this way, important phenomena may remain inaccessible to qualitative analysts.

Recently, sociologist Nelson [12] has advanced an analytical framework that aims at answering these challenges by integrating artificial intelligence (AI) techniques such as machine learning (ML) and natural language processing (NLP) into the qualitative data analysis process: computational grounded theory (CGT). CGT aims at using processing power and pattern recognition abilities of computers and integrate them with the theoretical sensitivity of the human analyst. This integrative idea is echoed by other work calling for distributing tasks in the data analysis process between human and computer so that they can complement each other in science education research [13,14].

In this paper, we present an applied example of CGT to probe the potentials and challenges of CGT in the context of PER. Specifically, we use CGT to investigate how students, and particularly participants of the Physics Olympiad, engage in physics problem solving.

II. BACKGROUND—QUALITATIVE RESEARCH METHODS

While qualitative methods are by no means limited to text as a data source, text is one of the most prevalent modalities of data used in qualitative PER. Therefore, and because CGT is most conveniently applied to textual data as fitting computational tools are widely available, and also because we draw on textual data in the applied example which we will present, we place a focus on text analysis in the background section. However, the argument for CGT specifically and the argument for distributing tasks between human and computer in the data analysis process more broadly pertain to other modalities of data such as audio or video data as well [15].

A. Text analysis

Text is considered an important data source in PER. In general, two broad approaches to text analysis are differentiated: qualitative and quantitative approaches.

Qualitative approaches typically employ an interpretative epistemic stance which acknowledges that reality and its meaning are the product of a context-dependent process of social co-construction. In this way, the rich descriptions of phenomena that qualitative methods produce are often fundamental to the development of abstract but data-driven theory [16]. Unfortunately, the application of qualitative methods in this way is typically very time consuming, limiting the amount of text that can be considered in such an analysis. This leads to methodological limitations. First, when only a small sample of text can be considered, there is a relatively large risk that the selected sample of text does not adequately cover the range of phenomena. For example, if one wants to investigate students' conceptions in a domain from open-ended answers, the sampled answers may simply not include the full range of existing conceptions. It is important to note that this limitation does not arise from the data not being available as collecting more open-ended answers from students is typically not the problem. Rather, the time and resources of researchers are limited. Second, there is an epistemic limitation regarding phenomena that only exist or become visible in large datasets [17]. Language data is described by long-tailed distributions. This means that, for a given language, there exists a set of words (and therefore phenomena) that occur frequently in language data, however, there exists a much larger set of words (and phenomena) that have a low probability of occurrence [18,19]. Therefore, rare phenomena eventually only occur in large text corpora which is why it is generally not sufficient to analyze only small samples to identify underlying patterns in textual data. Last, the decisions and subjective judgments of the researcher that are part of the analysis process can make it challenging to validate and reproduce the results of qualitative analyses [11].

Quantitative approaches to text analysis are typically easier to reproduce as the procedure is well documented in the form of computer code. Further, decisions in the analytical process, e.g., for the number of topics in topic models,[1] can often be justified on the basis of information theoretical criteria [21,22]. In addition, the processing power of computers allows for the analysis of large, unstructured textual data [23,24]. However, quantitative methods face the issue that the results of an analysis such as distributions of words over documents or syntactic networks do not speak for themselves [20]. In other words, just as qualitative approaches, quantitative approaches also require a substantial amount of human interpretation. At the same time, the complexity of modern quantitative approaches leads to an increased risk of researchers using methods in inadequate ways or misinterpreting the results which challenges the validity of the findings [25]. Further, when textual data is reduced to statistics in quantitative

[1] Topic models are probabilistic models that assume that texts consist of a mixture of topics. Fitting a topic model to a set of texts results in a description of the topics based on the words that define a topic and the prevalence of each topic in a given text. See [20] for more details.

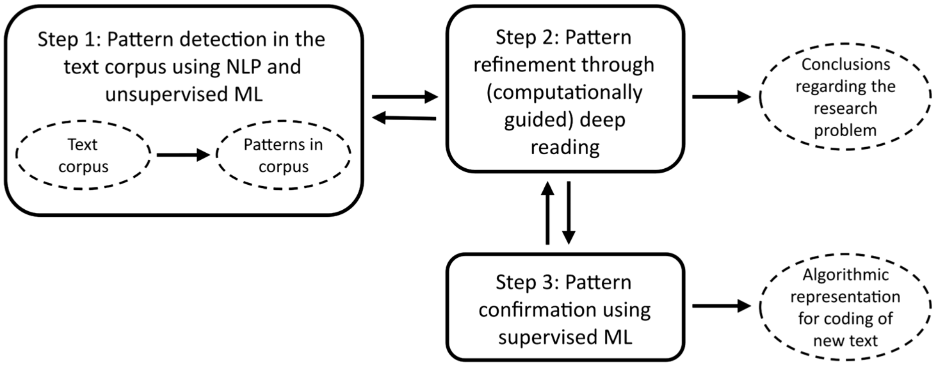
Step 1: Pattern detection in the text corpus using NLP and unsupervised ML
Text corpus
Patterns in corpus
Step 2: Pattern refinement through (computationally guided) deep reading
Conclusions regarding the research problem
Step 3: Pattern confirmation using supervised ML
Algorithmic representation for coding of new text

FIG. 1. Schematic representation (adapted from Kubsch *et al.* [13]) of the three steps of CGT based on Nelson [12].

analyses, important and contextual factors may get lost or nonexisting relationships may be suggested [26,27].

In summary, a picture emerges where the potentials and challenges of qualitative and quantitative methods for text analysis are at least partly complementary. In this light, the integration of quantitative, i.e., computational, and qualitative analytical procedures into an overarching analytical framework appears promising [13]. With CGT, Nelson [12] has introduced such a framework.

B. Computational grounded theory

Computational grounded theory as developed by Nelson [12] combines human analytic power and artificial intelligence-based methods such as natural language processing and machine learning within three consecutive steps: pattern detection, pattern refinement, and pattern confirmation (see Fig. 1).

In the *pattern detection step*, quantitative text analysis methods, especially from NLP and ML, are used in a complementary way for pattern detection in datasets prohibitively large for human analysis (see Fig. 2). NLP is the scientific field that utilizes computers to analyze, understand, and generate human (i.e., natural) language, oftentimes using technologies related to ML. As such, NLP provides models and techniques to systematically analyze natural language. This includes pretrained language models that are typically trained on massive linguistic corpora such as the Common Crawl corpus or Wikipedia. Such pretrained language models allow to transform textual data into numerical data which can then be processed using (more or less standard) quantitative methods. For example, sentences in a text corpus can be mapped to numerical vectors (of generally high dimensionality) which capture parts of the sentences' meanings (see upper part of Fig. 2). These numerical representations of sentences in the form of vectors are referred to as sentence embeddings and can easily be processed in further downstream tasks such as pattern detection through ML techniques [28]. ML is the scientific field mainly devoted to computational algorithms that can either learn the association between input data and corresponding labels (supervised ML) and algorithms that can identify patterns in unlabeled data (unsupervised ML). In both cases, predictions for new data can then be made on the basis of the learned association or on the basis of the identified patterns. Specifically, unsupervised ML techniques can be used to detect patterns in textual data, e.g., through clustering algorithms (see lower part of Fig. 2). Thus, NLP and unsupervised ML techniques complement each other for pattern detection in textual data: NLP techniques provide numerical sentence embeddings which can then be clustered through unsupervised ML techniques (see Fig. 2).

These detected patterns form the basis for the next step of *pattern refinement*. Through a process of "computationally guided deep reading of the text" [12], the patterns (e.g., clusters) found in the pattern detection step are interpreted by the human analyst, adding quality and theoretical depth to the quantity and breadth of the first step's results, i.e., through deep reading of representative texts for specific patterns, the human analyst can generate a description of the patterns driven by substantive theory. These first two steps are potentially iterative, i.e., the results of the pattern refinement step may lead to a revised pattern detection step which prompts a new pattern refinement step, incrementally refining the interpretations of patterns until convergence is reached.

The last step, *pattern confirmation*, aims at testing that the patterns found and interpreted during the computationally guided deep reading in the first two steps are not an artifact of a specific NLP or ML technique or a biased interpretation by the human analyst. In other words, the extent to which the detected and refined patterns hold throughout the whole dataset is tested. For this purpose, supervised ML techniques that learn the association between the textual input data (more precisely their numerical sentence embeddings) and their associated cluster labels can be used. If the supervised ML technique is able to correctly predict cluster membership of sentences to a high degree, this will be regarded as an indicator that the identified patterns hold throughout the whole dataset. If the identified patterns do not hold throughout the whole

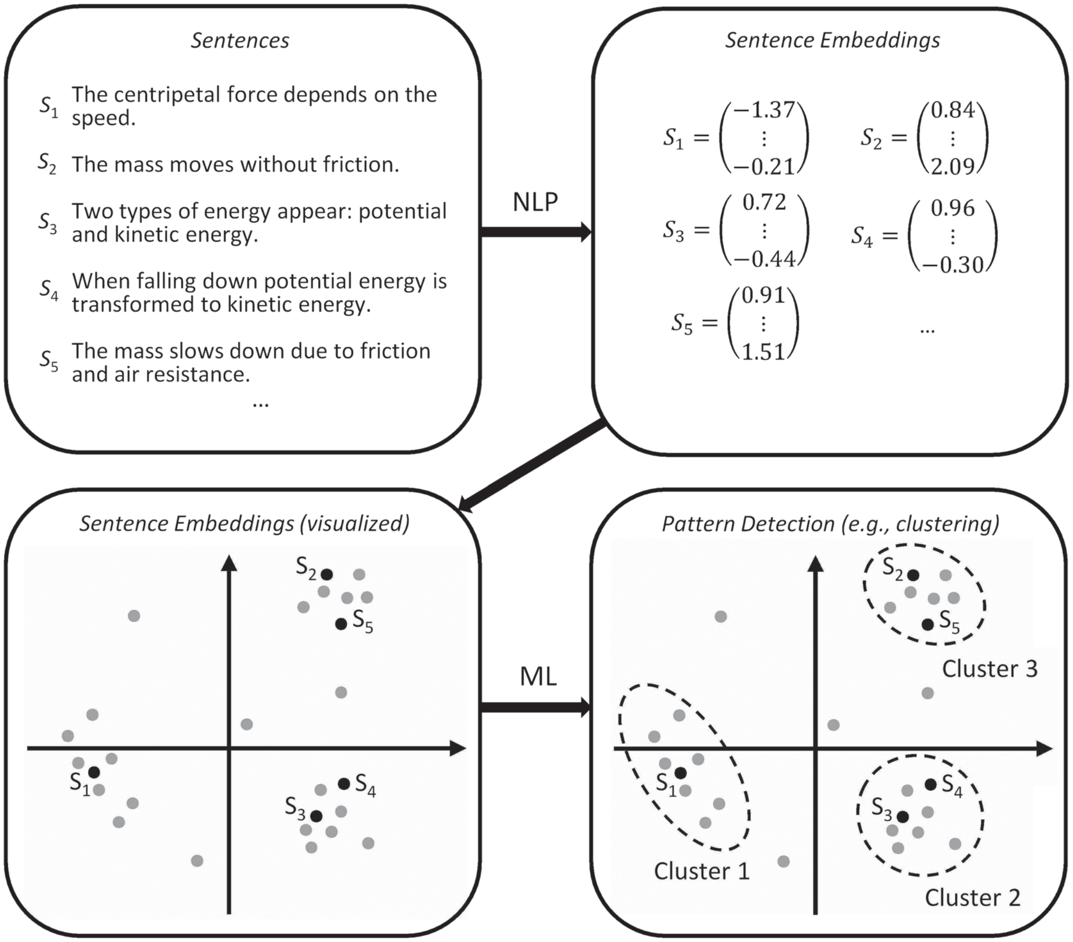
Sentences
S1 The centripetal force depends on the speed.
S2 The mass moves without friction.
S3 Two types of energy appear: potential and kinetic energy.
S4 When falling down potential energy is transformed to kinetic energy.
S5 The mass slows down due to friction and air resistance.
...
NLP
Sentence Embeddings
S1 = (−1.37 ⋮ −0.21)
S2 = (0.84 ⋮ 2.09)
S3 = (0.72 ⋮ −0.44)
S4 = (0.96 ⋮ −0.30)
S5 = (0.91 ⋮ 1.51)
...
Sentence Embeddings (visualized)
S1
S2
S3
S4
S5
ML
Pattern Detection (e.g., clustering)
Cluster 1
Cluster 2
Cluster 3

FIG. 2. Schematic representation of how natural language processing and machine learning serve as complementary tools for pattern detection in a sentence corpus.

dataset, one should use explanatory model analyses [29] and explainable AI techniques [30] to investigate this issue. This can then inform the (potentially iterative) revision of the pattern refinement step and—if necessary—the pattern detection step. Furthermore, the trained supervised ML model provides an algorithmic representation for automatic coding of new unseen textual data (taken from the same context). This could then be used by other researchers, however, generalizability of this automatic coding on unseen data should be checked first before confidently applying it.

