# Using text embeddings for deductive qualitative research at scale in physics education

Tor Ole B. Odden[1,†] Halvor Tyseng[2,*] Jonas Timmann Mjaaland[2,*]
Markus Fleten Kreutzer[2,*] and Anders Malthe-Sørenssen[1,2]

[1]*Center for Computing in Science Education, University of Oslo, NO-0316 Oslo, Norway*
[2]*Center for Interdisciplinary Education, University of Oslo, NO-0316 Oslo, Norway*

(Received 26 February 2024; accepted 13 November 2024; published 20 December 2024)

We propose a technique for performing deductive qualitative data analysis at scale on text-based data. Using a natural language processing technique known as text embeddings, we create vector-based representations of texts in a high-dimensional meaning space within which it is possible to quantify differences in meaning as vector distances. To apply the technique, we build off prior work that used topic modeling via latent Dirichlet allocation to thematically analyze 18 years of the Physics Education Research Conference Proceedings literature. We first extend this analysis through 2023. Next, we create embeddings of all texts and, using representative articles from the 10 topics found by the LDA analysis, define centroids in the meaning space. We calculate the distances between every article and centroid and use the inverted, scaled distances between these centroids and articles to create an alternate topic model. We benchmark this model against the LDA model results and show that this embeddings model recovers most of the trends from that analysis. Finally, to illustrate the versatility of the method, we define eight new topic centroids derived from a review of the physics education research literature by Docktor and Mestre and reanalyze the literature using these researcher-defined topics. Based on these analyses, we critically discuss the features, uses, and limitations of this method and argue that it holds promise for flexible deductive qualitative analysis of a wide variety of text-based data that avoids many of the drawbacks inherent to prior NLP methods.

DOI: 10.1103/PhysRevPhysEducRes.20.020151

## I. INTRODUCTION

Qualitative analysis is not easy. It is time-consuming and often difficult to replicate. In physics education research (PER), specifically, many qualitative studies aim to infer students' or instructors' understandings, framings, emotions, or social dynamics, and therefore, require fine-grained analysis of subtle details of dialogue, gesture, or speech. Categorizations can hinge on the use of specific phrases or words, the lack thereof, or even require researchers to go beyond surface content and context to focus on more gestalt aspects of the data.

Despite these challenges, qualitative analysis is often the best method we have for many of the research problems in a field as diverse, multifaceted, and complex as PER. Ours is a fairly new field, and as such is still in the process of identifying, exploring, and drawing in various methods and research questions [1]. It also addresses a sufficiently complex topic—human learning—that the methods used often look less like standard physics than biology or social sciences research, where qualitative methods are much more prevalent [2].

At the same time, there is plenty of room for methodological improvement. As theoretical paradigms and research questions become established, researchers often move from theoretical papers, small-N case studies, and existence proofs (e.g., [3,4]) to large-scale validation studies (e.g., [5,6]). But large qualitative studies often come at the cost of either significantly higher time investment or lower analytical grain size.

Over the past several years, the field of natural language processing (NLP) has made huge gains in the mathematical representation, manipulation, and analysis of text. These tools hold significant promise in supporting qualitative research on text-based data since computers are now able to reliably perform many of the types of content-based analyses that were previously only possible to do by a human researcher. Thus, these methods have the potential to help researchers scale up their qualitative analyses. In this article, we show a proof-of-concept of how one of these current, cutting-edge NLP methods can be used as an aid in performing deductive qualitative analysis at scale in PER.

*These authors contributed equally to this work.
†Contact author: t.o.b.odden@fys.uio.no

## II. NATURAL LANGUAGE PROCESSING AND QUALITATIVE RESEARCH IN PER

### A. Qualitative research in PER: Deductive and inductive approaches

Qualitative research is generally divided into two approaches: deductive and inductive. When using a deductive approach, researchers typically analyze a dataset using a preestablished categorization scheme: often a set of codes, categories, or a theoretical framework that defines the different categories used in the analysis. These codes may be mutually exclusive (i.e., each piece of data can only be given one code) or overlap; they may be applied as a binary, or using different levels of prevalence, confidence, or saturation. Once the data are coded, the codes may be used to organize or summarize the data [7] or form the basis for further quantitative analysis based on the prevalence and co-occurrence of codes in the data.

Inductive approaches, in contrast, allow researchers to make sense of data by deriving or discovering patterns, trends, or themes. Depending on the particular method, researchers may go into the inductive analysis with some ideas of what they are looking for or try to avoid bringing in any *a priori* ideas and let the data "speak for itself" (a variant known as grounded theory). They then review the data multiple times, building and refining their themes or categorizations with each pass through the dataset. Once the data are analyzed, the results can be used to describe theoretical or empirical phenomena, propose new frameworks (or variants of existing frameworks), or explain social dynamics or lived experiences.[1]

In PER, specifically, both deductive and inductive approaches are common [8]. Many of the foundational studies of student understanding of physics were based on interviews with students, which were by necessity analyzed qualitatively [1,9,10]. Even quantitative instruments to measure student attitudes or understandings, like the Force Concept Inventory [11], typically involve some amount of qualitative data analysis to establish which ideas or attitudes the instrument is to probe [12]. However, the actual process of qualitative data analysis often looks much the same today as it did 20 or even 40 years ago, aside from some digital advances in data storage and qualitative data analysis software.

Over the last years, the fields of artificial intelligence (AI) and natural language processing have seen huge advances with the development of large language models and generative AI. Much of this advancement is based on increasingly sophisticated ways of representing text as vectors, which allows the text to be mathematically manipulated using standard statistical and vector-mathematical methods [13–15]. Thus, we see a need to explore how these tools may help researchers qualitatively analyze their data.

### B. NLP analysis in PER—Supervised and unsupervised approaches

Machine learning methods, like those used in natural language processing and artificial intelligence, are generally divided into two types of methods: supervised and unsupervised. Supervised methods are trained on a dataset that has been analyzed or labeled ahead of time, and they aim to create a model which can predict patterns or analyze features in unseen data. In some cases, these data might simply be a collection of multiple statistics, like student demographics and time to graduation [16], while in others, it might involve human coders labeling images or pieces of text. Supervised methods are commonly used across many different fields and applications since they are straightforward to evaluate: given a labeled dataset, one can compare the predictions of the model on a subset of the data to the actual labels and quantify their level of agreement or disagreement. However, they can be expensive to train, as creating a suitable coded and labeled dataset often requires a significant investment of time, effort, and resources.

Unsupervised methods, in contrast, are trained on data where there is no *a priori* labeling or classification. For example, data might be grouped together using a clustering algorithm like $K$-means or a dimensional reduction algorithm like principal component analysis.[2] Unsupervised methods have the advantage that they do not require data to be labeled ahead of time in order to work; however, this has the disadvantage of making them much more difficult to interpret, evaluate, or benchmark.

When it comes to applying natural language processing techniques to qualitative analysis in PER, as well as some related work in science education research (SER), we can map out the space of prior work according to the axes of deductive and inductive and supervised and unsupervised to examine the different research paradigms and contextualize the present study. This mapping is shown in Fig. 1.

#### *1. Supervised NLP applied to deductive qualitative analysis*

This paradigm is defined by researchers creating models that learn and apply preexisting coding and classification schemes to textual datasets using standard supervised NLP methods, like logistic regression. According to a review by Zhai *et al.* of machine learning methods in science education assessment (usually a deductive task), most such studies use supervised methods [18].

Within PER, much of this work has focused on creating models that can learn to reproduce classifications made by

[1]For a thorough description of the qualitative research process in PER, we recommend Otero *et al.* [8].

[2]For a more thorough overview of machine learning methods applied to science education, we recommend Wulff [17].

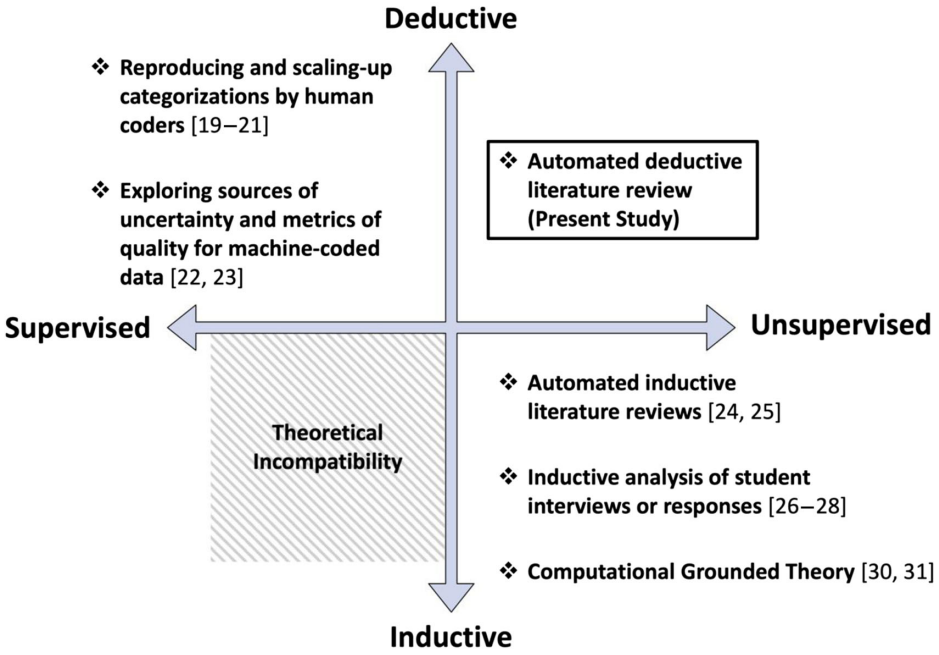


FIG. 1. Space of the literature in physics education research related to deductive/inductive qualitative analysis and supervised/unsupervised natural language processing.

human coders, in the hopes that such models could be used as a tool for rapid analysis of arbitrarily large future datasets. For example, Wilson *et al.* [19] created a logistic regression model that categorized student reasoning on the Physics Measurement Questionnaire. It was trained using a human-coded dataset of approximately 2450 responses, categorized using a coding scheme focused on different types of reasoning about measurement. The model produced similar levels of interrater reliability as that between two human coders.