When one compares CGT with traditional qualitative approaches, the iterative back and forth between the pattern detection and pattern refinement step is what sets CGT apart from traditional approaches the most. The pattern refinement step itself is very similar to traditional content analysis and the pattern confirmation step is very similar to establishing interrater reliability in qualitative analyses. In the back and forth between the pattern detection and pattern refinement step, however, the selection and contextualization of the material for analysis is guided by the computer. In a traditional approach, the analyst would engage differently with the material, often in a way guided by the structure of the material itself. Thus, computational guidance in CGT allows us to engage with very large data where lacking resources prohibit or limit traditional qualitative approaches. This guidance can lead to the analyst engaging differently with the data as how the material is approached is determined through the results of the computational pattern analysis and not by the decision of the analyst or the structure of the material. This way of engaging with the data may lead to insights which would have been unlikely to result from a traditional analysis.

Overall, CGT provides a framework that can help to avoid biased interpretations of qualitative data by human analysts and (too) shallow interpretations of qualitative data based only on quantitative properties. At this point, it is important to note that biased interpretations are not avoided because the quantitative approaches are bias free [31–33]. It is rather the iterative back and forth between human analyst and computational analysis that leads to this property of CGT.

III. AN APPLIED EXAMPLE

In the last section, we laid out the potentials that computational grounded theory offers as a framework for text analysis—a major and often challenging branch within qualitative methods. The main argument, however, holds for qualitative methods more generally: While there are various approaches to qualitative data analysis, at some point patterns are detected in the data. For human analysts, this step of pattern detection provides challenges, e.g., bias may slip in, subtle patterns may be hard to detect in large datasets, and some datasets are prohibitively large for human analysis. Specifically, with digital technologies on the rise we can expect educational (unstructured) datasets to become increasingly larger [34]. Artificial intelligence techniques such as natural language processing and machine learning have the potential to support human data analysts in the pattern detection step in large datasets. In this way, CGT provides a framework for integrating artificial intelligence-based methods into the qualitative data analysis process in a complementary way.

In this section, we will apply CGT to a PER problem. The analysis will highlight how CGT originally developed in sociology can be applied in PER and serve as a basis to discuss the potentials and challenges of CGT in PER—and the use of artificial intelligence in qualitative PER more generally. Specifically, we use CGT to investigate how students participating in the Physics Olympiad engage in problem solving and to what extent their approaches differ from students that do not participate in the Physics Olympiad.

In the next section, we first provide a brief overview of the relevant literature on problem solving in physics as well as its assessment and describe the context in which the textual data was gathered before diving into the actual application of CGT.

A. Research on problem solving in physics

Physics problem solving is recognized as an important ability to master when studying physics or engaging in a physics career [35,36], which is why it represents an important research topic in PER [22,37,38]. In the following, we consider problem solving for well-defined (i.e., textbook-style) physics problems. Those are problems in which students are confronted with an initial physics-related situation (e.g., throwing an object) and a goal state (e.g., determining the maximum height). The process of transforming the initial state to the goal state if the path to that goal is unknown is referred to as problem solving [39].

To model students' approaches to physics problem solving, researchers devised generic process models for solving well-defined physics problems that outline sequential phases in the problem solving process. These process models differentiate several phases in the problem solving process [40–42]. Even though all process model differ from each other to some extent, they all include a common core consisting of the phases: problem representation, strategy selection, execution of the strategy, and evaluation of the solution [40,42,43]. Adequately representing a problem is considered the crucial phase in problem solving [41] and requires processing lexical information as well as utilizing assumptions, idealizations, and physics concepts [44]. This phase of problem representation therefore involves transforming the given situation into a representation that makes sense from a physics point of view [40,41]. After the problem representation, problem solvers have to plan their solution strategy. Problem solvers can either access long-term memory problem schemata, i.e., abstracted solution procedures based on worked out examples on similar problems [3] or use their problem representation to more explicitly plan their solution [40]. Afterwards, the planned solution has to be executed. In physics, this typically involves quantitative aspects such as algebraic transformations and considerations on how to obtain relevant physical quantities [40]. Finally, the executed solution needs to be evaluated. These evaluations, for example, relate to checking for consistency of the solution with conservation laws, symmetries in the problem, or checking units of the solution [40].

Researchers particularly used expert-novice differences to understand differences in students' problem-solving approaches with the caveat that expert-novice dichotomies should not be considered mutually exclusive binary categories but rather a continuum that helps understanding group differences and hence provide diagnostic value. Compared to novices, expert problem solvers in physics were found to spend more time on representing the problem qualitatively based on fundamental physics concepts and to use more physics-specific assumptions and idealizations before applying mathematical operations [3,41,45,46]. Experts were found to have more conceptual knowledge elements than novices and to predominantly use problem schemata when solving problems while novices rely on single conceptual knowledge elements [47–49]. Previous research further established that experts and novices use different strategies to solve physics problems. Novices often start to solve problems by directly using mathematical operations and equations [42] without developing an adequate problem representation first [50].

Researchers commonly utilized language artifacts such as protocols (e.g., cognitive interviews or think-aloud studies) or constructed responses (e.g., written texts) to analyze students' problem-solving approaches since language-involved assessment formats can elicit important (qualitative) reasoning and knowledge of students [41]. Because of the generic nature of problem-solving process models and expert-novice differences, they are rather unspecific to the particulars of a specific physics problem at hand [51]. It would be desirable to employ principled and data-driven discovery methods that could systematically analyze constructed response formats (such as natural language data) while simultaneously accounting for a greater variety of students' problem-solving approaches. CGT may present a method of choice in this regard.

B. Example context

The applied example draws on data from a larger investigation into student science competitions (WinnerS) including the German Physics Olympiad [52]. Students who participated in the Physics Olympiad could voluntarily participate in a concurrent online study where they were asked to engage in a set of problem-solving tasks including the well-defined physics problem presented in Fig. 3. Moreover, data from a control group consisting of students who did not participate in the Physics Olympiad but were comparable in terms of school type and grade level were also collected but in a classic pen-and-paper format. As part of this study, students in both groups were given the following problem-solving task (translated from German to English) which also included the illustration presented in Fig. 3:

> A very small mass slides along a track with a vertical loop (see figure). The mass starts from a height above the highest point of the loop. Assume the motion to be frictionless. Determine the minimum starting height above the lowest point of the loop necessary for the mass to run through the loop without falling down. Describe clearly and in full sentences how you would solve this problem and what physics ideas you would use.

This task represents a typical, well-defined mechanics problem which—with variations—can be found in various physics textbooks [53]. By making simplifying assumptions (point mass, no friction, no rotation, loop is circular) and identifying the relevant physics concepts (conservation of energy, weight acting as centripetal force at highest point in loop), this problem can be solved using low-level mathematical operations.

Even though the problem instruction and solution space are rather well-defined, students' constructed responses exhibit variability, as is typical for language-based assessment even for simple instructions [54]. Language was coined to be ambiguous, noisy, and unsegmented [55]. Hence, we can expect any language products to expose a variety of expressions, terms, and text organizations that cannot easily be captured by holistic coding approaches [51,56]. To give an impression of the variability within students' textual descriptions of their problem-solving process, we decided to present three sample descriptions.[2]

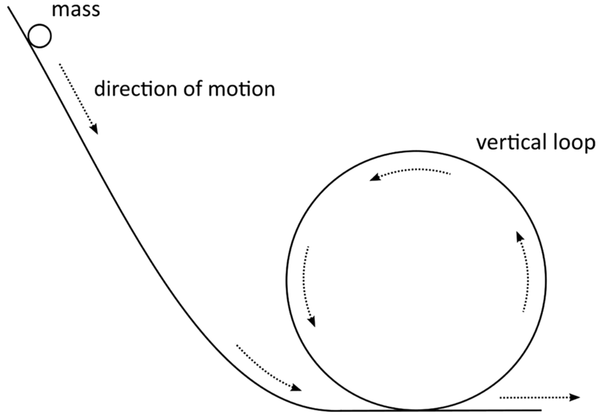

FIG. 3. Illustration of the vertical loop including the track of the sliding mass.

Student 1: *The mass should start from a point above the top of the loop to have enough momentum. The ball is then pressed against the track and cannot fall down.*

Student 2: *The mass performs a circular motion within the loop. At the upper point, the gravitational force must be equal to the centripetal force. The centripetal force can be calculated using the speed of the mass. This speed, in turn, can be calculated with the law of conservation of energy as the potential energy of the mass is converted into kinetic energy. In doing so, I assume that friction is negligible.*

Student 3: *First, calculate E_{pot} using $m \cdot g \cdot \Delta h$, then let the ball roll, i.e., E_{pot} becomes smaller, E_{kin} becomes larger* ($1/2\ \mathrm{mv}^2$). *At the point at the top, E_{kin} must be greater than the force pulling the ball downwards ($F = m \cdot g$) in order to not fall down.*

The sample descriptions also illustrate that the length of students' textual descriptions generally varies. In fact, in language analytics, text length is among the most predictive features for the quality of texts [57]. However, text length is not very informative and should be substituted by more substantive features. In particular, there exist major content-related differences between students' textual descriptions, i.e., students address different ideas within their textual description. Student 1, for example, directly states a solution with a short justification. However, this student did not adequately describe which physics ideas were used in deriving this solution. Student 2, on the other hand, emphasizes on physics ideas and (re-)states assumptions (i.e., circular motion within loop, no friction). Student 3 also focuses on physics ideas, however, this description is strongly interspersed with formulas. In particular, symbols of physics quantities are used as abbreviations for the corresponding quantity, e.g., $\mathrm{E}_{\mathrm{pot}}$ is used interchangeably for the potential energy of the mass. The three sample descriptions alone convey a first impression of how

[2] The original German texts contained varying amounts of spelling mistakes which were ignored for the translation into English.

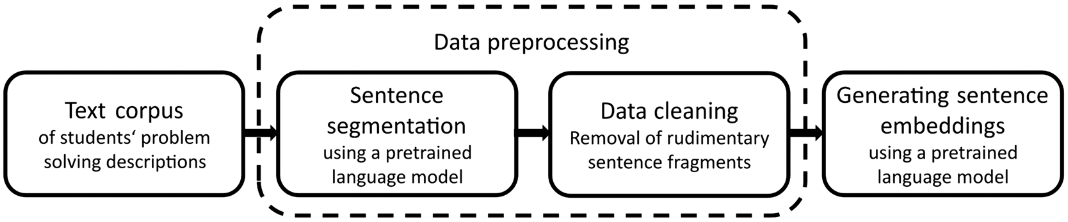
Data preprocessing
Text corpus
of students' problem solving descriptions
Sentence segmentation
using a pretrained language model
Data cleaning
Removal of rudimentary sentence fragments
Generating sentence embeddings
using a pretrained language model

FIG. 4. Visualization of the data preprocessing process which includes elements of NLP.

differently students engage in solving a typical physics problem. Moreover, it becomes evident that each description incorporates a different number of themes. Thus, when we intend to understand how students engage in physics problem solving, we cannot consider a student's description as one fixed unit, but have to consider it as being composed of several parts encompassing different themes [56].

C. The three steps of computational grounded theory

1. Pattern detection using human-centered computational exploratory analyses

Similar to traditional coding in qualitative data analyses, the first step of CGT also aims at detecting categories (or topics, themes, patterns, clusters, etc.) in a larger text corpus. However, in the context of CGT, not the researcher alone is trying to detect relevant categories in the text corpus as the researcher is supported through computational exploratory analysis techniques. However, to make the text corpus available to computational analyses, it needs to be preprocessed using NLP techniques. Specifically, all data (pre-) processing and analyzing was conducted using the free, widely used, and open source programming language Python [58].