Some researchers have also explored how much data are necessary to create robust supervised deductive models. For example, Wulff *et al.* [20] created a multinomial logistic regression model that classified written reflections by German preservice physics teachers, which had again been coded by human researchers ahead of time. Despite being trained on a relatively small amount of data (93 reflections), the model was able to achieve acceptable agreement with human coders on at least some of the code categories. Wulff *et al.* [21] performed a similar study with a larger dataset ($N = 270$ human-coded reflections from German preservice physics teachers) using a German pretrained large language model that was fine-tuned to the specific task using standard supervised machine learning methods, like hold-out and $k$-fold cross validation. When comparing the performance of this model with other deep learning architectures, they found that the fine-tuned classifier outperformed other deep-learning architectures.

Still other researchers are exploring the sources of uncertainty and metrics of quality in these types of models. Fussell *et al.* [22,23] trained two supervised machine learning models (a neural network and a logistic regression model, respectively) to apply an existing coding scheme to student responses on an open-ended survey about the trustworthiness of experimental results. Based on these trials, they propose a framework that describes the conditions and criteria necessary to consider machine coding trustworthy and reliable.

### 2. Supervised NLP applied to inductive qualitative analysis

This paradigm is defined by researchers using supervised NLP methods to create models that inductively learn or extract new themes from datasets. So far as we know, there is no research that directly fits this paradigm, as there is a fundamental incompatibility between the epistemological commitments of inductive qualitative analysis and supervised machine learning. That is, supervised machine learning requires one to approach the analysis *with* defined categories and some amount of prior analysis, while inductive qualitative analysis requires one to approach the data *without* a defined set of categories or prior analysis.

### 3. Unsupervised NLP applied to inductive qualitative analysis

This paradigm is defined by researchers using unsupervised NLP methods to extract themes or create new categories from the data that were not known *a priori*. A number of studies from PER and SER fit this paradigm. For example, Odden *et al.* [24]used an unsupervised topic modeling method called latent Dirichlet allocation (LDA) to perform a standard inductive qualitative research task: a

literature review. The literature base was 18 years of the *Physics Education Research Conference (PERC) Proceedings* (approximately 1300 articles). The analysis revealed ten distinct and recurring research topics and quantitatively mapped out their rise and fall over time. Odden *et al.* [25] used LDA to perform a similar analysis on approximately a century's worth of literature from the journal *Science Education* (approximately 5570 articles), again mapping large-scale shifts in that literature. The analysis showed that the journal had undergone a shift from science content-based articles, to studies focused primarily on teaching, and then to studies focused on student learning. It also showed that many of the advances in the field happened when science education borrowed theoretical frameworks or paradigms from other related fields like psychology and sociology.

Other researchers have used unsupervised methods for inductive analysis of student-level data. Clustering methods are especially popular in this area. For example, Sherin [26] used clustering algorithms to analyze segments of an interview transcript on the topic of how the Earth's orbit affects the seasons, finding that the clusters were thematically interpretable and suggestive that such methods could be used to enhance inductive qualitative analysis. Wulff *et al.* [27] used a clustering approach to categorize physics preservice teachers' written reflections ($N = 86$) of a teaching situation. Some amount of prior data analysis was performed using the trained classifier from Wulff *et al.* [21], which extracted relevant sentences according to a predefined coding scheme. These sentences were then grouped into topics using an unsupervised clustering algorithm, HDB-SCAN. The resulting clusters were found to be interpretable and allowed the researchers to organize and visualize different themes in the responses. Geiger *et al.* [28] used LDA to identify specific student ideas about circuits, based on a set of approximately 500 responses to a conceptual question about bulb brightness in a circuit with one vs two batteries. The resulting set of five ideas was quite interpretable, focusing on concepts like analogies to water flowing through circuits, different effects of electric potential, and quantitative relationships in Ohm's law.

Some researchers have explicitly positioned their analyses as a specific variety of qualitative analysis, such as grounded theory, and used that choice to guide the way they combine NLP and human analysis methods. For example, Rosenberg and Krist analyzed [30] ($N = 173$) responses from a seventh-grade chemistry assessment using a computational grounded theory approach, based on Nelson [29], in order to understand the students' epistemic ideas about generalizability. In this analysis, the authors first (inductively) produced an initial construct map that summarized their hypotheses about the data. Responses were then clustered using an unsupervised algorithm, and then these clusters were reviewed and compared to the construct map to produce a new coding scheme, which was again applied to the data. Finally, the authors used this coding scheme to train several supervised algorithms, which were found to have substantial agreement with human codes. Tschisgale *et al.* [31] also used a computational grounded theory approach in order to analyze student solutions to a German Physics Olympiad problem, comparing them to a control group of students who solved the same problem but did not participate in the Physics Olympiad. Their approach used a dimensional reduction technique called UMAP, followed by the unsupervised clustering technique HDB-SCAN to categorize 1127 sentences from the dataset. The researchers then qualitatively interpreted the resulting clusters and grouped them into themes. This analysis revealed that Olympiad participants showed more expert-like problem-solving behavior (referring more to assumptions and idealizations, conceptual aspects, and quantitative aspects of problem solutions) compared to nonparticipants.

### *4. Unsupervised NLP applied to deductive qualitative analysis*

This paradigm is defined by researchers analyzing a dataset by applying preexisting categories to it using unsupervised NLP methods. Although there might appear to be another theoretical incompatibility between these two approaches, this is only the case when one is using completely unsupervised methods (i.e., those without any input from a researcher) to perform deductive analysis. In practice, most unsupervised methods require some amount of researcher input, both in initialization and in tuning of model hyperparameters; for example, topic modeling methods like LDA allow researchers to specify the number of topics and the general "mixedness" of topics, while clustering methods allow researchers to specify the number of clusters and choose (if they wish) the initial centroid of each cluster. Thus, certain unsupervised NLP methods could, in principle, be used for deductive qualitative analysis.

To our knowledge, this paradigm has yet to be explored in the PER and SER literature. This represents a significant gap, as such a paradigm could provide significant advantages in flexibility and ease of use, combining the resource efficiency of unsupervised methods (which do not require humans to label large amounts of data ahead of time) with the many applications of deductive qualitative analysis.

In this article, we show a proof of concept of this kind of analysis. To do so, we have used a cutting-edge NLP technique called text embeddings, which allows one to turn texts into high-dimensional vectors and then perform vector operations on them, like calculating distance and similarity measures.

In what follows, we describe the theory behind text embeddings and contrast them with the features of a previously used technique, latent Dirichlet allocation. We then use text embeddings to perform an inductive review of PERC proceedings articles, benchmarking the analysis

against a prior analysis of the same dataset [24] that we have extended up to the most recent published year as of writing (2023). This provides a benchmark for the technique and a proof of concept that it can be deductively applied to reproduce prior results. It also allows us to evaluate its affordances and drawbacks. We then extend the analysis by performing a second literature review of PERC proceedings articles using researcher-defined topics, in order to show the versatility of the method.

### C. NLP theory: LDA and embeddings

Both LDA and text embeddings rely on the fundamental idea of representing text as vectors. This representation allows researchers to quantitatively explore different aspects of the text—such as word co-occurrence, meaning, and sentiment—using standard tools for vector manipulation, such as addition, subtraction, projection, and decomposition. However, there are different ways to construct these kind of representations, which hinge on the size and meaning of the vector space into which one projects the text.

Latent Dirichlet allocation (LDA) [32,33] is based on a bag of words (BoW) model. This approach, which has a long history in NLP research [34], represents each text in the dataset as the count of the unique words in that document. Thus, when creating a bag of words out of a set of texts, each document will be represented as a vector, with a dimensionality equal to the total number of unique words across all the texts. Each entry in a document's vector corresponds to a particular word, and the entry's value is the number of times that word occurs in that document.[3]

From a vector perspective, a BoW model thus treats each word as an independent dimension in a high-dimensional vector space encompassing all the words in all the documents (in practice, often many thousands of dimensions). A document is then represented by a specific point (or vector) in this space. Depending on the number, length, and heterogeneity of documents, such vector spaces may be sparsely populated. But, with enough documents and enough words, statistical regularities can emerge—that is, certain documents may use similar words, putting them at similar "locations" in the high-dimensional "word space." Topic modeling techniques like LDA allow one to use these regularities to extract latent *topics*, in the form of sets of words that frequently occur together [32,33]. Each document can then be scored based on the amount (prevalence) of these topics in its text, which, in turn, allows one to look for trends in topic prevalence across the dataset. For example, if one has a large number of documents over a significant time period, one can evaluate how the prevalence of different topics has varied as a function of time [24,25,35].

However, both LDA and the underlying bag of words model have some limitations and implementation challenges. For starters, such analyses often require extensive data preprocessing and filtering, since words that do not help to distinguish semantic meaning (commonly known as *stop words*, like "is," "and," "but," etc.) must usually be filtered out, along with words that are over-represented in the dataset [24]. Because LDA is an unsupervised algorithm, it can also be a challenge to interpret identified topics, and even when topics are interpretable they may not necessarily reflect meaningful trends from a researcher's perspective. For example, in Odden *et al.*'s analysis of PERC proceedings, one of the identified topics focused primarily on student difficulties with quantum mechanics. This topic likely emerged because these types of studies use a particular vocabulary that is distinct from that in the most other PERC Proceedings articles; however, from a researcher's perspective, it would have been more helpful to have a topic that focused, for example, on student conceptual difficulties across multiple physics topics. Finally, topic modeling techniques are often unstable, in that models are randomly initialized and results may be sensitive to these initialization states [24].

Embeddings are a more sophisticated method for representing words and text, which do not treat words as independent dimensions. Instead, the words themselves are transformed, using a large language model, into high-dimensional vectors (usually several hundred dimensions), which live in a semantic space that encodes their relative meanings. This allows words to be manipulated using classic vector-handling techniques: for example, one can add and subtract word vectors, such as taking the word vector for "king," subtracting "man," and adding "woman" to construct a vector that lies closest in the semantic space to the word "queen" [13]. One can also use dot products to evaluate their similarity (often known as a cosine similarity metric) and find the distance between two or more vectors using various distance metrics (like Euclidean distance or Manhattan distance).

Many of the recent advances in large language models and AI chatbots are based on representing words and texts using embeddings. In 2017, the foundations of these breakthroughs were laid when researchers developed a model that allowed one to encode relationships between word vectors, using a so-called *attention mechanism* [15]. For example, in a sentence like "the physics teacher wrote on the whiteboard," the word "physics" modifies the meaning of the word "teacher," and the word "wrote" is related to both "teacher" and "whiteboard." The development of the attention mechanism allowed language models to represent these types of relationships. Language models that incorporated these kinds of relationships were thereby

[3]For example, the zeroth entry might correspond to the word "physics," and the first entry might correspond to the word "education;" a document that uses the word "physics" 10 times and "education" 12 times would have 10 as the first entry in its vector, 12 as the second, and so forth.

able to create embeddings from longer texts, up to hundreds of words in length, by "pooling" or compressing the meanings of the words into a single vectorized representation. One of the first such models was the bidirectional encoder representations from transformers (BERT) model [36]. Over time, researchers created variations on these models, eventually developing the ability to fine-tune text vectors so that sentences (or other chunks) with overall similar meanings would be located closer to one another in the embeddings space; one early such model was sentence-BERT or SBERT [37].

In practice, current-generation text embeddings can encode not just the content or meaning of a text but also potentially deeper features like sentiment and writing style [38] and often use vectors of several hundred dimensions (commonly 512, 768, or 1536). By representing increasingly longer texts as vectors, entire documents could be compared using the vector-handling techniques described above—for example, one can compare the similarity of documents based on their relative distances in a high-dimensional space.

Embeddings, therefore, offer an alternative approach for comparing and extracting latent features across documents, where these vectorized representations of documents can encode more information than pure counts of word occurrence. This opens the possibility of doing analyses similar to those that can be done with a BoW model, like topic modeling.

### D. Research questions

In this study, we investigate this possibility. Rather than using embeddings to do inductive qualitative analysis, where we let an algorithm group the texts together according to some metric (as with LDA or clustering), we instead explore how one could use them deductively, choosing a set of archetypal texts from the data, aggregating them, and then using the relative distance between particular texts and these aggregated examples to classify the remainder of the data in an unsupervised way. Our research questions are, therefore, as follows:

1. How can embeddings be used to deductively reproduce prior inductive qualitative research on scientific literature like the PERC proceedings?
2. How can embeddings be used to perform deductive analyses of literature using researcher-defined categories?
3. What trends do embeddings-based analyses reveal in the development of the literature as a function of time?

## III. METHODS

### A. Dataset and update to prior LDA analysis

We based the present study on a prior dataset and study performed by *Odden et al.* [24], which analyzed PERC proceedings published from 2001–2018 using latent Dirichlet allocation. That analysis found the following ten recurring research topics:

(1) *Representations*: This topic focuses on research on student understanding of and difficulties with representations and the physics content usually used to teach representational fluency, like electricity and magnetism.
(2) *Problem solving*: This topic focuses on research on problem solving in physics.
(3) *Laboratory instruction*: This topic focuses on research on physics laboratory instruction.
(4) *Assessment (concepts)*: This topic focuses on the use of quantitative methods to measure student conceptual understanding, such as concept inventories.
(5) *K-12*: This topic focuses on research on precollege physics education, such as teacher professional development and pedagogy.
(6) *Quantum difficulties*: This topic focuses on student understanding of and difficulties with quantum mechanics, although the topic also shows up in articles related to student difficulties in other subjects like E&M.
(7) *Qualitative theory*: This topic focuses on theoretical case studies of physics student cognition, usually qualitative in nature, with a specific focus on student discourse and explanation.
(8) *RBIS*: This topic focuses on studies related to research-based instructional strategies, such as tutorials and clickers, in lectures and recitation sections.
(9) *Assessment (demographics)*: This topic focuses on quantitative research that unpacks the effects of different factors in student learning, such as gender and ethnicity.
(10) *Community and identity*: This topic focuses on research that uses more sociocultural perspectives, such as communities of practice, institutional change, and physics identity.

The analysis also quantified the prevalence of these topics in the data corpus of approximately 1300 PERC proceedings articles published from 2001 to 2018, by scoring each article based on the percent of each of these 10 topics in the article's text. Thus, one article might, for example, include 50% RBIS, 20% laboratory instruction, and 30% problem-solving topics, while another article might contain 80% community and identity and 20% assessment (demographics).

In line with our first two research questions, our first goal was to explore whether embeddings could be used to deductively reproduce this prior NLP-based inductive literature review of PERC proceedings. To explore this question, the remaining PERC literature published 2019–2023 was downloaded, and the text was extracted using the same text extraction tools as in the 2020 analysis (the

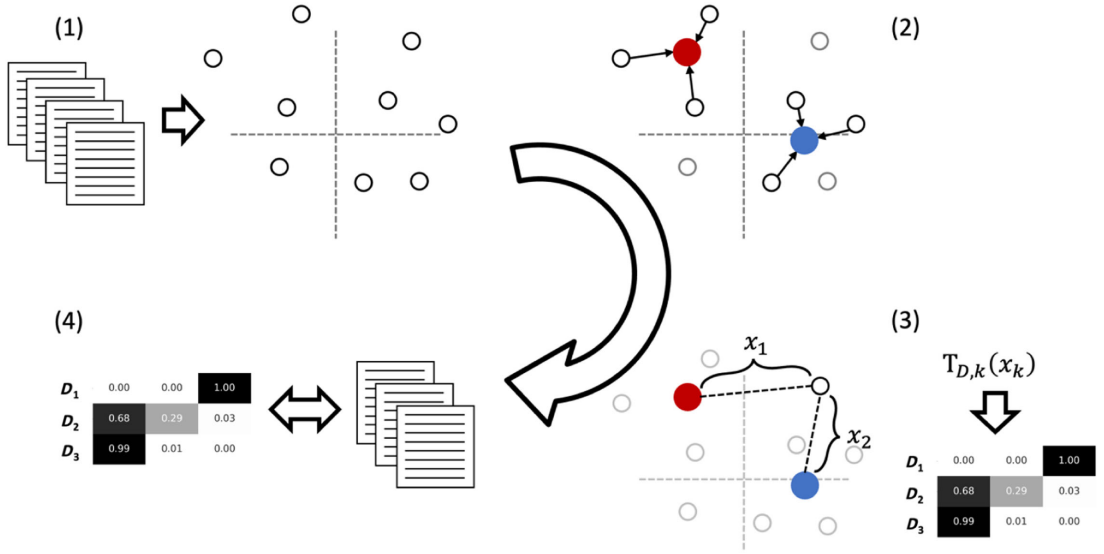


FIG. 2. Flowchart diagram of a method for performing deductive qualitative analysis using embeddings: (1) Creating embeddings vectors from texts; (2) defining topic centroids based on a sample of texts featuring that topic; (3) calculating distances from texts to topic centroids and transforming these into topic scores using a transformation function; and (4) evaluating results of the analysis based on face validity of topic scores and general trends.

Python library pdfminer.six). This resulted in a data corpus of 1745 articles. Next, an identical data cleaning procedure to that from Odden *et al.* was used on the 2019–2023 literature to produce an up-to-date dataset for LDA analysis, including steps of removing references and acknowledgments; removing numbers, symbols, and punctuation; lowercasing all words; connecting broken words; removing stop words; tokenizing; and lemmatizing; creating bigrams; and filtering words based on frequency. Using the LDA model created by Odden *et al.*, we then scored the 2019–2023 articles based on these 10 topics (using built-in functionality from the Gensim implementation of LDA) and added these topic scores to the dataset of LDA scores on articles published in 2001–2018 from Odden *et al.*'s original analysis. We present the results of this analysis in Sec. IV A; however, the primary purpose of this analysis was to provide a set of scores to benchmark our deductive analysis using embeddings.

## B. Description, implementation, and benchmarking of method

The method for this study consisted of four primary steps: (1) create embeddings vectors out of texts; (2) define topic centroids based on a selected sample of texts that exemplify desired features; (3) calculate and transform distances to produce topic scores, using scaling parameters to determine "mixedness" of the model; and (4) evaluate results of the analysis based on face validity of topic scores and general trends. We visualize these steps in Fig. 2 and elaborate on each below:

### *1. Creating embeddings vectors out of texts*

To begin, we created embeddings out of the raw text of each article, using an open-source large-language model called *Jina Embeddings 2* [39]. This model was chosen because it is one of the few open-source models, which allows one to create embeddings out of texts of up to 8192 tokens (individual words, numbers, etc.) in length. It can be run locally using an implementation from the HuggingFace repository.[4] The large language model took in the text of each article and produced a 512-dimensional vector with vector entries varying between −1 and 1 which represented the meaning of the text in the 512-dimensional semantic space.

We note that here, in contrast to the LDA analysis, we did not perform any of the preprocessing steps described above, such as removing numbers and punction or lemmatization, prior to embedding these texts. Texts were thus embedded in their raw form. In other words, this embeddings analysis technique required minimal data cleaning and preprocessing.