(a) Data preprocessing.—To make students' raw textual descriptions processable for diverse kinds of computational analysis techniques, each text answer must be transformed into a numerical representation which is referred to as an embedding (see Fig. 4 for a visualization of the data preprocessing process). In line with standard NLP procedures and in accordance to the examination of the three sample descriptions in the last section, we concluded that a single student answer generally comprises different themes. Therefore, it seemed reasonable to break each student answer down into single segments such that each segment is assumed to relate to a specific theme. For practical reasons, we decided to split the student answers into single sentences as there exist tools that automatically perform this task. Specifically, we used a pretrained German language model [59] reporting a precision of 95% for sentence segmentation. Even though performing this task by hand is even less prone to error, it costs quite some time and is hardly scalable when processing large amounts of lengthy textual data. As a next step, all sentences which consisted of less than 20 characters were removed as we assumed that generally meaningful sentences are of greater length. Consequently, short noninformative sentences such as “No idea.” or “I don't know.” were removed from the sentence corpus. Furthermore, this approach also tackled another issue: Often, students interspersed their responses with an elaboration of physics concepts through mathematical formulas. However, the used language model for sentence segmentation was trained using newspaper and Wikipedia articles which is why formulas and equations were not always handled effectively by the model. More precisely, formulas and equations in students' textual descriptions were on some occasions torn apart resulting in sentence fragments such as “$r =$”, “E”, and “$m \cdot g \cdot h$”. Such fragments were generally of small length and therefore desirably removed by this step. However, a large share of sentences including formulas and equations was correctly extracted and therefore remained within the sentence corpus for the upcoming analyses.

After data preprocessing, a total of 1127 sentences belonging to the textual descriptions of $N = 417$ students (41% identified as female; grade level: $M = 11.1$, $SD = 0.8$) remained. From the overall 417 textual descriptions, 145 (35%) belonged to the group of Physics Olympiad participants (30% identified as female; grade level: $M = 11.1$, $SD = 1.0$), while the other 272 (65%) descriptions belonged to the group of nonparticipating students (47% identified as female; grade level: $M = 11.0$, $SD = 0.7$). The length of

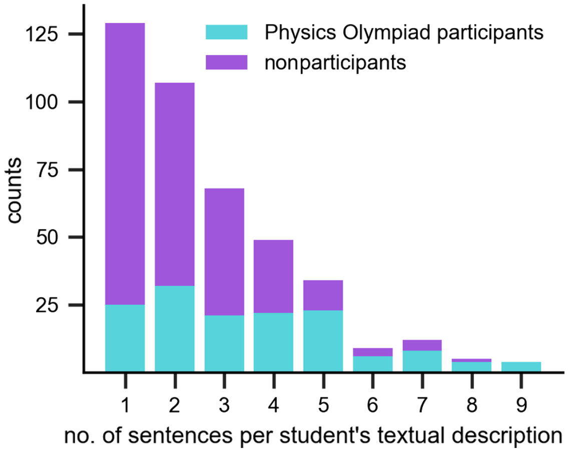

FIG. 5. Frequency distribution showing the number of sentences per textual description.

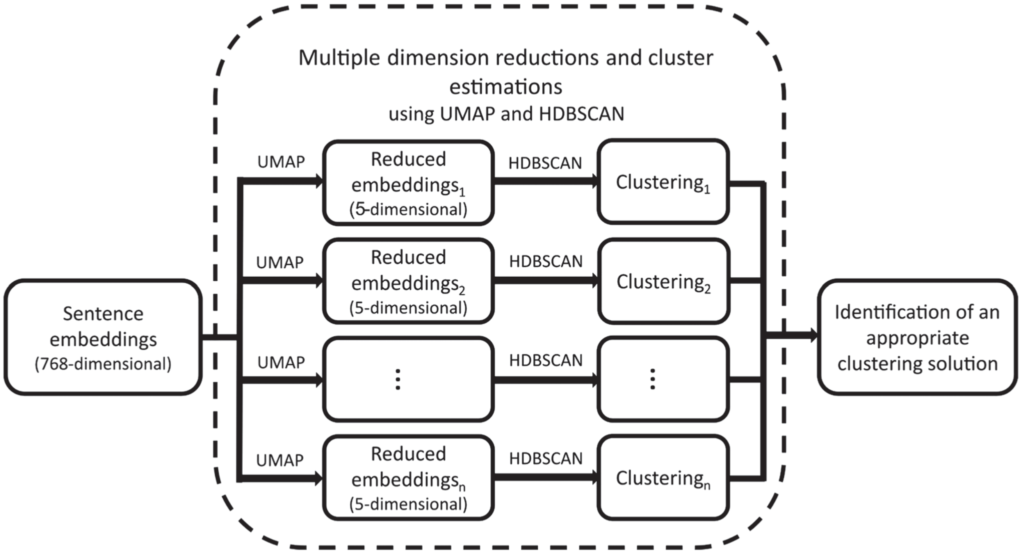

FIG. 6. Visualization of the process for obtaining an appropriate clustering solution using the dimension reduction procedure UMAP and the clustering procedure HDBSCAN.

students' textual descriptions showed high variability and shorter descriptions occurred more frequently than longer descriptions (see Fig. 5; No. of sentences per textual description: $M = 2.7$, $SD = 1.8$). A first comparison between Physics Olympiad participants and nonparticipants revealed that Physics Olympiad participants wrote on average more sentences per description ($M = 3.6$, $SD = 2.1$) than nonparticipants ($M = 2.2$, $SD = 1.4$).

(b) Generating sentence embeddings.—To tackle our research problem, the preprocessed data in the form of a sentence corpus must be transformed into some numerical representation which can then be used to detect patterns by applying quantitative methods (see Fig. 2). For this purpose, a variety of NLP methods are available—from simple bag-of-words models [60] up to more elaborated pretrained language models [28].

For our study, we used the Python framework and corresponding open-source library SentenceTransformers [61] as well as the pretrained language model German BERT [62–64] to handle students' textual descriptions in the German language. As schematically illustrated in Fig. 2, each of the 1127 sentences is transformed into a high-dimensional vector which encompasses the sentence's meaning. More precisely and technically speaking, all sentences were embedded into a 768-dimensional vector space, i.e., each sentence was mapped to a vector with 768 components.[3]

(c) Exploratory computational analysis.—We strive to cluster sentences based on their corresponding sentence embeddings. In order to obtain an appropriate clustering solution, a combined approach consisting of dimension reduction followed by clustering was employed (see Fig. 6).

We first inspected the nature of the data at hand. As a reminder, we had 1127 sentence embeddings—which may appear to be a lot already. However, these 1127 embeddings exist within a 768-dimensional embedding space. This poses a serious problem, as with increasing dimensionality, the volume of the embedding space increases exponentially fast so that the available data quickly become sparse (curse of dimensionality, see, e.g., Ref. [65]) which, furthermore, impedes locating dense clusters. Moreover, language and language data can be regarded as a complex dynamical system. Such systems often exhibit the relevant dynamics only in a few dimensions of their respective phase space [66] which in the case of language data can be considered the embedding space. Hence, we decided to initially perform a dimension reduction technique to find the lower-dimensional space that exhibits the relevant dynamics or language features in order to increase the performance of subsequent clustering approaches [67].

For this purpose, we applied a nonlinear dimension reduction technique called uniform manifold approximation and projection (UMAP) instead of relying on traditional linear dimension reduction techniques (such as principal component analysis or factor analysis) since there exists no argument (to our knowledge) that embeddings of sentences from the same topics can be described by linear subspaces within the high-dimensional embedding space. UMAP as a nonlinear technique seeks to simultaneously preserve the local structure of the embeddings while also revealing the global structure [68]. Specifically, it was shown that UMAP could remarkably improve the performance of well-known clustering procedures [69]. More precisely, the researcher specifies the target dimension of

[3] The dimension of the embedding space is predefined by the used pretrained language model.

this dimension reduction technique by selecting a value for the hyperparameter n_components [70]. We chose the dimension of the reduced embedding space as five similar to Wulff *et al.* [71] who performed similar analyses with a comparable dataset and retrieved well-interpretable and robust clusters. So UMAP provided us with five-dimensional reduced embeddings.[4]

Having reduced the dimension of the sentence embeddings, we intended to identify clusters in the newly obtained five-dimensional embedding space. At this point, the researcher must decide for one of many possible clustering procedures (e.g., K-means, hierarchical agglomerative clustering, etc.). The decision for a specific procedure depends on diverse aspects, e.g.: Is hard or soft clustering preferred? Must every sentence be categorized to a cluster or are "uncategorized" sentences allowed? Is the number of clusters that ought to be detected known beforehand?

For the applied example, we decided to use hierarchical density based spatial clustering of applications with noise (HDBSCAN; see [72–74]). This choice was made for two practical reasons: First, HDBSCAN does not require a predefined number of clusters which ought to be detected in the data. As our approach is exploratory by nature, we did not know *a priori* how many clusters to expect and also did not want to bias the clustering outcome by *a priori* specifying a concrete number of clusters. Second, this procedure does not compulsively assign all sentences (embeddings) to a cluster as nonassignable data are allowed to exist. The existence of such so-called noise data is particularly suitable for our purposes as students' textual descriptions of their problem-solving approaches are typically noisy in the sense that some sentences are difficult to categorize. For example, ca. 15% of text segments in a similar study on physics problem solving by Docktor *et al.* [51] were not captured by their coding manual. In general, HDBSCAN searches for dense regions of data within the embedding space. More figuratively speaking, this clustering procedure tries to find islands of higher density amidst a sea of noise where the sea level can be adjusted by HDBSCAN's hyperparameters, i.e., parameters that must be specified before applying the procedure.

There are two hyperparameters in HDBSCAN which are particularly important for the researcher. First, the researcher must decide on a minimum cluster size (min_cluster_size) which acts as a threshold distinguishing real (dense) clusters from noise. There exists no concrete advice for specifying this parameter in the literature, however, we propose two advices that might help researchers who intend to apply this method. First, it is known that choosing a too small value for this parameter results in too many small clusters that are often hardly distinguishable from noise while too high values result in only few quite large clusters (see Supplemental Material, Part A [75]). Therefore, we first advise that this parameter should be chosen somewhere in-between these two extreme cases. Second, the researcher should think about the minimum number of sentences needed in a cluster to reliably interpret those. In the applied example, we decided that there should be at least 15 sentences per cluster to make sense of them. Other researchers, however, might choose a different value for this hyperparameter based on the characteristics of their data and research question.

Second, the researcher can also control to some extent how conservative the clustering will be by altering the hyperparameter min_samples. By default, the value of this parameter equals the value of min_cluster_size. Manually increasing min_samples results in a more conservative clustering, i.e., more points would be declared as noise and clusters would be restricted to even more dense regions, while decreasing this hyperparameter results in a less conservative clustering, i.e., less noise and likely a greater number of clusters. We decided against such a less conservative clustering (and hence used the default value of min_samples) based on the following reasoning: Both more and less conservative clustering can accurately cluster sentences that are clearly related in content. They differ, however, in their handling of noisy sentences, i.e., sentences that HDBSCAN cannot straightforwardly categorize into a specific cluster. More conservative clustering does not force such sentences into a specific cluster and declares them as belonging to the noise category, leaving them for human intervention in the second CGT step of pattern refinement (see Fig. 1). Contrary, by applying a less conservative clustering, HDBSCAN more frequently forces such noisy sentences into a specific cluster. We now argue that a human coder (typically familiar with the domain) is superior in accurately categorizing noisy sentences compared to HDBSCAN, because a human coder possesses relevant domain knowledge in addition to hermeneutic skills while HDBSCAN solely relies on domain-unspecific sentence embeddings. Thus, it makes more sense to apply a more conservative clustering (i.e., better interpretable, less noisy clusters and a larger noise category for human intervention) than a less conservative clustering (i.e., less interpretable clusters because HDBSCAN forced more noisy sentences into clusters). For exploratory reasons, we nevertheless applied the least conservative clustering by setting min_samples to its smallest possible value (see Supplemental Material, Part A [75]).

An issue results from the dimension reduction technique UMAP as it incorporates a stochastic component resulting in slightly different embeddings from run to run. These slightly different embeddings, in turn, result in slightly different clustering solutions using HDBSCAN, i.e., the

[4] A sensitivity analysis at the end of the pattern detection step revealed that increasing the dimension of the reduced embeddings notably increased computation time, however, similar clusters emerged.