### *2. Defining topic centroids based on a selected sample of texts that exemplify desired features*

Next, we defined specific locations in the semantic space that corresponded to particular topics. To do so, for each of the 10 topics from the LDA analysis, we selected the 15 most representative articles (that is, the articles with the

[4]This avoids issues with other proprietary embeddings models, such as those from OpenAI and Google, which require one to upload articles to their systems using an API.

highest percentage score of that topic, often 90% or more) and took the mean of their embeddings vectors. This produced ten mean embeddings, one for each topic, which we refer to as the *topic centroids*. Since these topics were meant to capture similar meaning to the LDA topics, we used the same topic names and descriptors.

### 3. Calculating and transforming vector differences to produce topic scores

Next, we aimed to classify each article based on the relative distance between it and the various topic centroids in the semantic space. To do so, we used a similarity-based approach [40,41], similar in some ways to the technique known as semantic search [37], in which we first computed the distance between every individual article in the corpus and each topic centroid. For the purposes of exploration, we computed these distances multiple times using several distance metrics: Euclidean distance, cosine distance (1 − dot product), and Cityblock (Manhattan distance).

Due to the high dimensionality of the space, these distances were typically very similar (for example, dot product scores of 0.91–0.96 or equivalent cosine distances of 0.04–0.09), which yielded small-but-meaningful differences in distance from every paper to each topic centroid. So, to transform these distances into topic scores we developed a novel technique in which we applied a transformation function to the distances in order to invert them (since smaller distances should produce larger topic scores), scale them up, and normalize the scaled distances for each document. This produced a set of topic scores that summed to 1, with each individual score ranging from 0 to 1. We tried out two different transformation functions:

$$T(x) = e^{-\alpha x}, \tag{1}$$

$$T(x) = x^{-\alpha}. \tag{2}$$

Here, $\alpha$ is a hyperparameter, which in practical terms controls the degree of homogeneity and heterogeneity of the resulting topic distributions, similar to the $\alpha$ hyperparameter in LDA [32]. A low $\alpha$ produces scores that are approximately equal for each topic, while a high alpha will tend to push the model to give high scores in a single topic.

Embeddings model results were then benchmarked and evaluated by computing and summing the Jensen-Shannon divergence (JS divergence) between the embeddings topic scores and the LDA topic scores across the entire data corpus. JS divergence is a measure often used in data science applications to quantify the differences between two distributions, based on methods from the field of information theory. A value of 0 indicates that the two distributions are identical, and increasing amounts of difference are represented with increasingly higher divergence values. In our case, the distributions in question were the percent topic scores by LDA compared to the same topic scores by the embeddings model for each paper (illustrated in Figs. 6 and 7); we thus computed the JS divergence on each article and summed the results to get an aggregate divergence for the complete data corpus.

After running approximately 500 models to evaluate all combinations of the two transformation functions and three distance metrics over a range of $\alpha$ values (0–200), we found that the Jensen-Shannon divergence was minimized when we used the normalized exponential function [Eq. (1)] with approximately equal minimum values for cosine distance ($\alpha = 94$) and Euclidean distance ($\alpha = 34$), as shown in Fig. 3. We, therefore, chose to use the cosine distance metric, since it often performs better in high-dimensional datasets [42] and is recommended by OpenAI for use in their embeddings models [43].

For a given document $D$ and topic $k \in \{1, \ldots, K\}$, we, therefore, define the normalized topic score from the embeddings model, $T_{D,k}$, as

$$T_{D,k}(x_k) = \frac{1}{\sum_{k=1}^{K} e^{-\alpha x_k}} e^{-\alpha x_k}, \tag{3}$$

where $x_k$ is the cosine distance between the document and the $k$th topic centroid, calculated by taking 1 − dot product between the vectors; and $\alpha$ is a hyperparameter controlling the relative "mixedness" of the topics. This score can take a value of 0–1 (corresponding to the percentage of the $k$th topic in document $D$), and the scores of all topics $k = 1, \ldots, K$ for each document sum to 1. For reference, this divergence was also compared to that produced by uniform data and the mean of 2000 runs of random data. The minimum JS divergence was found at $\alpha = 94$, as shown in Fig. 3.

### 4. Evaluating results of the analysis based on model comparisons, face validity of topic scores, and general trends

Figure 3 shows that, in addition to the chosen model (exponential scaling/cosine distance metric/$\alpha = 94$), there were two other models that had similar minimum JS divergences with the LDA topic scores: exponential scaling with Cityblock distance metric and $\alpha = 1.9$; and exponential scaling with Euclidean distance metric and $\alpha = 34$. Exploratory modeling showed that these models differed most obviously in their primary topic scores, that is to say, the topic to which the model assigned the highest score for each paper. To evaluate our results' sensitivity to model choice, we therefore calculated the mean variation in the primary topic scores given to each paper, $T_i$, for each combination of these three models:

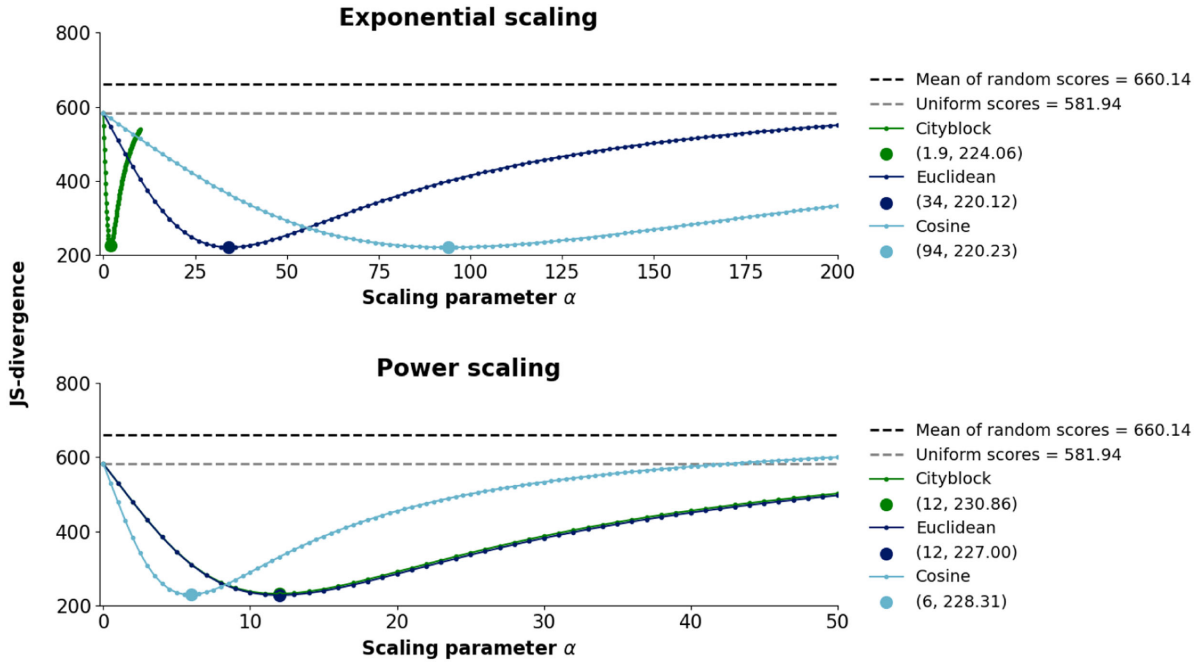


FIG. 3. Jensen-Shannon divergence between LDA topics and LDA-derived embeddings topics, using exponential and power scaling, as a function of scaling parameter $\alpha$. Distances were calculated using three metrics: Cityblock (Manhattan distance), Euclidean, and cosine distance (1 − dot product). JS divergence was also calculated with uniform topic distributions and random data (mean of 2000 runs) for reference. Minimum divergence is found with cosine distance metric, using exponential scaling function, at $\alpha = 94$.

$$\langle |\Delta T_{\text{models}}| \rangle = \frac{1}{3n} \left( \sum_{i=1}^{n} |T_{i,\text{Euclidean}} - T_{i,\text{Cityblock}}| + \sum_{i=1}^{n} |T_{i,\text{Cosine}} - T_{i,\text{Cityblock}}| + \sum_{i=1}^{n} |T_{i,\text{Cosine}} - T_{i,\text{Euclidean}}| \right), \tag{4}$$

where $\langle |\Delta T_{\text{models}}| \rangle$ is the mean variation in primary topic scores across the three compared models; $n$ is the total number of documents; $T_{i,\text{Euclidean}}$ is the primary topic for document $i$ calculated using the *Euclidean* distance metric; and the two other sets of primary topic scores are denoted by $T_{i,\text{Cityblock}}$ and $T_{i,\text{Cosine}}$.

This calculation produced a mean variation of $\langle |\Delta T_{\text{models}}| \rangle = 0.029$. The mean primary topic score across all three models was $\langle T_{\text{models}} \rangle = 0.47$, so dividing this variation by the mean value indicates that, on average, these models differed in their primary classifications by about 6%.

We also evaluated the sensitivity of the results to variations in $\alpha$, by computing $\langle |\Delta T_\alpha| \rangle$, the mean variation in the primary topic score as a function of alpha. To do this, we used the scaling function and distance metrics that provided the lowest JS divergence (exponential scaling and cosine distance), where this minimum was found at $\alpha' = 94$ (we used $\alpha'$ to denote that this was our chosen value of alpha for the analysis). Next, we computed topic scores with $\alpha$ values at approximately ±10% that of the $\alpha'$, which is to say $\alpha = 84$ and $\alpha = 104$. We then computed the mean variation in primary topic score across this range:

$$\langle |\Delta T_\alpha| \rangle = \frac{1}{n} \left( \sum_{i=1}^{n} |T_{i,\alpha'+10\%} - T_{i,\alpha'-10\%}| \right). \tag{5}$$

This resulted in a mean variation of $\langle |\Delta T_\alpha| \rangle = 0.07$ for a range of $\Delta\alpha = 20$, giving a sensitivity in primary topic score per unit $\alpha$ of $\langle |\Delta T_\alpha| \rangle / \Delta\alpha = 0.0035$. The mean primary topic score at $\alpha' = 94$ is $\langle T_{\alpha'} \rangle = 0.47$, so by dividing this sensitivity by the mean score we can interpret this sensitivity as a 0.7% change in primary topic score per unit $\alpha$ around that value.

Further, to unpack how the embeddings and LDA-derived models differed in their scores of specific topics, we calculated the correlation between LDA and embeddings model scores for each topic using Kendall's Tau (computed using the Scipy stats package). Kendall's Tau is an appropriate metric for data that do not pass the test of normality and which includes many ties when data are ranked, both of which were the case for this dataset. This analysis can be found in Supplemental Material of the article [44].

Finally, the first author checked the face validity of the model by comparing the scores of a representative set of individual articles by LDA and the embeddings model.

This embeddings model was then used to analyze the topic prevalence in PERC proceedings as a function of time, by aggregating and normalizing the percent of each topic from all articles in each year. Results were qualitatively compared with those from the updated LDA analysis, and differences were unpacked in light of the different textual features incorporated by the two methods. Results were also qualitatively interpreted to address research question 3.

### C. Extension of the analysis: Embeddings-based literature review of PERC proceedings based on Docktor and Mestre

Having used the LDA analysis as a benchmark to explore the use of embeddings for qualitative analysis, we next set out to perform a new analysis, using fully researcher-defined topics based on the PER literature review by Docktor and Mestre [45]. To do so, we first redownloaded the entire dataset and extracted the text using a more up-to-date text extraction library (the Python library PyMuPDF), which provided better quality texts sensitive to the two-column format of most PERC proceedings articles. This allowed us to retain several articles that had been removed from the prior analysis due to text extraction issues, bringing to complete corpus to 1761 articles.

Next, we repeated all of the methodological steps described above, except for benchmarking of results against prior analyses. That is, we first created new embeddings vectors of all 1761 articles using the same *Jina Embeddings 2* large language model described above [39]. We then defined a new set of topic centroids by operationalizing the categories and subcategories proposed by Docktor and Mestre in their review and synthesis of physics education research:

(1) *Conceptual understanding*: Including research on (a) naïve theories or misconceptions; (b) knowledge in pieces or conceptual resources; and (c) ontological categories or metaphors.
(2) *Problem solving*: Including research on (a) expert-novice differences; (b) worked examples; (c) representations; (d) mathematics in problem solving; and (e) instructional strategies for problem solving.
(3) *Curriculum and instruction*: Including research on (a) lecture-based methods; (b) recitations and discussion-based methods; (c) laboratory instruction; (d) structural changes to classroom environments; and (e) general instructional strategies and materials.
(4) *Assessment*: Including research on (a) development of concept inventories; (b) comparing scores across multiple measures; (c) comparing scores across multiple populations and demographics; (d) course exams and homework; and (e) rubrics for assessment.[5]
(5) *Cognitive psychology*: Including research on (a) knowledge and memory; (b) attention; (c) reasoning and problem solving; and (d) learning and transfer.
(6) *Attitudes and beliefs*: Including research on (a) student attitudes and beliefs; (b) faculty attitudes and beliefs; and (c) teaching assistant and learning assistant attitudes and beliefs.[6]

On top of these categories, we added two additional categories encompassing literature not included in that review, either due to explicit exclusion criteria (7) or because it had not yet emerged as a major theme in the literature base when that article was published (8):

(7) *Precollege physics education*: Including research on (a) teaching tools; (b) teacher preparation; (c) teaching strategies; and (d) student experiences.
(8) *Identity and equity*: Including research on (a) *physics identity;* (b) *equity;* (c) *race;* and (d) *gender.*

To define new topic centroids, the first author selected four representative articles for each category's subtopic from a list of all articles in the data corpus, aiming for a variety of authors, years, and study topics in each subcluster. The complete list of chosen articles can be found in Supplemental Material [44]. Articles were chosen based on whether they referenced some aspect of the subtopic in their title: for example, using the terminology "expert or novice" in a study of problem solving or "physics identity" in a study on identity and equity.[7] Thereafter, we constructed topic centroid vectors for each category by taking the collective mean of the embeddings vectors of all articles from every associated subtopic and computed topic scores using the same distance and transformation functions as before [Eq. (3)] with a high alpha value ($\alpha = 200$, approximately double the $\alpha$ from the prior analysis) to encourage the model to assign high primary topic scores.

To evaluate the sensitivity of this model to our chosen $\alpha$, we performed a similar procedure to that done with the LDA-based model: first, we reran the model with $\pm 10\%$ of the chosen value of $\alpha' = 200$ (that is to say $\alpha = 180$ and $\alpha = 220$) and computed the mean variance in primary topic scores. This produced a mean variation of $\langle |\Delta T_\alpha| \rangle = 0.06$ over a range of $\Delta\alpha = 40$, giving a sensitivity in primary topic score per unit alpha of $\langle |\Delta T_\alpha| \rangle / \Delta\alpha = 0.0015$. The mean primary topic score at $\alpha' = 200$ was calculated to be $\langle T_{\alpha'} \rangle = 0.58$, and dividing this sensitivity by the mean

[5]In topic 4, the subtopic *complex models of student learning* was dropped due to the narrowness of the topic and difficulty in operationalizing it.

[6]In topic 6, the subtopic *instructor implementations of reformed curricula* was dropped due to thematic inconsistency (i.e., this topic seemed more appropriate with curriculum and instruction)

[7]We note that most of the referenced articles in Docktor and Mestre's literature review came from outside the PERC proceedings. In this analysis, we restricted ourselves to defining centroids using articles within the dataset to see how the model would classify those articles.

gives us a percent sensitivity of approximately 0.5% change in primary topic score per unit $\alpha$ around 200.

To evaluate the sensitivity of results to our particular choice of model, we chose to vary the papers used to define the centroids rather than the scaling function or distance metric, since there were no *a priori* ways to benchmark those parameters. In other words, inherent to this technique is a sensitivity to the particular articles used to define the topic centroids, which likely will have a significant effect on model results. To test this sensitivity, we conducted a centroid resampling procedure, in which we created new topic centroids by repeatedly sampling the subtopic articles from each topic. Specifically, for each of the eight topics, we sampled sets of three articles from the four in each subtopic and took the collective mean of their embeddings vectors to define alternative topic centroids, which were then used to compute topic scores using the method described above. This procedure was repeated 1000 times to provide a measure of the spread of resulting topic scores.

Upon investigation, the mean topic scores from these 1000 models were most aligned with the set of topic centroids based on all available text data (that is, averaging all four article vectors from every subtopic). We, therefore, used topic centroids constructed from all articles on each topic to produce a final analysis of the topical prevalence as a function of time in the data corpus. The standard deviation of the 1000 resampling models was used to set error bars on this prevalence. We also produced plots of normalized topic prevalence over time and heatmaps of topic scores for individual articles to establish model face validity.

## IV. RESULTS

### A. Updated LDA analysis and embeddings analysis using LDA-derived topics and centroids

As discussed, to address research question 1, our first methodological step was to extend the LDA analysis up to the most recent published year (2023, as of writing) and to compare the performance of an embeddings analysis, applied deductively using topic centroids based on LDA model output, to these results. Figures 4 and 5 show stacked area plots for the updated LDA model and the embeddings model, respectively. In both cases, topic prevalence has been normalized each year to account for the changing number of publications as a function of time, which varies between a minimum of 28 papers in 2002 and a maximum of 120 papers in 2017 (an average of 76 papers per year). This normalized topic prevalence thus represents the percent of each topic in the total literature within each year.

Inspection of the two plots shows that they are very similar, which indicates that the embeddings model recovers a significant amount of the aggregate trends found by LDA: for example, the significant growth in the community and identity topic over time, a small early bump in qualitative theory studies, and surge in problem-solving research in the middle years, as well as the fluctuation of research on representations. Additionally, many of the smaller spikes and dips in topic prevalence from the LDA plot are present in the embeddings model results.

To measure divergence between the two models, in addition to the Jensen-Shannon divergence, we also calculated the summed absolute value of the topic-by-topic difference between LDA and embeddings scores for each article. For a particular article, this value could hypothetically range from 0 (if both the LDA and embeddings models assigned identical topic scores) to 2 (if there was no overlap between topic scores), meaning the aggregate score could range from 0 to 3490 for the entire corpus. At the chosen $\alpha = 94$, the calculated aggregate difference was approximately 1061, indicating that the embeddings model seems to have recovered approximately 70% (2429/3490) of the cumulative topic scores assigned by LDA. We note that this measure, when plotted as a function of $\alpha$, also had a minimum at approximately the same $\alpha$ value as the JS-divergence plot. Thus, the similarities in Figs. 4 and 5 are to be expected, since the embeddings model has recovered approximately 70% of the total topic scores assigned by LDA.

However, as can be seen from Figs. 4 and 5, there are some large-scale differences between the two models. The embeddings analysis shows a higher prevalence of research related to lab instruction, problem solving, and assessment (concepts) topics compared to LDA, and a lower overall prevalence of research on representations, community and identity, assessment (demographics), and qualitative theory.

To understand why the models differ, in Fig. 6, we present a snapshot of articles with approximately mean JS divergence (mean JS divergence = 0.14, standard deviation 0.09) between the topic scores from the two models. These articles can be considered representative of how the two models scored articles in the data corpus on average. As one can see, the topic scores by LDA and embeddings overlap significantly in many cases, which explains why the general features of the two graphs are so similar. However, the details of the topic scores sometimes differ. For example, article 6(a) *Studying Expert Practices to Create Learning Goals for Electronics Labs* was scored by both LDA and embeddings as approximately 43%–48% lab instruction, but LDA additionally scored it as 35% community and identity while the embeddings model scored it as 28% RBIS. Both of these choices make sense, as the article focuses on expert practice (a community of practice perspective) and learning goals-driven design. Article 6(c), *Observing Teaching Assistant Differences in Tutorials and Inquiry-Based Labs,* was scored by LDA as 77% RBIS and 15% lab instruction, while the embeddings model scored it as 44% RBIS and 26% lab instruction, with an additional 11% problem solving (likely due to the focus on tutorials).