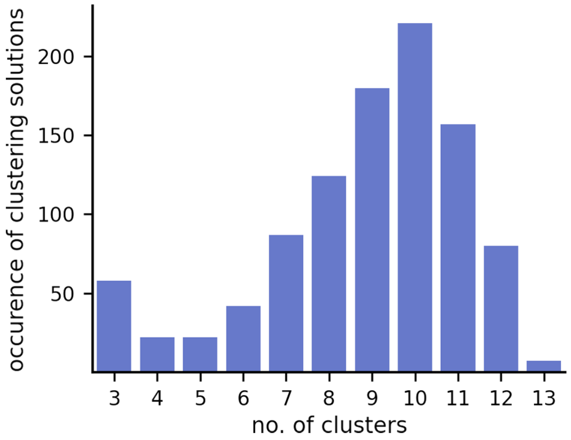

FIG. 7. Frequency distribution of the number of clusters within the 1000 different clustering solutions.

total number of clusters as well as the specific form of clusters varies. However, specific outcomes can be reproduced by fixing the random seed parameter of UMAP. In practice, in order to obtain a clustering solution in which all frequently occurring clusters appear, we did the following (see also Fig. 6): First, using different random seeds, we applied UMAP directly followed by HDBSCAN for $n = 1000$ runs. This resulted in 1000 different clustering solutions. Second, we examined how often solutions with a specific number of clusters appeared. Results are depicted in Fig. 7 and show that solutions with ten clusters appeared the most (i.e., ten is the mode of the distribution in Fig. 7). Third, 20 of those clustering solutions having ten clusters were randomly sampled and visually inspected using two-dimensional embedding plots while taking the similarity in the three most relevant words per cluster into account (for a more detailed description of this visual inspection, see Supplemental Material, Part B [75]). Alternatively, more rigorous and data-driven procedures can be constructed, however, identifying corresponding clusters across solutions is not trivial, specifically if a cluster in one solution is split into two or more smaller clusters (i.e., constituting clusters) in another solution. Hence, we decided to rely on a more visual inspection of clustering solutions which is largely unambiguous if there are only few and largely separable clusters (as in our applied example). For obtaining two-dimensional embedding plots, the original (768-dimensional) sentence embeddings were reduced to two dimensions using UMAP, while keeping the assigned cluster labels by HDBSCAN on the reduced five-dimensional embeddings. The three most relevant words per cluster in each solution were determined using TF-IDF scores (see next section). Following the visual inspection procedure, we identified 12 constituting clusters that occurred frequently among the inspected 20 solutions, i.e., clusters formed by merging constituting clusters were not considered. We decided to consider all these 12 constituting clusters for further investigations as considering more clusters may give rise to interesting findings which a human might overlook when reading larger text corpora. Moreover, if some of these clusters proved useless further on, they could be discarded or merged together in the second step of CGT. Thus, a clustering solution consisting of the 12 previously identified constituting clusters was chosen for the upcoming analyses.

The chosen clustering solution is graphically represented in Fig. 8 using a two-dimensional embedding plot in which the identified clusters are highlighted in colors. Even in two dimensions, the structure of the five-dimensional clusters seems to be preserved. Moreover, a condensed tree plot of the final clustering solution is shown in Fig. 9. Such

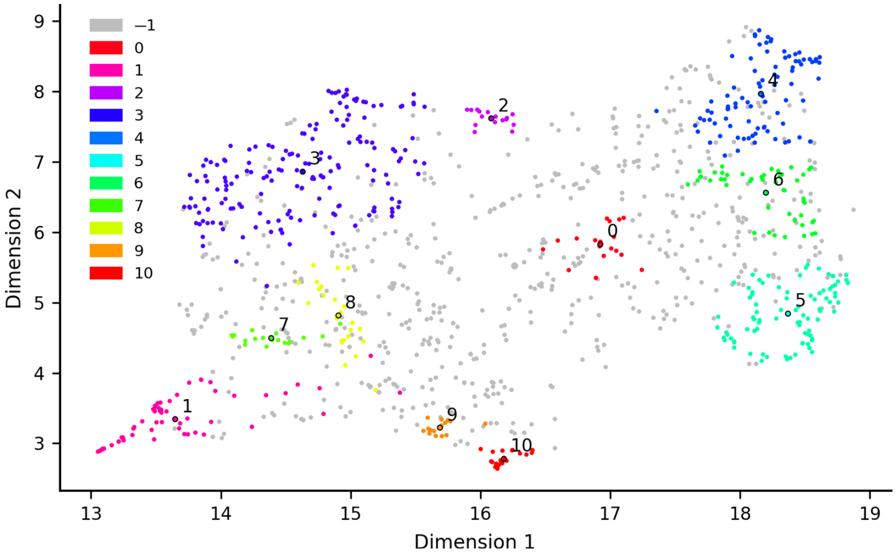

FIG. 8. Representation of detected clusters within a two-dimensional embedding space. Overall, eleven clusters labeled from 0 to 10 and an additional noise cluster labeled as −1 were detected. Black bordered points represent cluster centroids.

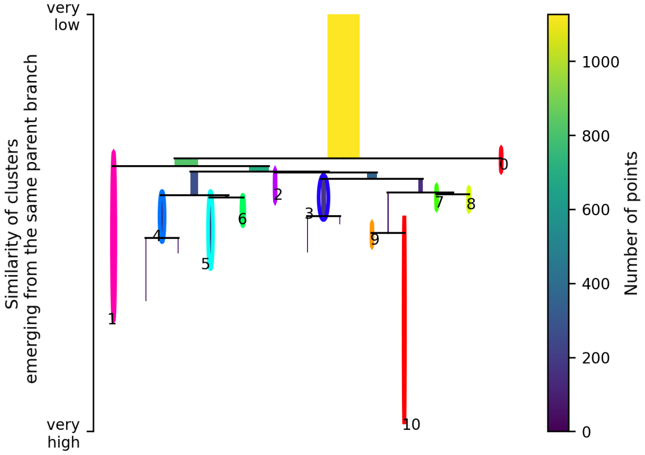

FIG. 9. Condensed tree plot of the final clustering solution.

condensed tree plots visualize the cluster hierarchy, i.e., they illustrate data-based similarities between the detected clusters. Specifically, the vertical position of a parent branch from which two clusters emerge represents a measure for the data-based similarity of both these clusters (see left axis of Fig. 9). For example, the parent branch connecting clusters 9 and 10 is located furthest down which means that both these clusters are the most similar among all detected clusters. The next two most similar clusters are then clusters 5 and 6. These data-based similarities generally translate to semantic similarities when interpreting the detected clusters. However, they can also be misleading, i.e., two clusters may mainly use similar words resulting in higher data-based similarity even though the clusters differ essentially in meaning. Therefore, there exists no precise rule prescribing in which case two clusters must be merged based on the tree plot. In other words, a condensed tree plot should not be considered the driver for the decision on merging clusters. Rather, hypotheses for meaningful merging should be formulated by the human analysts with regards to the research problem based on deep reading and substantive theory, and then buttressed in a data-based way, for example, by using condensed tree or embedding plots. This way, condensed tree plots represent a helpful tool in the sense of triangulating evidence when thinking about merging clusters in the following pattern refinement step of CGT.

2. Pattern refinement using human-centered interpretation

In the second step of CGT, the focus shifts from detecting clusters to qualitatively analyzing them with the aim to confirm the clusters' plausibility, add interpretation, and potentially refine the clusters [12]. In this step, the researchers used inductive qualitative content analysis, i.e., they read parts of the text corpus in what Nelson [12] calls a "computationally guided deep reading". Texts for deep reading were generated based on the clusters detected in the preceding step in two ways: (i) determining the most relevant words in a cluster and (ii) determining the most representative sentences for a given cluster.

For determining the most relevant words in a cluster, a class-based variant of TF-IDF (term frequency-inverse document frequency) scores was used [76–78]. First, a single document for each cluster was created by joining all sentences belonging to a specific cluster. Second, words that do not add much meaning to a sentence (so-called stopwords, e.g., articles) were automatically removed. Third and finally, TF-IDF scores were computed for each word in every cluster-specific document [78]. Essentially, this score compares the relative frequency of a word in a (cluster-specific) document to the inverse frequency of that word in all documents. Therefore, words within a cluster-specific document that have the highest TF-IDF scores are technically the most relevant for that specific cluster as these words appear particularly often within a specific cluster while also appearing relatively less often in the overall text corpus. The ten most relevant words per cluster determined by this procedure can be found in Table I.

To determine the most representative sentences per cluster, cluster centroids, i.e., mean vectors of all sentence embeddings belonging to a specific cluster, were computed. Then, for each sentence in a cluster, the (Euclidean) distance between this sentence embedding and the corresponding cluster centroid was computed. The smaller the distance between a sentence embedding and its corresponding cluster centroid, the more this sentence was considered representative of that specific cluster [79]. In this way, the 15 most representative sentences per cluster were determined (see Table I for a sample representative sentence per cluster).

Based on the most relevant words and most representative sentences per cluster, the authors of this manuscript and an additional research associate independently developed a characterization for each cluster. Here, the analysts engaged with the computationally selected texts as they would engage with text in any other inductive qualitative content analysis. The researchers began with an open, exploratory coding, allowing for the development of new codes and themes as they arose. Once the initial codes had been developed, the researchers grouped them into broader themes that captured the essence of the findings. These themes were continuously refined and reorganized as the analysis progressed. Then, the independently developed characterizations were triangulated, i.e., they were compared and a consensus characterization for each cluster was determined (see Table I). In this process, the involved researchers' substantive knowledge of the domain (problem solving in physics) as well as their expertise as teachers helped to inform the decisions. In the following, each cluster will be briefly described. Additionally, the number of sentences within each cluster will be given in brackets.

Cluster 0 ($n = 18$) contains *assumptions* and *idealizations* that are made explicit by the students. Mostly, it is

TABLE I. Sizes (number of sentences), ten most relevant words, sample sentences, and devised definitions of the detected clusters. Note that there are diverse reasons for multiple occurrences of the same word within the list of ten most relevant words in a cluster: (1) A German word appeared correctly as well as incorrectly spelled in the list, however, this misspelling was ignored during translation to English. (2) The same German word appeared in different grammatical cases in the list, however, the different grammatical cases have the same English translation. (3) There existed different German words having the same meaning in the list, however, for all of them there existed only a single corresponding English word.

Cluster	Size	Ten most relevant words	Sample of a representative sentence	Definition
−1	565	⋯	⋯	Noise cluster
0	18	friction, air resistance, neglected, air resistance, decrease, but, present, gained, first, already	All following calculations ignored friction and air resistance.	Assumptions or idealizations such as neglecting friction and air resistance are made.
1	58	2r, v^2, pot, kin, energy, mg, holds, solve, formula, hold	Energy in P1 is only potential energy $E_{\text{pot1}} = \text{m} \cdot \text{g} \cdot \text{h}$.	Physics ideas which are strongly interspersed with formulas and equations.
2	16	think, loop, height, say, higher, enough, minimum, ca, loop, believe	I believe that the minimum starting height must be higher than the loop in order to build up enough velocity.	Guesses for the minimum starting height, mainly without explanations but occasionally with short explanations and including colloquial wordings.
3	169	velocity, calculate, calculate, mass, at first, loop, acceleration, required, height, calculate	At first, I would calculate the velocity reached at the beginning of the loop.	General descriptions of high variability on how students would proceed at specific points during problem solving.
4	87	loop, height, start, starting height, at least, minimum, point, highest, highest, loop	The mass must start at a starting height that is at least at the height of the highest point of the loop.	Formulations of an initial hypothesis or a final solution for the minimum starting height.
5	86	weight, greater, centrifugal force, centripetal force, mass, point, gravitational force, acts, equal, loop	To prevent the mass from falling down at the highest point of the loop, the mass must have a centrifugal force greater than or equal to its own weight at the highest point.	Statements of what forces act on the mass within the loop and/or how they must relate for the mass to pass the loop.
6	48	mass, enough, loop, velocity, loop, down, reaches, high, fast, possible	At the highest point of the loop the mass must have a velocity so that the mass is fast enough to not be pulled down by gravity.	Qualitative formulations of conditions that must be met so that the mass can pass the loop.
7	18	equation, results, both, equate, equated, multiplied, solved, rearrange, forces, obtains	Consequently, both forces are equated and the equation is simplified.	The usage of mathematical operations is described.
8	24	can be, calculate, velocity, calculated, dependence, with the help of, radius, height, derive, using	By calculating the centrifugal force on the ball and comparing it to the gravitational force, the mass of the ball can be determined and its starting height can be calculated.	Descriptions on how one physical quantity can be determined from other physical quantities or physics concepts, in some instances formulas are verbalized.