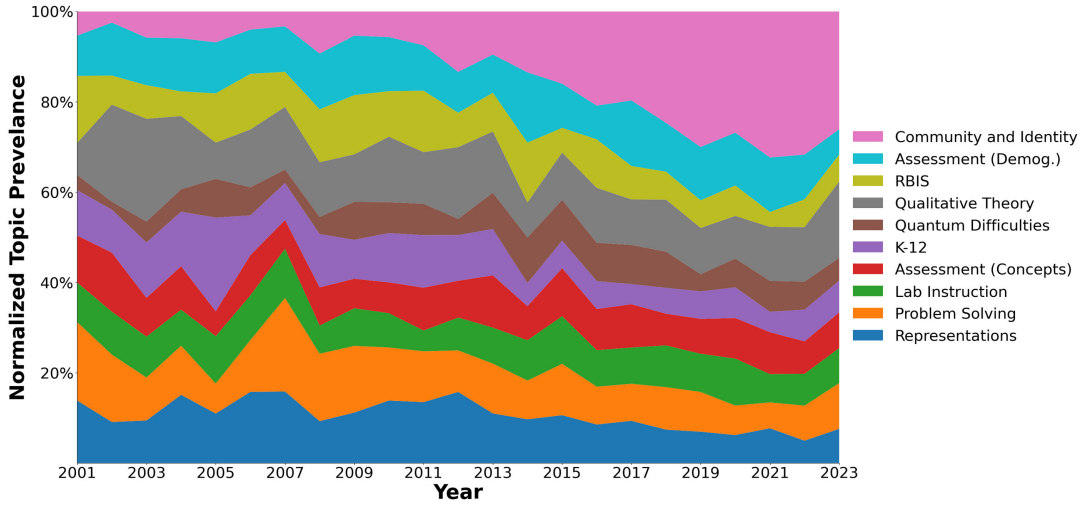


FIG. 4. LDA-based calculation of normalized topic prevalence as a function time in PERC Proceedings 2001–2023, based on LDA model from Odden *et al.* [23].

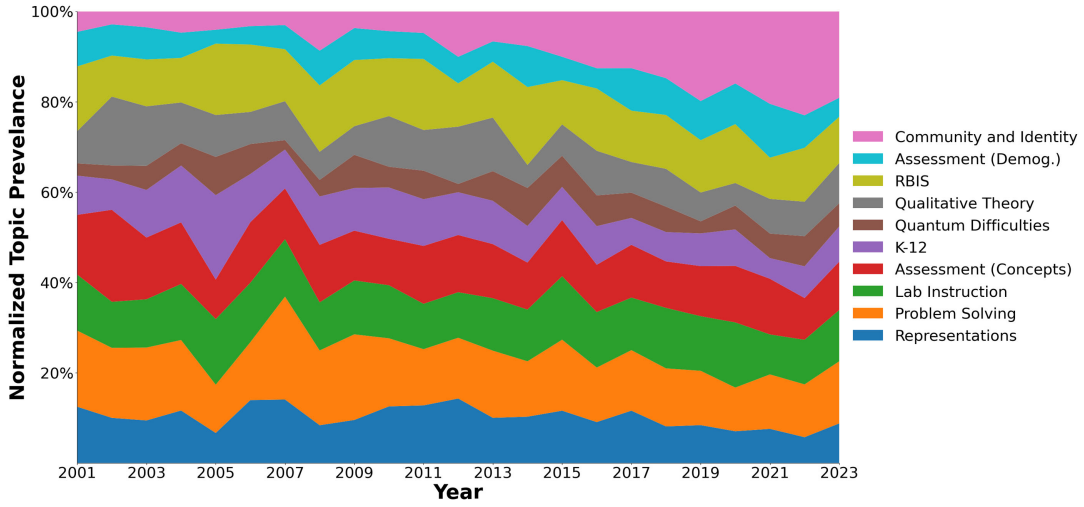


FIG. 5. Embeddings-based calculation of normalized topic prevalence as a function time in PERC Proceedings 2001–2023, using LDA-derived centroids.

Figure 6 also reveals that the primary difference between the models seems often to come from cases in which one model has smeared topic scores over several topics, giving low scores across many categories. For example, both models have similar primary topic scores for article 6(d), *3rd grade English language learners making sense of sound:* 49%–54% qualitative theory, likely due to the focus in the study on student mechanistic reasoning and discourse. However, LDA has additionally scored this paper as 32% K-12 and 11% community and identity, while the embeddings analysis has scored it as 13% K-12 and assigned low scores over most of the remaining categories. A similar dynamic is visible in article 6(e), *Examining the consistency of student errors in vector operations using module analysis*, which was scored by both LDA and the embeddings model as 37%–41% representations and 22%–28% assessment (concepts), likely because the study analyzed student multiple-choice responses; however, LDA has smeared the remaining topics scores over quantum difficulties (likely due to the focus on vectors), assessment (demographics), and community and Identity, while the embeddings model has assigned a 19% score of problem solving.

Examination of the articles with high JS divergence, shown in Fig. 7, shows a similar dynamic, in which some discrepancies can be traced to interpretable differences in model focus and others seem to be due to models smearing scores across multiple categories. For example, LDA classified four of the five articles as primarily quantum difficulties: 7(a) *Students' Understanding of the Concepts of Vector Components and Vector Products*; 7(b) *Student Difficulties with unit vectors and scalar multiplication of a*

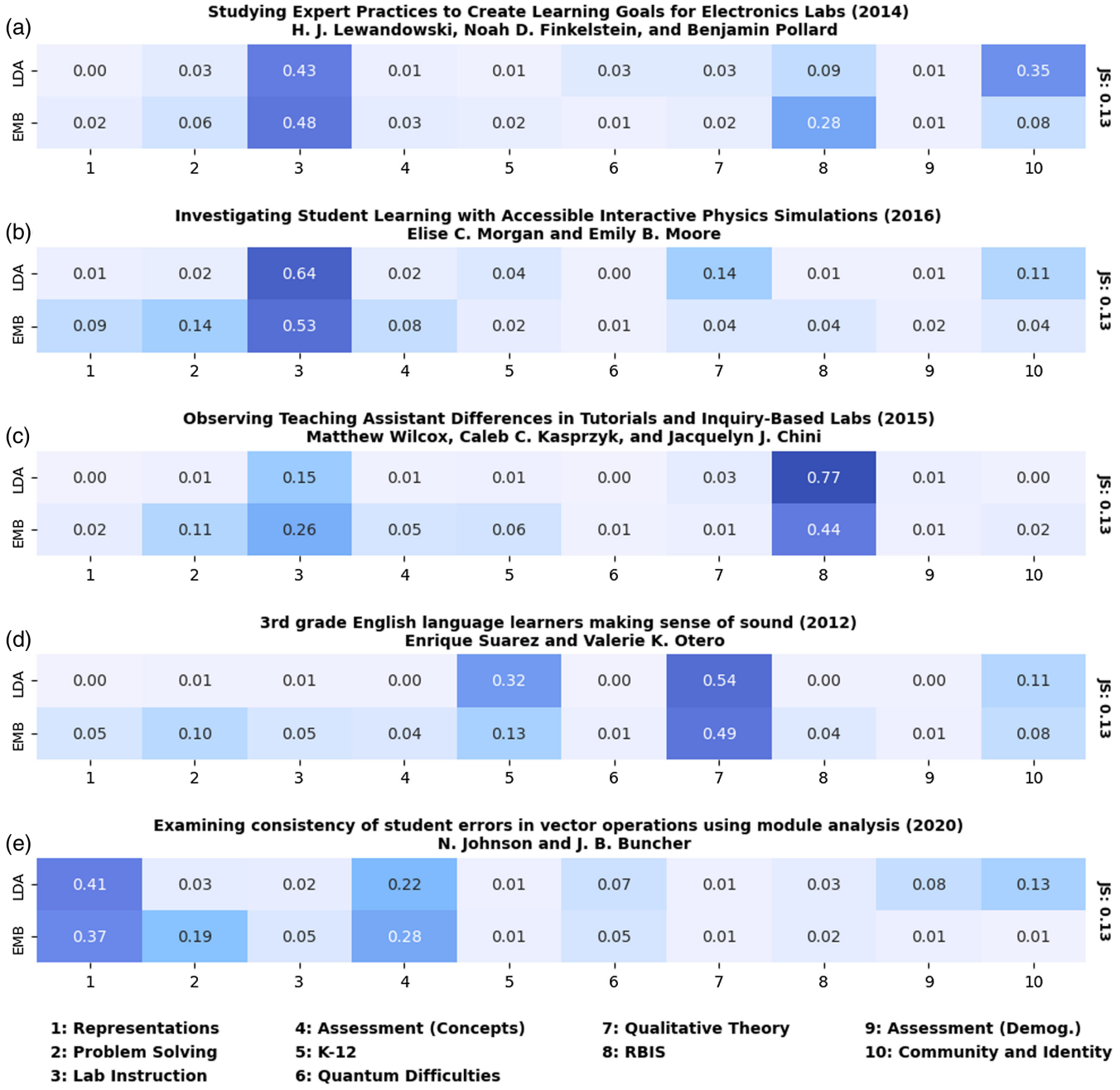


FIG. 6. Heatmap of topic scores for five articles (a)–(e), scored using ten topics by LDA and embeddings models. Articles were chosen because they have approximately mean Jensen-Shannon divergence (a measuring difference between scores) between LDA and embeddings topic scores. Darker colors indicate a higher percent of individual topics.

*vector*; 7(d) *Quantum Mechanics Students' Understanding of Normalization*; and 7(e) *Students understanding of dot products as a projection in no-context, work, and electric flux problems*. This is likely because the vocabulary of these articles frequently included words related to student conceptual difficulties and vectors, both of which are characteristic of this topic. The embeddings model primarily classified all four articles as focusing on representations, which also makes sense given the focus on vectors and normalization; this is likely because the embeddings model has more sensitivity to the semantic meaning of the text as a whole. Interestingly, three of these articles come from the same research group, suggesting that the models have consistent disagreement on articles about similar topics by similar authors. In the final case, LDA has scored the article primarily as one topic—assessment (demographics)—while the embeddings model has smeared the topic scores over 4–5 different topics.

To get a sense of how representative these differences are across the two models, Fig. 8 shows a violin plot of the

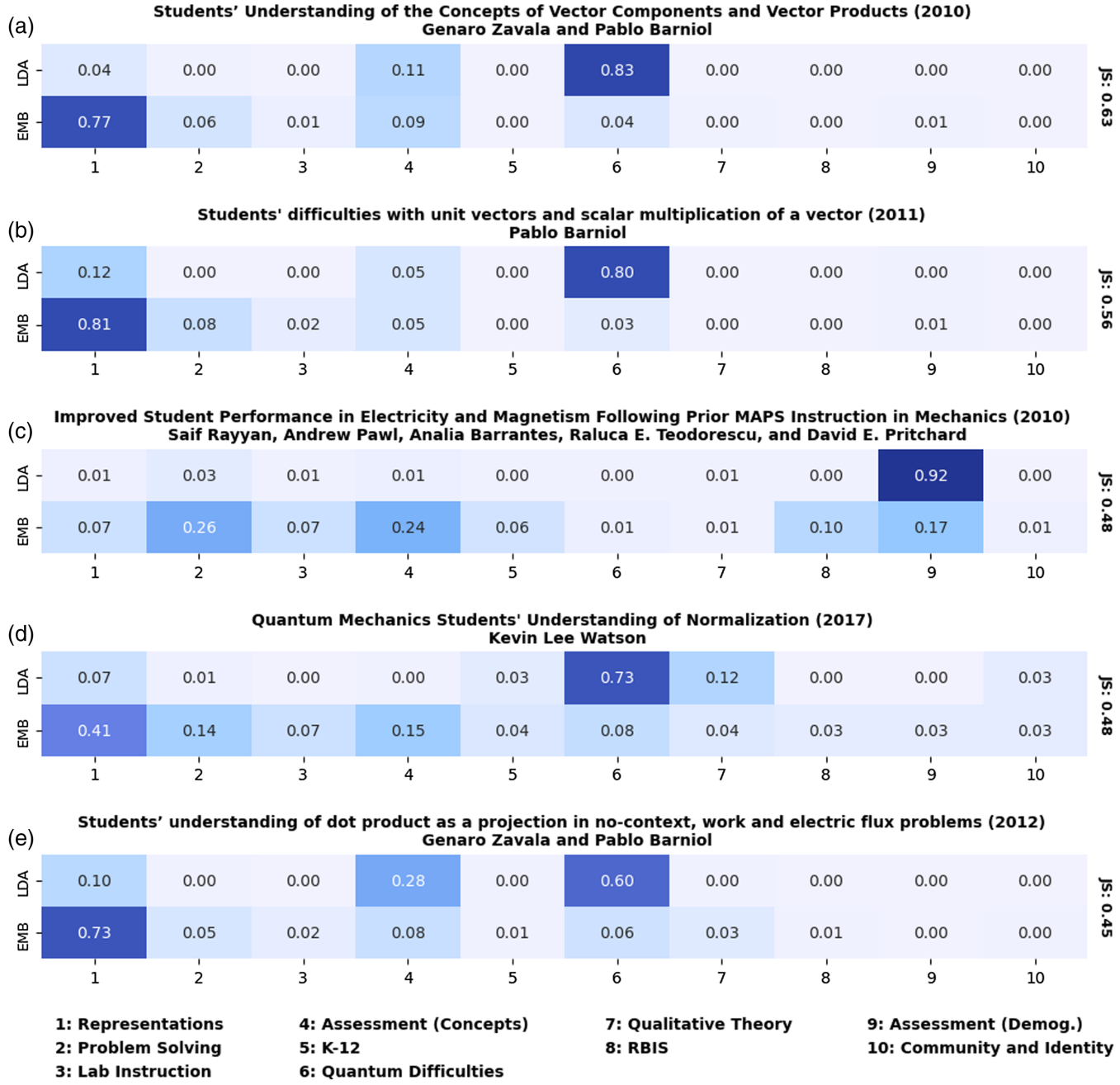


FIG. 7. Heatmap of topic scores for five articles (a)–(e), scored using ten topics by LDA and embeddings models. Articles were chosen because they had the highest Jensen-Shannon divergence (that is, topic disagreement) between LDA and embeddings topic scores. Darker colors indicate a higher percent of individual topics.

distributions of primary, secondary, and tertiary topic scores—that is, the highest, second highest, and third highest topic scores on each paper—from both the LDA and embeddings analyses. This plot shows that at the chosen $\alpha$ value (94), the embeddings analysis tended to produce more mixed topic scores than LDA, with a lower median value for primary topic scores (0.42 versus 0.50 for LDA) and an overall lower peak to the distribution. Although not shown in this figure, mean values for primary topics also followed this pattern, with LDA's assigned primary topic scores having a mean value of 0.53 and standard deviation of 0.17, as compared to a mean score of 0.47 and standard deviation of 0.20 for the embeddings model. This indicates that LDA had a tendency to produce more distinct and focused topic scores than the embeddings model (at the chosen alpha value), since it tended to assign higher scores to individual (primary) topics on average.

We also note that two of the articles in Figs. 6 and 7, articles 6(c) and 7(c), were in fact used to define the centroids for the embeddings analysis. This shows that even

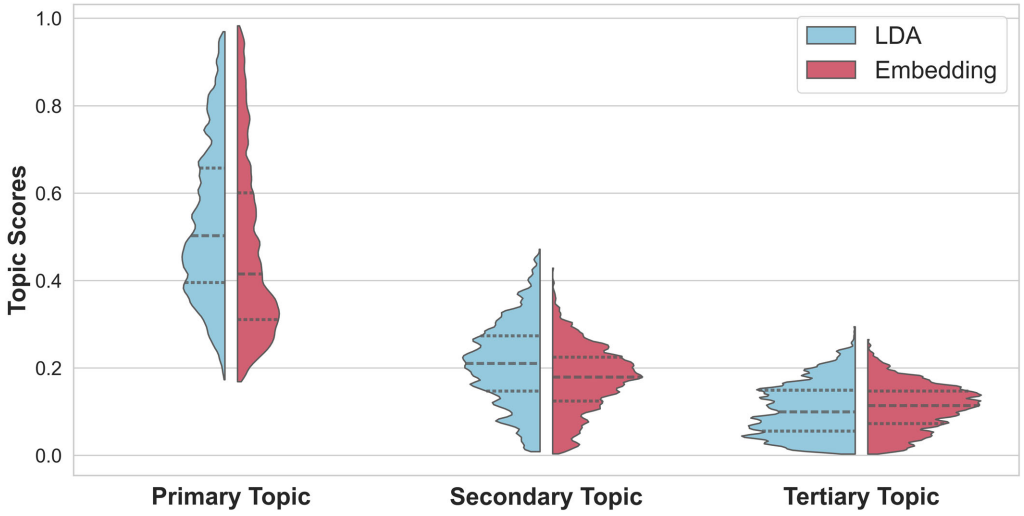


FIG. 8. Violin plot of primary, secondary, and tertiary topic scores for each article, classified by LDA and LDA-derived embeddings model with $\alpha = 94$. Widths of curves show the approximate distributions of topic scores by each model across the entire dataset. Within the curves, heavier dashed lines show the median values, and lighter dashed lines show the interquartile range of the distributions.

articles used to define centroids can still receive mixed topic scores from the embeddings model. This, however, is to be expected since many papers focus on multiple topics, and different models will weight these topics differently in their scores. By defining a centroid as the mean of several articles' embeddings vectors, we hope to "wash out" these secondary topics in order to retain only the meaning of the primary topic of the centroid.

What do these differences mean for our interpretation of the different models? First, we stress that this tendency in the embeddings model is due in large part to our choice of $\alpha$ scaling parameter. Exploratory modeling showed that with increasing $\alpha$, the mean and peak of the primary topic distribution for the embeddings model shifted toward 1 (as is shown in Fig. 10), and significantly more articles were classified as nearly 100% of a single topic. Furthermore, some differences are to be expected, as the two models are attending to different features of the text and arriving at their topic scores using very different mathematical models: LDA is attending to vocabulary use in the texts, and essentially performing a kind of probabilistic matrix factorization in order to extract a small number of "basis vectors" for this vocabulary [33]. Embeddings, in contrast, focus on the meanings of the texts (operationalized as aggregate relationships between the words) and provide topic scores based on text similarities in a high-dimensional "meaning space." Thus, it seems natural to expect that the two models will often differ in their particulars, much as two human coders might differ when attending to slightly different features of a dataset. Even given these differences, the two models have a high degree of overlap (approximately 70%), both in aggregate features (Figs. 4 and 5) and categorization of specific articles.

It should also be stressed that, prior to the current study, it was not clear that a tool like embeddings, applied deductively using a small sample of articles (15 per topic), would be able to reproduce *any* significant features of a topic analysis like that produced by LDA. Thus, this result also acts as a proof-of-concept that such an analysis is possible, as well as offering initial ideas on how to benchmark embeddings-based results to other human or machine-derived analyses. Although there are clear differences in results, we argue that these differences should be used for comparison of the two models and understanding of the embeddings analysis technique, but they do not speak to any specific "ground truth" regarding either model [35]. What can be said is that both analysis methods agree on some general trends in the data: for instance, a significant increase in studies focusing on student identities and communities of practice since around 2011; an early focus in the field on qualitative theory building; and a focus in middle years on research related to problem solving; along with an ongoing focus on all other topics (to varying degrees).