(Table continued)

TABLE I. *(Continued)*

Cluster	Size	Ten most relevant words	Sample of a representative sentence	Definition
9	16	energy, kinetic, kinetic, potential, sum, potential, potential, equated, potential, starting point	Additionally, a consideration of energy yields that the sum of potential and kinetic energy of the mass is constant at all times.	Problem is modeled from an energy perspective, i.e., relevant energy forms are identified and a focus is placed on the conservation of energy.
10	22	converted, kinetic, energy, potential, converted, at, potential energy, completely, mechanical, downhill	When the mass reaches the ground, all the potential energy has been converted into kinetic energy.	Problem is modeled from an energy perspective while focusing on the conversion of potential to kinetic energy and vice versa.

restated from the problem description that friction is neglected, which also includes neglecting air resistance. Other assumptions and idealizations occur less often and include, for example, considering the mass as a point mass, neglecting rotational energy, and modeling the loop as a circular path.

Clusters 5, 6, 9, and 10 ($n = 86$, 48, 16, and 22) all include sentences that explicitly or implicitly focus on *conceptual aspects*, i.e., on physics concepts which are relevant for solving the underlying physics problem (i.e., conservation of energy, weight acting as centripetal force at highest point in loop). Cluster 6 incorporates sentences in which these relevant concepts are indirectly addressed as conditions are formulated that must be met so that the mass can pass the loop. By contrast, cluster 5 includes concrete statements on what forces act on the mass within the loop and how these forces must relate for the mass to pass the loop. Therefore, this cluster explicitly focuses on the force concept which is necessary to successfully solve the physics problem in a classical fashion. In clusters 9 and 10, the physics problem is modeled from an energy perspective and relevant forms of energy are identified. Specifically, there exists a small distinction between those two clusters as cluster 9 places a focus on the conservation aspect of energy, while cluster 10 focuses on the conversion of energy. At this point, it must be said that, even though these clusters incorporate physics concepts and ideas, some of those are inaccurate, ambiguous, or even physically incorrect.

Clusters 1, 7, and 8 ($n = 58$, 18, and 24) combine more *quantitative aspects* of the students' textual descriptions on how to solve the problem at hand. Of those clusters, cluster 1 can be considered the most extreme as it contains sentences that are strongly interspersed with formulas and equations. By contrast, cluster 7 includes sentences describing the usage of mathematical operations such as equating physical quantities or rearranging and solving equations. Cluster 8 includes descriptions on how one physical quantity can be determined from other quantities or by using a specific concept which also includes verbalizations of formulas.

Clusters 2 and 4 ($n = 16$ and 87) both include sentences in which a *solution* to the problem at hand or a *hypothesis* for the solution is formulated. Both clusters are highly similar, however, statements in cluster 2 seem more like guesses which are sometimes justified by short explanations that also include colloquial wordings. Moreover, in more than half the cases, the sentences of cluster 2 belonged to a description consisting of only one sentence. By contrast, sentences in cluster 4 belonged more often to descriptions that consisted of more than one sentence. In those cases, the sentences in cluster 4 were mainly the first or the last sentence in the complete description, which is why these sentences could be considered an initial hypothesis or a final solution statement.

Cluster 3 ($n = 169$) was by far the largest cluster if noise is excluded and consisted of *general descriptions* of high variability regarding specific phases when solving the problem at hand.

However, approximately half of all sentences (i.e., 565 of the overall 1127 sentences; 50.1%) were categorized as noise which might seem disadvantageous at first glance. However, it must be emphasized that the goal of the pattern detection step in CGT is not to categorize as many sentences as possible. Quality comes before quantity in the sense that it is only necessary to get a minimum number of sentences per cluster that are categorized confidently, such that the human analyst can develop valid descriptions of clusters. Moreover, in this pattern refinement step of CGT, researcher should closely examine this noise category. By doing so, we realized that a large proportion of noise sentences could be well assigned to one of the other detected clusters. Moreover, for some sentences another similar cluster seemed more reasonable than the computationally assigned cluster. So having developed an enhanced understanding of the detected clusters beforehand, all sentences in the noise category (as well as sentences in the other clusters) were reviewed by a human coder and eventually relabeled based on the developed cluster characterizations (see Table I).

Overall, 74.0% of sentences (418 of 565) originally categorized as noise were relabeled while 34.5% of sentences (194 of 562) in non-noise categories were relabeled. Most sentences that were relabeled required a low level of inference [13]. For example, the two sentences "For the mass to not fall down, the centripetal force must be greater than or equal to the weight of the mass" and "Since the mass is a point mass, there is no rotational but only translational energy" were originally labeled as noise in the pattern detection step. However, by observing the developed cluster characterizations in Table I, one can confidently tell that the first sentence addresses the force concept and therefore should be relabeled as belonging to cluster 5. In a similar manner, one realizes that the second sentence states an assumption which is why it was relabeled as belonging to cluster 0. Moreover, some sentences such as "Since the movement is frictionless, no energy is lost" were labeled as noise. For a human coder, such sentences proved ambiguous in the sense that they could be categorized as belonging to one or another cluster. For example, the above sentence incorporates the idealization of a frictionless motion (cluster 0) but also the concept of conservation of energy (cluster 9). A human coder must decide in such instances which cluster (or category, etc.) suits best, however, different coders will likely make different decisions even if an existing coding manual might be overly verbose. An example for a sentence in a non-noise cluster that was relabeled is "I assume that friction is negligible." This sentence was categorized into cluster 3 (general description) even though it represents an assumption (cluster 0) which is why it was relabeled. In summary, given the mainly low level of inference [13] required for relabeling sentences in our study we considered one coder to be sufficient. However, reliability of this step can be generally improved by having multiple coders. Thus, we advise anyone who is applying CGT in their own research to carefully consider the required number of coders to ensure reliable relabeling.

At some point, considerations should be given on whether the detected clusters were sufficiently distinct and meaningful for the purpose of tackling the research problem. Based on the relevant theory on problem-solving processes in domains such as physics and the developed cluster characterizations (e.g., see Table I) as well as in view of our research problem, it seemed reasonable to reduce the complexity of the clustering and merge clusters that are similar in content. Specifically, as we were interested in how students participating in the Physics Olympiad engage in problem solving and to what extent their approaches differ from students that do not participate in the Physics Olympiad, we decided to merge clusters that seemed to represent the same phases or have the same function within a typical problem-solving process.

The developed cluster characterizations (Table I) point out that both cluster 9 and cluster 10 revolve around the physics concept of energy. Moreover, they are the most similar clusters based on the condensed tree plot (Fig. 9) which is why merging those two clusters seems reasonable. Merging clusters 5 and 6 also seems reasonable as they both (directly and indirectly) revolve around the relation of relevant forces for the mass to pass the loop and the condensed tree plot buttresses this similarity between clusters. Thus, we decided to merge clusters 5, 6, 9, and 10 into the global cluster *Conceptual Aspects* as they all relate to the application of physics concepts and it is irrelevant in view of the research problem to differentiate between the energy concept (clusters 9 and 10) and the force concept (clusters 5 and 6). Clusters 7 and 8 are also substantively similar as they both revolve around quantitative aspects appearing in typical problem-solving processes. The condensed tree plot also adds evidence for the similarity between those clusters. However, cluster 1 includes sentences strongly interspersed with formulas and equations and therefore also has a quantitative focus. Even though the condensed tree plot suggests no similarity between cluster 1 and clusters 7 and 8, they are all three situated in the same region within the embedding space (see Fig. 8) indicating some data-based similarity. Based on these considerations, we decided to merge clusters 1, 7, and 8 into the global cluster *Quantitative Aspects* as all three clusters revolve around formulas, equations, and how they can be manipulated. Lastly, even though neither the condensed tree plot nor the two-dimensional embedding plot suggest a data-based similarity between clusters 2 and 4, we decided to merge them into the global cluster *Formulation of a Solution* due to their strong substantive similarity that we established during deep-reading. Cluster 0 and cluster 3 remained untouched and made up the global clusters *Assumptions and Idealizations* and *General Descriptions*.

In summary, the second step of the CGT framework provided us with five final themes in students' problem-solving approaches that inductively emerged from the data based on the human interpretation of the computationally identified patterns with the help of additional tools such as embedding plots and condensed tree plots. Those resulting themes and their corresponding global clusters are illustrated in Fig. 10.

The resulting five global clusters are illustrated in Fig. 10 and represent specific themes within the problem-solving process.

3. Pattern confirmation

As the final set of five themes might be an artifact of the specific clustering procedure used in the first CGT step or a consequence of biased interpretations or inappropriate

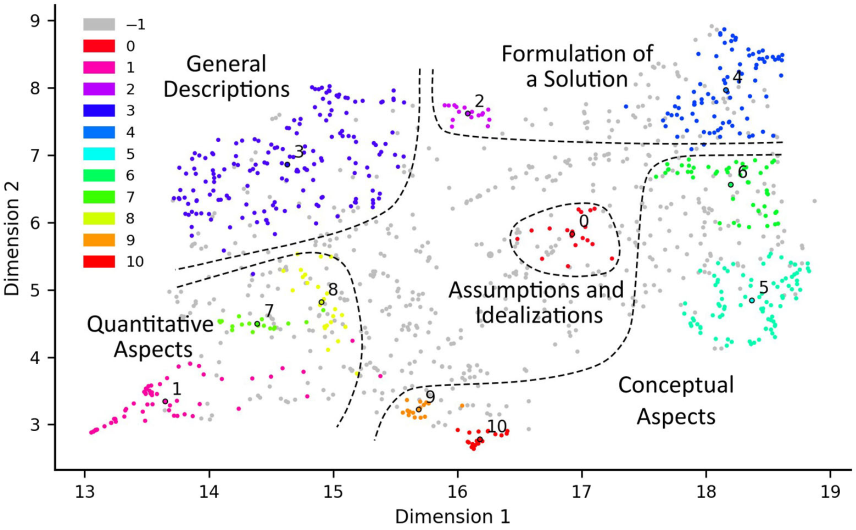

FIG. 10. Final five global clusters representing specific themes within the problem-solving process and how these global clusters relate to the initial detected clusters as indicated by the dashed black lines.

summarizations in the second CGT step, it is necessary to test whether the inductively identified themes hold throughout the entire text corpus. More precisely, the computational techniques in the pattern detection step may struggle with interpretative aspects of language such as ambiguity, colloquialism, and irony. These more interpretive aspects of language are to some extent checked by the human expert in the pattern refinement step. However, only a subset of data is generally used for interpretive deep reading in this step which impedes reliability in addition to the issue of bias incorporated by the human expert. Therefore, this step can be considered a final reliability check for the CGT process on the analyzed text corpus [12].

For this purpose, a supervised ML technique called relevance vector machine (RVM) was employed as this technique particularly allows a probabilistic prediction of themes (classes, clusters, etc.) for input data [65]. In terms of the applied example, the RVM represents a classifier and allowed to estimate probabilities of membership to each of the five identified themes based on five-dimensional reduced sentence embeddings as input data. The RVM was trained with all except the noise data as the noise data's high substantive variability would complicate model training and likely downgrade subsequent predictions of themes.

To determine the accuracy of the RVM we used tenfold cross validation [65,80]. The training dataset (sentence embeddings and corresponding themes) is split into 10 (approximately equally sized) bunches. Now, the RVM is trained using all those bunches except one. The RVM is then tested by predicting the themes of the sentence embeddings in the one remaining bunch. This procedure is repeated so that themes of sentence embeddings were predicted for each bunch while the remaining bunches were always used for model training. So in total, tenfold cross validation involved training ten additional RVM—each was trained on approximately 90% of the data while the remaining 10% were used for testing. Finally, the overall predictive accuracy, i.e., the percentage agreement between the predicted themes and the previously assigned themes, is computed across all bunches. Predictive accuracy was 0.76, i.e., around three-quarters of the RVM's predictions were correct (for further performance measures see Supplemental Material, Part C [75]).

As ML algorithms can exhibit bias, i.e., in the form of gender or racial bias [33,81], we also investigated to what extent the final model exhibited bias. Because of the known systematic differences regarding gender in science competitions [82], we focused on gender and computed predictive accuracy separately for students who identified as female and male. Similar accuracies were found for both the female and the male gender group with accuracies of 0.77 and 0.76, respectively.

All the above results suggest that the identified patterns are a characteristic of the dataset and not an artifact of the specific clustering procedure or a biased interpretation of the clusters. Therefore, the identified global clusters can be reliably used to tackle the research problem. To do so, the relative frequencies of the five themes and the "uncategorized" category for the Physics Olympiad participants and nonparticipants were compared (see Fig. 11).

In Fig. 11, one observes that the identified themes vary in their frequencies within the textual descriptions of Physics Olympiad participants and nonparticipants. A chi-square test (for homogeneity) revealed that the two frequency distributions are indeed significantly different: $\chi^2(df = 5) = 135.2$; $p < 0.001$. Regarding differences in occurrence of themes between Physics Olympiad participants and nonparticipating students, it can be seen that Physics Olympiad participants referred three times more frequently to *assumptions and idealizations* in their descriptions than nonparticipants.

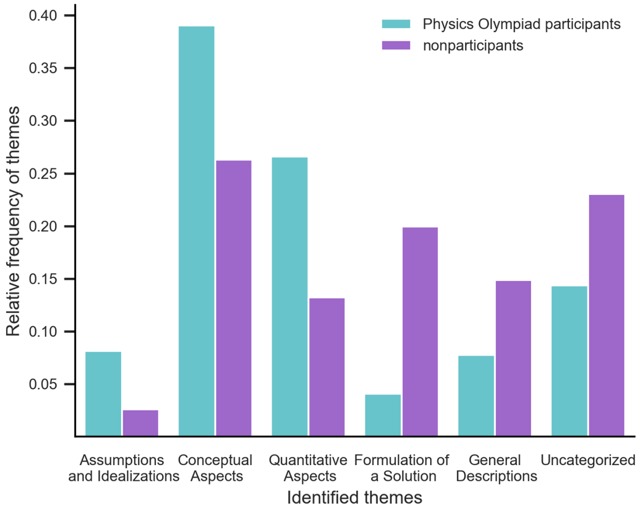

FIG. 11. Relative frequencies of the five themes (and uncategorized sentences) within textual descriptions of Physics Olympiad participants and nonparticipants.

Physics Olympiad participants also referred 50% more often to *conceptual aspects* and twice as often to *quantitative aspects* than nonparticipants. Nonparticipants performed almost five times as often a *formulation of a solution* and gave approximately twice as often *general descriptions* than Physics Olympiad participants within their textual descriptions. Lastly, nonparticipants' textual descriptions included 50% more often sentences that did not belong to any of the five themes than descriptions of Physics Olympiad participants.

Before continuing with a discussion of these results, we want to highlight that we obtained a trained classifier (i.e., an algorithmic representation for coding of new text, see Fig. 1) as a by-product of training the RVM in the third step of CGT. Specifically, the RVM is able to predict themes of unseen input data (in the form of embeddings) and thus could be used for automatic coding of students' textual descriptions to the problem at hand. However, we want to clarify that in this case and more generally, it would be necessary to test the RVM on an unseen data corpus to test generalizability before confidently using it for automatic coding purposes.

Nevertheless, in order to demonstrate this classifier's abilities, we used it to categorize some unseen input data. Specifically, we obtained an output table including the extracted sentences from the input data, corresponding probabilities of membership to each theme, and the final assigned theme based on the maximum membership probability (see Table II). In this example, the classifier performs quite well except for sentence No. 5 in which the classifier is unsure between two themes.

D. Discussion of the applied example

In this section, we discuss the results of our analysis in light of the research problem before delving into a more general discussion of CGT for PER.

Overall, we found that when students engage in problem solving, their approaches can be well described using existing process models of problem solving in physics [40–43]. We identified five themes that map well onto the important phases in problem solving [40]. The assumptions

TABLE II. Output of the trained classifier using an unseen textual description as input. Abbrevations are assumptions and idealizations (AI), conceptual aspects (CA), quantitative aspects (QA), formulation of a solution (FS), and general descriptions (GD).

		Probability of membership to					
No.	Sentences	AI	CA	QA	FS	GD	Assigned themes
1	I assume the motion to be frictionless and circular.	0.95	0.01	0.03	0.02	0.0	AI
2	Furthermore, I consider the mass as a point mass so that no rotation takes place.	0.73	0.22	0.04	0.01	0.0	AI
3	Then I think about which physics laws play a role here.	0.01	0.08	0.10	0.01	0.81	GD
4	At the highest point of the loop, the weight of the mass must be smaller than the centripetal force.	0.0	0.93	0.06	0.01	0.0	CA
5	I think that the centripetal force only depends on the velocity of the mass and on the radius of the loop.	0.0	0.41	0.56	0.02	0.01	QA
6	The velocity at the highest point can be calculated using the law of conservation of energy.	0.0	0.20	0.71	0.02	0.07	QA
7	The initial potential energy is converted to kinetic energy, and at the highest point of the loop the mass has both kinetic and potential energy.	0.0	0.84	0.15	0.0	0.0	CA
8	By solving the energy equation for the velocity and plugging this into the equation for the forces, one gets the minimum starting height after rearranging the resulting equation.	0.0	0.05	0.90	0.01	0.04	QA
9	In any case, the required starting height is greater than the height of the loop.	0.01	0.01	0.06	0.93	0.0	FS

and idealizations theme and the conceptual aspects theme map well onto the problem representation phase, while elements of the general descriptions theme and again the conceptual aspects theme fit well with the phase of strategy selection. Physics concepts play an important role for problem representation and also for strategy selection, while in practice, it can generally be hard to distinguish between these phases [40]. The quantitative aspects theme maps onto the execution of the solution phase. The evaluation phase is somewhat underrepresented but can also be found to some extent in the formulation of a solution theme. This underrepresentation can be understood if we recall that the students were not asked to solve the physics problem completely but rather to describe how they would solve it. Therefore, this prompt already suggests that students' textual descriptions would generally focus on the representation of the problem and strategy selection, while the execution and also the evaluation of the solution as later phases within problem solving might be of minor importance within textual descriptions. To summarize, the fact that the themes that emerged from the data through our analyses map to existing phases in problem solving adds further support to the usefulness of the existing problem-solving process models [40–43] in PER.

Regarding the differences in problem-solving approaches between students participating in the Physics Olympiad and nonparticipating students, our results indicate that Physics Olympiad participants showed on average more expertlike problem-solving behavior as their problem solving seemed to be characterized on average by a more prominent use of fundamental physics concepts [48] and, on average, by a better articulation of the problem representation in terms of assumptions, idealizations, and physics concepts [3,41,44]. Specifically, we found during deep reading that Physics Olympiad participants more frequently identified both relevant physics concepts for the problem at hand instead of just one or none which relates to expert problem solvers better developed conceptual knowledge [47–49]. Contrary, nonparticipating students seemed to exhibit on average more novicelike problem-solving behavior. Specifically, nonparticipating students fall on average short of assumptions and idealizations even though making assumptions is essential during problem solving [44]. An explanation for this might be that they seem to be less frequently aware that specific assumptions are prerequisites for the application of specific physics concepts (e.g., conservation of energy requires frictionless motion), contrary to expert problem solvers [83]. Further, Physics Olympiad participants far more frequently introduced quantitative aspects in their descriptions compared to the nonparticipants. This can, on one hand, be ascribed to the Physics Olympiad participants on average greater knowledge base as they recall formulas and dependencies between physical quantities [84]. On the other hand, mental imaginative power and flexibility are required to describe precise quantitative aspects such as manipulation of specific equations without explicitly doing so. Lastly, and somewhat unexpectedly, nonparticipants more frequently formulated a concrete solution or hypothesis regarding the solution of the problem. This may be explained by the finding of problem-solving research that expert problem solvers are generally more focused on the representation of the problem at hand while novice problem solvers often are more focused on reaching the goal state [48]. Additionally, nonparticipants generally gave shorter descriptions ($M = 2.2$ sentences) than Physics Olympiad participants ($M = 3.6$ sentences). Specifically, nonparticipants more often just formulated a hypothesis including a short justification or explanation only, i.e., they did not elaborate on physics concepts or quantitative aspects at all. To summarize, our results indicate that Physics Olympiad participants exhibit on average more expertlike problem-solving behavior while nonparticipants show on average more novicelike behavior. We admit that using the traditional expert-novice dichotomy to describe the two subgroups limits our conclusions as we do not consider stages of expertise in between as well as variance within the subgroups (e.g., experts among the nonparticipants). Future research might benefit from, for example, more holistic approaches (such as latent profile analysis; e.g., [85]) to identify different groups of students with similar distributions of themes within their problem-solving approaches. This would allow us to establish a more differentiated picture of students' problem solving and go beyond the traditional expert-novice dichotomy.

While overall the findings regarding students' problem-solving approaches and the differences between Physics Olympiad participants and nonparticipants were rather expectable based on the existing literature, the ML model trained in the pattern confirmation step of the CGT approach leads to an intriguing byproduct of the analysis: New textual descriptions of students' problem-solving approaches can now be automatically characterized based on the five themes identified in this study. If this turns out to work reliably on unseen data (which should definitely be checked), then this would allow us to automatically provide feedback based on the characteristics of the descriptions which was found to effectively support students' development of problem-solving strategies [86]. Moreover, potential for future research is given by investigating sequences of themes in students' problem-solving descriptions by means of sequence analyses [87]. This would allow us to understand to what extent specific sequences of themes are more predictive of successful problem solving. Specifically, this could further improve automatic feedback. For example, if a student's actual sequence of themes (based on what the student has written so far) is characteristic for unsuccessful problem solving, then the system could automatically generate feedback "on the fly" to maneuver the student back to a more promising sequence.

IV. GENERAL DISCUSSION

In this paper, we presented an applied example of computational grounded theory—a method for analyzing qualitative data that involves the application of artificial intelligence techniques to assist with the coding, categorization, and analysis of the data—to probe its potentials and challenges for physics education research. CGT conceptualizes a process of how researchers can efficiently and effectively analyze large amounts of qualitative data with the help of computational methods such as natural language processing and machine learning, which provides new research opportunities and allows asking new questions that depend on prohibitively large datasets for human analysis. Further, CGT promises to provide new insights as the usage of NLP and ML techniques during coding, categorization, and analysis may reveal things inaccessible to the human analyst. Finally, CGT promises to make qualitative research more rigorous by enhancing the reliability and reproducibility of the analysis. To what extent can CGT fulfill these promises?

CGT certainly allowed us to analyze a large amount of qualitative data, i.e., a total of 1127 sentences from 417 students, in a very efficient way. For someone familiar with the computational techniques, the pattern recognition and pattern confirmation steps are straightforward and can be easily conducted within a few hours. This time can be even further reduced if the structure of the data allows us to reuse existing analysis pipelines that can easily be made available for interested researchers by sharing the project's code through online repositories (see [70] for our project's code and data). However, a potential caveat here is processing time. Depending on the computational power available, computation on very large datasets or using complex ML techniques can require a lot of time or money for powerful hardware, which also taxes the environment [88]. For the pattern refinement step, the human analysts altogether needed approximately five hours. Thus, the total time required for the analysis is in the range of ten hours. Based on our experiences as qualitative researchers, we estimate that a traditional qualitative analysis of the data would easily have taken at least twice as long. In this way, using CGT was not only time effective, but, as all the computational analyses were carried out using open source tools, also very cost effective, even for a single analysis. This is somewhat in contrast to the conclusion of Nehm and Haertig [89] who still had to rely on commercial tools and thus suggested that computational techniques become cost effective only for repeated analyses on very large datasets (e.g., in the context of admission tests).

However, cost and time effectiveness are only of interest if the results are valid. The close alignment of the identified themes in the data with substantive theory suggests the principal validity of the approach. At the same time, this affirms that the applied computational techniques did not reveal new patterns in problem-solving processes that human analysts did not already find in the past. Does this mean that the promise of gaining new insights as NLP and ML techniques can reveal things inaccessible to the human analyst does not hold? This conclusion can hardly be drawn based on one applied example. Specifically, Nelson [12], Hope and Witmore [26], and Rosenberg and Krist [14] have demonstrated how computational pattern detection can reveal patterns that are likely too complex for the human analyst to identify. Nelson, the sociologist who proposed CGT, used CGT to shed new light on the different approaches to facilitating social change underlying the women's movements in New York City and Chicago between 1865 and 1975. This analysis was conducted on a text corpus which under usual circumstances would have been prohibitively large for a single individual to investigate. Hope and Witmore applied a computational approach to Shakespeare's Macbeth and found that the heterodox usage of "the" within Macbeth plays a major role in constructing the uncanny atmosphere of the play. Lastly, Rosenberg and Krist applied the CGT methodology to develop a refined rubric for assessing students' ideas about generality in model-based explanations. A challenge for future applications of CGT will certainly occur when CGT-based studies come to results that deviate further from or even contradict existing literature. Here, technical expertise with the involved computational methods and skilled analysts will both be required to provide strong validity arguments. Here, current and developing techniques of explanatory model analysis [29] and explainable AI [30] are promising tools to provide validity arguments for the computational methods.

When NLP and ML techniques are used for pattern detection, bias can be an issue [31,33]. Especially pretrained language models can forward biases in their embeddings, which would hamper further analyses [90]. Not coincidentally, these language models exhibit similar biases as humans [91]. Similar to Ha and Nehm [92], our bias analysis did not reveal bias with regards to gender as a result of the used computational techniques. This does not mean, however, that CGT is bias proof. Rather, it means that CGT provides tools to easily conduct bias analyses and that those should become a standard component in CGT applications [93,94].

Another kind of bias, i.e., bias in a nondiscriminatory meaning, is also often a concern in qualitative analyses. As Nelson [12] writes "It is difficult to get the same person to code the same article in the same way twice, let alone train an entirely new team to code a corpus in the same way as a previous team." The usage of computational tools in the steps of CGT partly address this issue. There are human decisions and actions involved when applying computational tools (see Table III) which must be made transparent by the analysts for the sake of reproducibility. Completely computational parts of the analysis can be considered reproducible as the analysts' decisions and action are mostly documented directly or indirectly in the respective

TABLE III. Overview of the computational aspects and corresponding human analyst decisions and actions within our applied CGT example.

	Computational procedure	Computational output(-s)	Involved human analyst decisions and actions
Pattern detection	(a) *Data preprocessing*		
	Sentence segmentation (using a pretrained language model)	uncleaned sentence corpus including single sentences and (rudimentary) sentence fragments	deciding on a sentence's minimum (character) length below which a sentence (fragment) is removed from the corpus based on scanning of the uncleaned sentence corpus and theoretical arguments
	Data cleaning	cleaned sentence corpus ready for further processing	⋯
	Computing some descriptive statistics	e.g., frequency distribution showing the number of sentences per text, average text length, …	⋯
	(b) *Generating sentence embeddings* (using a pretrained language model)	high-dimensional sentence embeddings	⋯
	(c) *Exploratory computational analysis*		
	Performing dimension reduction (UMAP) and clustering (HDBSCAN) multiple times	n clustering solutions; frequency distribution of the number of clusters in each clustering solution; mode M of this distribution	deciding on the number n of clustering solutions to be generated, selecting hyperparameter values for UMAP and HDBSCAN based on good practice examples and heuristics as oftentimes explicated by the authors of the algorithms or using an exploratory approach
	Computationally guided visual inspection of selected clustering solutions	two-dimensional embedding plots (including top three words per cluster) of m (randomly sampled) clustering solutions having M clusters each	deciding on the number m of clustering solutions to further inspect, identifying constituting clusters among these solutions by visually inspecting their embedding plots and taking the similarity of the three most relevant words per cluster into account
	Visualizing the optimal clustering solution	final clustering solution including two-dimensional embedding plot and condensed tree plot	determining a clustering solution among the n solutions that (best) includes all constituting clusters
Pattern refinement	(a) *Determining most relevant words* (based on TF-IDF scores)	table of most relevant words per cluster	deciding how many relevant words or representative sentences are retrieved, developing definitions of detected clusters based on the output tables (further deep-reading beyond table output is possible),
	(b) *Determining most representative sentences* (based on cluster centroids and distances)	table of most representative sentences per cluster	pattern refinement based on developed definitions, which includes: partial relabeling of sentences (particularly in the noise cluster), merging clusters based on substantive theory and buttressed in a data-based way by using the embedding and condensed tree plot, conclusions regarding the research question based on the (refined) detected patterns (only after *Step 3—Pattern confirmation* confirmed reliability of CGT process so far)

(Table continued)

TABLE III. *(Continued)*

	Computational procedure	Computational output(-s)	Involved human analyst decisions and actions
Pattern confirmation	(a) *K-fold cross validation* (involving training of *K* classifiers/RVMs)	reliability measures such as predictive accuracy (overall and for specified subgroups, e.g., by gender)	low overall reliability or differences in reliability for subgroups (indicating possible bias) should lead to revising the pattern refinement (and possibly pattern detection) step based on additional exploratory analyses and/or explainable AI techniques
	(b) *Training of final classifier*	trained RVM classifier that can be applied to unseen input data	deciding on whether the trained classifier can be used in practical applications based on inspecting the classifier's generalizability by applying it to unseen data

code. Whenever human judgments or interpretations are involved, especially in the pattern refinement step, reproducibility is limited to that of any other qualitative analysis. In addition, the pattern confirmation step provides a reproducible version of the coding in the form of a computational model which can be easily shared and applied to new data by other researchers.

Overall, we argue that the computational aspects of CGT and the resulting documentation of large parts of the analysis in the form of analysis scripts or computational notebooks provides new and exciting ways for sharing qualitative work in the research community. If analysts shared their data and analysis scripts, others can easily dive into the analysis. For example, researchers can then explore how sensitive the presented results are to changes in the preprocessing of the textual data, fostering open, reproducible, and productive research practices in this way [95].

At this point, we want to highlight some cornerstones for improving CGT as presented in our applied example for future applications. First, we used pretrained language models which were domain independent. Using language models that were trained on in-domain data could substantially increase performance of NLP methods and downstream tasks [96]. If available for the correct language and domain, researchers should employ such domain-specific pretrained language models (e.g., SciBERT which was trained on scientific textual data in the English language [97]). Second, we had assumed that a sentence in a textual description always corresponds to exactly one theme. However, it is typically not uncommon that two or more themes are embedded within a single sentence. A finer grain size of segments might mitigate this limitation. For example, more elaborated segmentation algorithms could be used to segment sentences further into clauses. Third, we mentioned that our employed language model could not always effectively handle formulas and equations. Specifically, our model could not extract any meaning from formulas and equations which presents a serious limitation as students' problem-solving descriptions typically include many formulas and equations that can be seen as representative for the relevant physics concepts. Advanced NLP technologies such as ChatGPT are even able to interpret formulas and equations within textual problem-solving descriptions which represents a huge advantage compared to our employed pretrained language model. If these technologies become more available (and data protection issues are resolved), we advise researchers for incorporating these models into the CGT framework in PER in order to far more effectively handle formulas and equations automatically. Fourth, to probe generalizability of the trained ML classifier in the third CGT step, one would have to test the classifier on an unseen dataset (one that was not used during model training or development) as is common practice in ML research. This should be included as a standard in the CGT framework if the researchers plan to use this classifier in practical applications beyond the research context.

Taken together, we argue that CGT presents physics education researchers a valuable tool to scale up qualitative analyses and perform data-driven discovery. We were able to analyze a large dataset very effectively, the results are valid without evidence of bias, and the resulting category system can be used in a reproducible manner as it is encoded in a trained ML model. Following Kubsch *et al.* [13], we argue that the reason for this is that CGT as a framework distributes the tasks in qualitative data analysis between human analysts and computational tools in a way in which both can effectively complement each other. Human analysts offer expert knowledge and hermeneutic skills for pattern refinement while computational tools offer processing power to facilitate pattern detection and confirmation.

V. CONCLUSION

In this paper, we have presented CGT as a novel methodology that makes use of computational tools to enhance traditional qualitative analysis. The applied example has demonstrated that CGT held many of its promises. Still, CGT is no magic bullet to replace all other qualitative approaches. Rather, we argue that CGT seems especially valuable in two areas: (i) discovery in prohibitively large and unstructured datasets [15,98] and (ii) scaling analyses to investigate the generalizability and variance of phenomena. While using CGT with ten 5 min interview transcripts is certainly possible, the benefits compared to traditional approaches may be limited. Especially, since researchers curious to apply CGT should be aware that without the skills to use the computational tools appropriately, one can also be misled tremendously [12]. In consequence, this calls for collaboration with respective experts and respective learning opportunities for physics education researchers, e.g., in the form of methods courses in Ph.D. programs or conference workshops. In this spirit, we hope that the present article can serve as a first step towards making CGT available as a powerful research method for the PER community.

ACKNOWLEDGMENTS

We thank Holger Maus for his support and expertise in interpreting the clustering solutions. This work was supported by the Leibniz Association, Germany under Grant No. K194/2015.

[1] D. M. Watts, Some alternative views of energy, Phys. Educ. **18**, 213 (1983).

[2] L. Ding and P. Zhang, Making of epistemologically sophisticated physics teachers: A cross-sequential study of epistemological progression from preservice to in-service teachers, Phys. Rev. Phys. Educ. Res. **12**, 020137 (2016).

[3] M. T. H. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. **5**, 121 (1981).

[4] H. B. Carlone and A. Johnson, Understanding the science experiences of successful women of color: Science identity as an analytic lens, J. Res. Sci. Teach. **44**, 1187 (2007).

[5] G. M. Quan, C. Turpen, and A. Elby, Analyzing identity trajectories within the physics community, Phys. Rev. Phys. Educ. Res. **18**, 020125 (2022).

[6] L. Ivanjek, L. Morris, T. Schubatzky, M. Hopf, J.-P. Burde, C. Haagen-Schützenhöfer, L. Dopatka, V. Spatz, and T. Wilhelm, Development of a two-tier instrument on simple electric circuits, Phys. Rev. Phys. Educ. Res. **17**, 020123 (2021).

[7] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*, 1st ed. (Routledge, London, 2017).

[8] P. Mayring, *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution* (Klagenfurt, Austria, 2014), https://www.ssoar.info/ssoar/handle/document/39517?locale-attribute=en.

[9] L. A. Barth-Cohen, S. K. Braden, T. G. Young, and S. Gailey, Reasoning with evidence while modeling: Successes at the middle school level, Phys. Rev. Phys. Educ. Res. **17**, 020106 (2021).

[10] A. A. diSessa, Knowledge in pieces, in *Constructivism in the Computer Age*, edited by G. Forman and P. B. Pufall, 1st ed. (Lawrence Erlbaum Associates, Hillsdale, 1988).

[11] R. Biernacki, *Reinventing Evidence in Social Inquiry: Decoding Facts and Variables*, 1st ed. (Springer, Palgrave Macmillan, New York, 2012), https://doi.org/10.1057/9781137007285.

[12] L. K. Nelson, Computational grounded theory: A *Methodological Framework*, Sociol. Methods Res. **49**, 3 (2020).

[13] M. Kubsch, C. Krist, and J. M. Rosenberg, Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research, J. Res. Sci. Teach. **60**, 423 (2023).

[14] J. M. Rosenberg and C. Krist, Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations, J. Sci. Educ. Technol. **30**, 255 (2021).

[15] C. Krist, E. Dyer, J. Rosenberg, C. Palaguachi, and E. Cox, Leveraging computationally generated descriptions of audio features to enrich qualitative examinations of sustained

uncertainty, in *Proceedings Of The International Conference Of The Learning Sciences* (to be published).

[16] K. Charmaz, *Constructing Grounded Theory* (Sage, Los Angeles, 2014).

[17] C. A. Bail, The cultural environment: Measuring culture with big data, Theory Soc. **43**, 465 (2014).

[18] M. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys. **46**, 323 (2005).

[19] P. Wulff, Network analysis of terms in the natural sciences insights from wikipedia through natural language processing and network analysis, Educ. Inf. Technol. (2023).

[20] D. M. Blei, Probabilistic topic models, Commun. ACM **55**, 77 (2012).

[21] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, Optimizing semantic coherence in topic models, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011), pp. 262–272.

[22] T. O. B. Odden, A. Marin, and M. D. Caballero, Thematic analysis of 18 years of physics education research conference proceedings using natural language processing, Phys. Rev. Phys. Educ. Res. **16**, 010142 (2020).

[23] N. H. Christianson, A. Sizemore Blevins, and D. S. Bassett, Architecture and evolution of semantic networks in mathematics texts, Proc. R. Soc. A **476**, 20190741 (2020).

[24] T. O. B. Odden, A. Marin, and J. L. Rudolph, How has science education changed over the last 100 years? An analysis using natural language processing, Sci. Educ. **105**, 653 (2021).

[25] E. P. S. Baumer, D. Mimno, S. Guha, E. Quan, and G. K. Gay, Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?, J. Assoc. Inf. Sci. Technol. **68**, 1397 (2017).

[26] J. Hope and M. Witmore, The language of Macbeth, in *Macbeth: The State of Play*, edited by A. Thompson (Bloomsbury (Arden), London, 2014), pp. 183–208.

[27] M. Rost, L. Barkhuus, H. Cramer, and B. Brown, Representation and communication: Challenges in interpreting large social media datasets, in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (ACM, San Antonio Texas USA, 2013), pp. 357–362.

[28] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, Pre-trained models for natural language processing: A survey, Sci. China Technol. Sci. **63**, 1872 (2021).

[29] P. Biecek and T. Burzykowski, *Explanatory model analysis: Explore, explain, and examine predictive models*, 1st ed. (CRC Press, Boca Raton, 2021).

[30] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, editors, *Explainable AI: Interpreting, Explaining And Visualizing Deep Learning* (Springer International Publishing, Cham, 2019), Vol. 11700.

[31] T. Cheuk, Can AI be racist? Color-evasiveness in the application of machine learning to science assessments, Sci. Educ. **105**, 825 (2021).

[32] K. Crawford, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, New Haven, Connecticut, 2021).

[33] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown Publishers, New York, 2017).

[34] M. I. Baig, L. Shuib, and E. Yadegaridehkordi, Big data in education: A state of the art, limitations, and future research directions, Int. J. Educ. Technol. High. Educ. **17**, 44 (2020).

[35] P. Mulvey and J. Pold, *Physics doctorates: Skills used and satisfaction with employment* (American Institute of Physics, New York, 2020).

[36] H. Jang, Identifying 21st century STEM competencies using workplace data, J. Sci. Educ. Technol. **25**, 284 (2016).

[37] R. F. Frey, C. J. Brame, A. Fink, and P. P. Lemons, Teaching discipline-based problem solving, CBE Life Sci. Educ. **21**, 1 (2022).

[38] L. Hsu, E. Brewe, T. M. Foster, and K. A. Harper, Resource letter RPS-1: Research in problem solving, Am. J. Phys. **72**, 1147 (2004).

[39] M. E. Martinez, What is problem solving?, Phi Delta Kappan **79**, 605 (1988), https://www.jstor.org/stable/20439287.

[40] G. Friege, *Wissen Und Problemlösen: Eine Empirische Untersuchung Des Wissenszentrierten Problemlösens Im Gebiet Der Elektizitätslehre Auf Der Grundlage Des Experten-Novizen-Vergleichs [Knowledge and problem solving: An empirical investigation of knowledge-centered problem solving in the field of electricity based on expert-novice comparison]* (Logos-Verlag, Berlin, Germany, 2001).

[41] E. R. Savelsbergh, M. G. M. Ferguson-Hessler, and T. de Jong, The importance of an enhanced problem representation: On the role of elaborations in physics problem solving, University of Twente Faculty of Educational Science and Technology, Department of Instructional Technology, 1997.

[42] G. S. Selçuk and S. Çalýskan, The effects of problem solving instruction on physics achievement, problem solving performance and strategy use, Lat. Am. J. Phys. Educ. **2**, 151 (2008), http://www.lajpe.org/sep08/01_Gamze_Sezgin.pdf.

[43] G. Polya, *How to Solve It—A New Aspect of Mathematical Method* (Princeton University Press, Princeton, Oxford, 1945).

[44] D. Fortus, The importance of learning to make assumptions, Sci. Educ. **93**, 86 (2009).

[45] P. T. Hardiman, R. Dufresne, and J. P. Mestre, The relation between problem categorization and problem solving among experts and novices, Memory Cogn. **17**, 627 (1989).

[46] J. H. P. Van Weeren, F. F. M. De Mul, M. J. Peters, H. Kramers-Pals, and H. J. Roossink, Teaching problem-solving in physics: A course in electromagnetism, Am. J. Phys. **50**, 725 (1982).

[47] G. Friege and G. Lind, Types and qualities of knowledge and their relations to problem solving in physics, Int. J. Sci. Math. Educ. **4**, 437 (2006).

[48] J. H. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, Science **208**, 1335 (1980).

[49] P. Reinhold, G. Lind, and G. Friege, Wissenszentriertes problemlösen in physik, Z. Für Did. Der Naturwissenschaften **5**, 41 (1999) [Knowledge-based problem-solving in physics].

[50] C. A. Ogilvie, Changes in students' problem-solving strategies in a course that includes context-rich, multifaceted problems, Phys. Rev. ST Phys. Educ. Res. **5**, 020102 (2009).

[51] J. L. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson, A. Mason, Q. X. Ryan, and J. Yang, Assessing student written problem solutions: A problem-solving rubric with application to introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 010130 (2016).

[52] S. Petersen and P. Wulff, The German Physics Olympiad—Identifying and Inspiring Talents, Eur. J. Phys. **38**, 034005 (2017).

[53] J. Walker, R. Resnick, and D. Halliday, *Fundamentals of Physics*, 10th ed. (Wiley, Hoboken, NJ, 2014).

[54] D. Meurers, *Natural Language Processing and Language Learning, in The Encyclopedia of Applied Linguistics*, edited by C. A. Chapelle, 1st ed. (Wiley, New York, 2021), pp. 1–15.

[55] D. Jurafsky, Probabilistic Modeling in Psycholinguistics. Linguistic Comprehension, and Production, in *Probabilistic Linguistics*, edited by J. Hay, R. Bod, and S. Jannedy (MIT Press, Cambridge, MA, 2003), pp. 39–95.

[56] L. N. Jescovitch, E. E. Scott, J. A. Cerchiara, J. Merrill, M. Urban-Lurain, J. H. Doherty, and K. C. Haudek, Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression, J. Sci. Educ. Technol. **30**, 150 (2021).

[57] D. E. Powers, "Wordiness": A selective review of its influence, and suggestions for investigating its relevance in tests requiring extended written responses, No. RM-04-08, ETS, 2005.

[58] G. van Rossum, *Python Reference Manual* (CWI, Amsterdam, Netherlands, 1995), https://ir.cwi.nl/pub/5008.

[59] A. Boyd, SpaCy: Industrial-Strength NLP (2022), https://github.com/explosion/spaCy.

[60] Y. Zhang, R. Jin, and Z.-H. Zhou, Understanding bag-of-words model: A statistical framework, Int. J. Mach. Learn. Cyber. **1**, 43 (2010).

[61] N. Reimers and I. Gurevych, Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks, arXiv:1908.10084v1.

[62] B. Chan, S. Schweter, and T. Möller, German's next language model, in *Proceedings of the 28th International Conference on Computational Linguistics* (International Committee on Computational Linguistics, Barcelona, Spain, 2020), pp. 6788–6796.

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805v2.

[64] I. Tenney, D. Das, and E. Pavlick, BERT rediscovers the classical NLP pipeline, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Florence, Italy, 2019).

[65] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

[66] S. L. Brunton and J. N. Kutz, Data-driven science and engineering: Machine learning, *Dynamical Systems, and Control*, 1st ed. (Cambridge University Press, Cambridge, England, 2019).

[67] S. Ayesha, M. K. Hanif, and R. Talib, Overview and comparative study of dimensionality reduction techniques for high dimensional data, Inform. Fusion **59**, 44 (2020).

[68] L. McInnes, J. Healy, and J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv:1802.03426v3.

[69] M. Allaoui, M. L. Kherfi, and A. Cheriet, Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study, in *Image and Signal Processing*, edited by A. El Moataz, D. Mammass, A. Mansouri, and F. Nouboud (Springer International Publishing, Cham, 2020), Vol. 12119, pp. 317–325, https://doi.org/10.1007/978-3-030-51935-3_34.

[70] P. Tschisgale, Computational grounded theory in physics education research, OSF, 10.17605/OSF.IO/D68CH (2023).

[71] P. Wulff, D. Buschhüter, A. Westphal, L. Mientus, A. Nowak, and A. Borowski, Bridging the gap between qualitative and quantitative assessment in science education research with machine learning—A case for pretrained language models-based clustering, J. Sci. Educ. Technol. **31**, 490 (2022).

[72] R. J. G. B. Campello, D. Moulavi, and J. Sander, Density-based clustering based on hierarchical density estimates, in *Advances in Knowledge Discovery and Data Mining*, edited by J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Vol. 7819 (Springer, Berlin, Heidelberg, 2013), pp. 160–172, https://doi.org/10.1007/978-3-642-37456-2_14.

[73] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, Hierarchical density estimates for data clustering, visualization, and outlier detection, ACM Trans. Knowl. Discov. Data **10**, 1 (2015).

[74] L. McInnes, J. Healy, and S. Astels, Hdbscan: Hierarchical density based clustering, J. Open Source Softw. **2**, 205 (2017).

[75] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.19.020123 for (a) further analyses on hyperparameter sensitivity, (b) two-dimensional embedding plots of 20 clustering solutions with ten clusters each, and (c) further performance measures of the relevance vector machine.

[76] S. Qaiser and R. Ali, Text mining: Use of TF-IDF to examine the relevance of words to documents, Int. J. Comput. Appl. Technol. **181**, 25 (2018).

[77] J. Ramos, Using TF-IDF to Determine Word Relevance in Document Queries, in *Proceedings of the First Instructional Conference on Machine Learning* (2003), Vol. 242, No. 1, pp. 29–48, http://www.liber.ufpe.br/tg/modules/tg/docs/ramos.pdf.

[78] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv:2203.05794v1.

[79] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand, and P. K. Soni, *Natural Language Processing (NLP) Based Text Summarization—A Survey* (IEEE, Coimbatore, India, 2021), pp. 1310–1317, https://doi.org/10.1109/ICICT50816.2021.9358703.

[80] R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, 2nd ed. (CRC Press, Boca Raton, 2020).

[81] A. Grimm, A. Steegh, J. Çolakoğlu, M. Kubsch, and K. Neumann, Positioning responsible learning analytics in the context of STEM identities of under-served students, Front. Educ. **7**, 1082748 (2023).

[82] A. Steegh, T. N. Höffler, M. M. Keller, and I. Parchmann, Gender differences in mathematics and science competitions: A systematic review, J. Res. Sci. Teach. **56**, 1431 (2019).

[83] F. Reif and S. Allen, Cognition for interpreting scientific concepts: A study of acceleration, Cognit. Instr. **9**, 1 (1992).

[84] J. H. Larkin, The Role of problem representation in physics, in *Mental Models*, edited by D. Gentner and A. L. Stevens, 1st ed. (Psychology Press, New York, 1983).

[85] D. Spurk, A. Hirschi, M. Wang, D. Valero, and S. Kauffeld, Latent profile analysis: A review and "how to" guide of its application within vocational behavior research, J. Vocat. Behav. **120**, 103445 (2020).

[86] W. J. Leonard, R. J. Dufresne, and J. P. Mestre, Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems, Am. J. Phys. **64**, 1495 (1996).

[87] M. Raab and E. Struffolino, *Sequence Analysis* (Sage, Los Angeles, 2023).

[88] J. Dodge, T. Prewitt, R. Tachet des Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, and W. Buchanan, Measuring the carbon intensity of AI in cloud instances, *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, Seoul Republic of Korea, 2022), pp. 1877–1894, https://doi.org/10.1145/3531146.3533234.

[89] R. H. Nehm and H. Haertig, Human vs. computer diagnosis of students' natural selection knowledge: Testing the efficacy of text analytic software, J. Sci. Educ. Technol. **21**, 56 (2012).

[90] L. Weidinger *et al.*, Ethical and social risks of harm from language models, arXiv:2112.04359v1.

[91] A. Caliskan, J. J. Bryson, and A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science **356**, 183 (2017).

[92] M. Ha and R. H. Nehm, The impact of misspelled words on automated computer scoring: A case study of scientific explanations, J. Sci. Educ. Technol. **25**, 358 (2016).

[93] T. Cerratto Pargman, C. McGrath, O. Viberg, K. Kitto, S. Knight, and R. Ferguson, Responsible learning analytics: Creating just, ethical, and caring LA systems, in *Companion Proceedings of the 11th International Conference on Learning Analytics & Knowledge* (Society for Learning Analytics Research, 2021), https://oro.open.ac.uk/75925/1/Responsible%20Learning%20Analytics.pdf.

[94] O. L. Liu, J. A. Rios, M. Heilman, L. Gerard, and M. C. Linn, Validation of automated scoring of science assessments, J. Res. Sci. Teach. **53**, 215 (2016).

[95] L. Figueiredo, C. Scherer, and J. S. Cabral, A simple kit to use computational notebooks for more openness, reproducibility, and productivity in research, PLoS Comput. Biol. **18**, e1010356 (2022).

[96] D. Banman, Natural language processing for the long tail, in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2006), pp. 120–128, https://www.semanticscholar.org/paper/Natural-Language-Processing-for-the-Long-Tail-Bamman/a2504ebc51f9a5a4ba73d6f1dc26961a8b8b27ef.

[97] I. Beltagy, K. Lo, and A. Cohan, SciBERT: A pretrained language model for scientific text, arXiv:1903.10676v3.

[98] K. Hall, What did we (not) say? Using a computaional grounded theory approach to map the evolution of "equity" in a 107 years of science education editorials (unpublished).