### B. Embeddings analysis using human-derived topics and centroids

We now turn to our second analysis, in which we used the same embeddings-based approach to reanalyze the PERC proceedings literature using fully researcher-defined categories based on the review and synthesis of discipline-based education research in physics by Docktor and Mestre [45], plus the additional categories of *precollege physics education* and *identity and equity*. This analysis addresses our second research question.

We first examine the face validity of this model in order to establish whether it is, in fact, producing interpretable results. To begin, Fig. 9 shows a heat map of the individual articles with highest topic prevalence (not used to define the topic centroid) for each of the eight topics. These articles

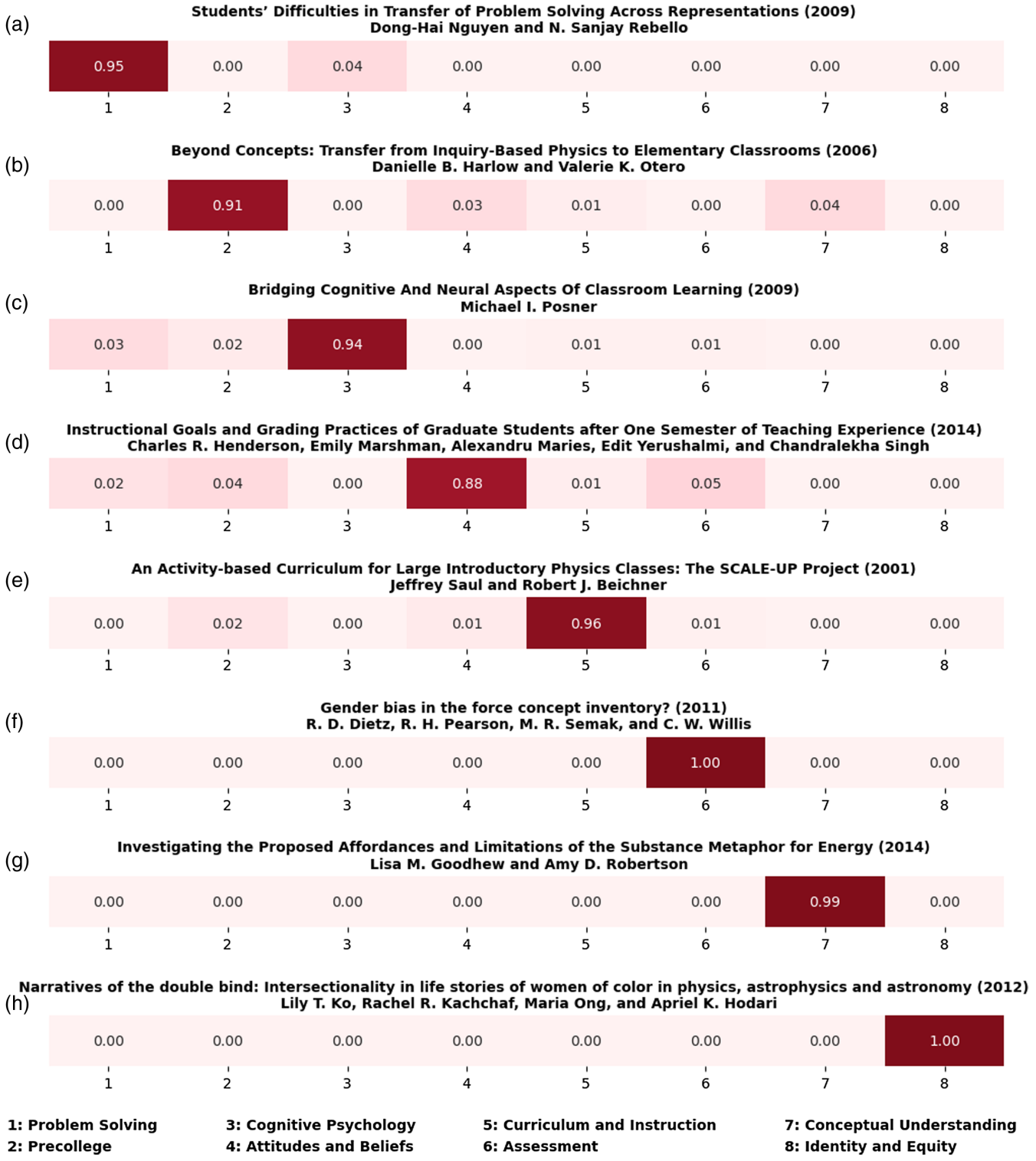


FIG. 9. Heatmap of embeddings model topic scores for eight articles (a)–(h), scored according to eight researcher-defined topics. Articles were chosen because they have the highest prevalence of each researcher-defined topic but were not included in centroid definitions. Darker colors indicate a higher percent of individual topics.

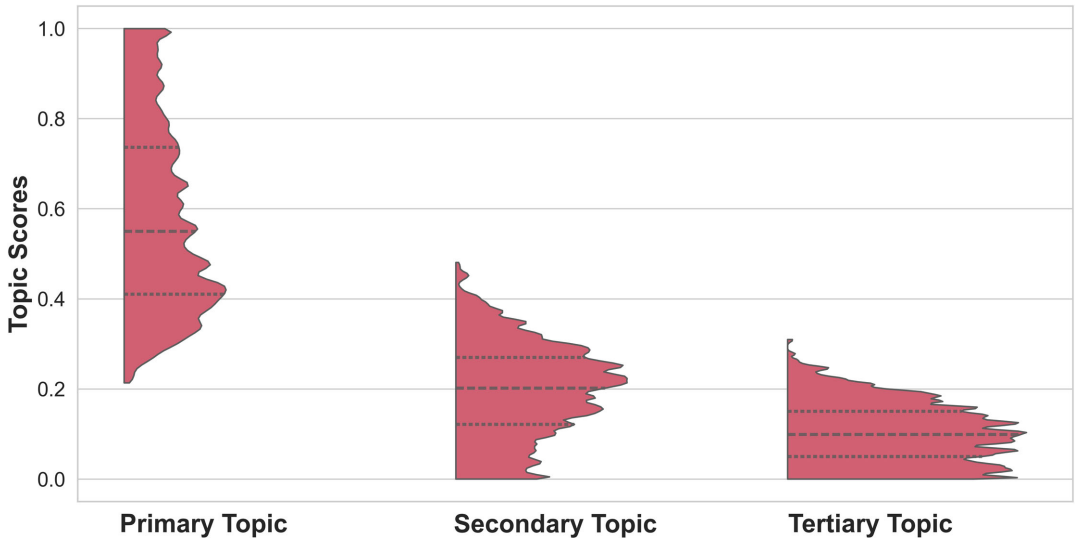


FIG. 10. Violin plot of primary, secondary, and tertiary topic scores for each PERC proceedings article, produced using researcher-defined topic centroids based on Docktor and Mestre with exponential scaling, cosine distance, and $\alpha = 200$. Widths of curves show the approximate distributions of topic scores by each model across the entire dataset. Within the curves, heavier dashed lines show the median values, and lighter dashed lines show the interquartile range of the distributions.

can be considered characteristic of each topic, and an inspection of these articles confirms that they all are substantively related to their primary topic, often referencing named research methods, pedagogical strategies, or theoretical frameworks in their titles. For example, article 9(f), *Gender bias in the force concept inventory?,* was classified as 100% assessment; this makes sense, given the reference to the force concept inventory. Article 9(e), *An Activity-Based Curriculum for Large Introductory Physics Classes: The SCALE-UP Project,* was classified at 96% curriculum and instruction, which fits well considering the fact that it focuses on the well-established SCALE-UP pedagogy. Article 9(c), *Bridging Cognitive And Neural Aspects Of Classroom Learning,* was classified as 94% cognitive psychology, which also makes sense considering the study's focus on cognition and neuroscience. The other articles show similar levels of interpretability, solidly matching specific subtopics from the literature review by Docktor and Mestre.

Importantly, this embeddings-based analysis is still a mixed-topic model, as shown in the histogram of Fig. 10. Even though the scaling factor alpha has been set to a high level ($\alpha = 200$), the distribution of primary topics for articles varies between 0.2 and 1, with a median of approximately 0.55. 95 articles in the dataset have a primary topic score of 95% or above, which produces a spike at the top of the primary topic distribution. This is to say, despite the high $\alpha$ value, most articles have been assigned a mixture of topics.

To further explore these classifications, in Fig. 11, we present a sample of articles whose primary topic has a score of around 50%, which provides a second face-validity check.

Figure 11 shows that articles with a mixture of topics are also usually quite interpretable. For example, article 11(a), *Understanding student computational thinking with computational modeling,* was rated as 50% problem solving, 25% precollege physics education, and 17% curriculum and instruction. This article analyzed how ninth-grade students learned computational thinking practices while taking a variation of Arizona State University's modeling instruction curriculum, assessed using a proctored programming assignment; it, therefore, makes sense that the article would be scored as a combination of problem solving (due to the programming assignment), precollege physics education (due to the ninth-grade students) and curriculum and instruction (due to the modeling curriculum). Article 11(c), *Effects of Visual Cues and Video Solutions on Conceptual Tasks*, analyzed how students solved physics transfer tasks using different kinds of video solutions; it thus makes sense that this article was classified as 48% cognitive science (due to the transfer task focus) and 48% problem solving (due to the worked solutions). Article 11(g), *A case of resources-oriented instruction in calculus-based introductory physics*, was scored as 51% conceptual understanding, 12% curriculum and instruction, and 23% precollege physics education. This study focused on a curricular approach based on the conceptual resources model, and the authors had an explicit focus on TA professional development based on the literature from K-12 teacher training, which explains these scores. Article 11(h), *Physics teachers integrating social justice with science content*, was rated as 50% equity and identity, 45% precollege physics education, a reasonable score for a paper about high

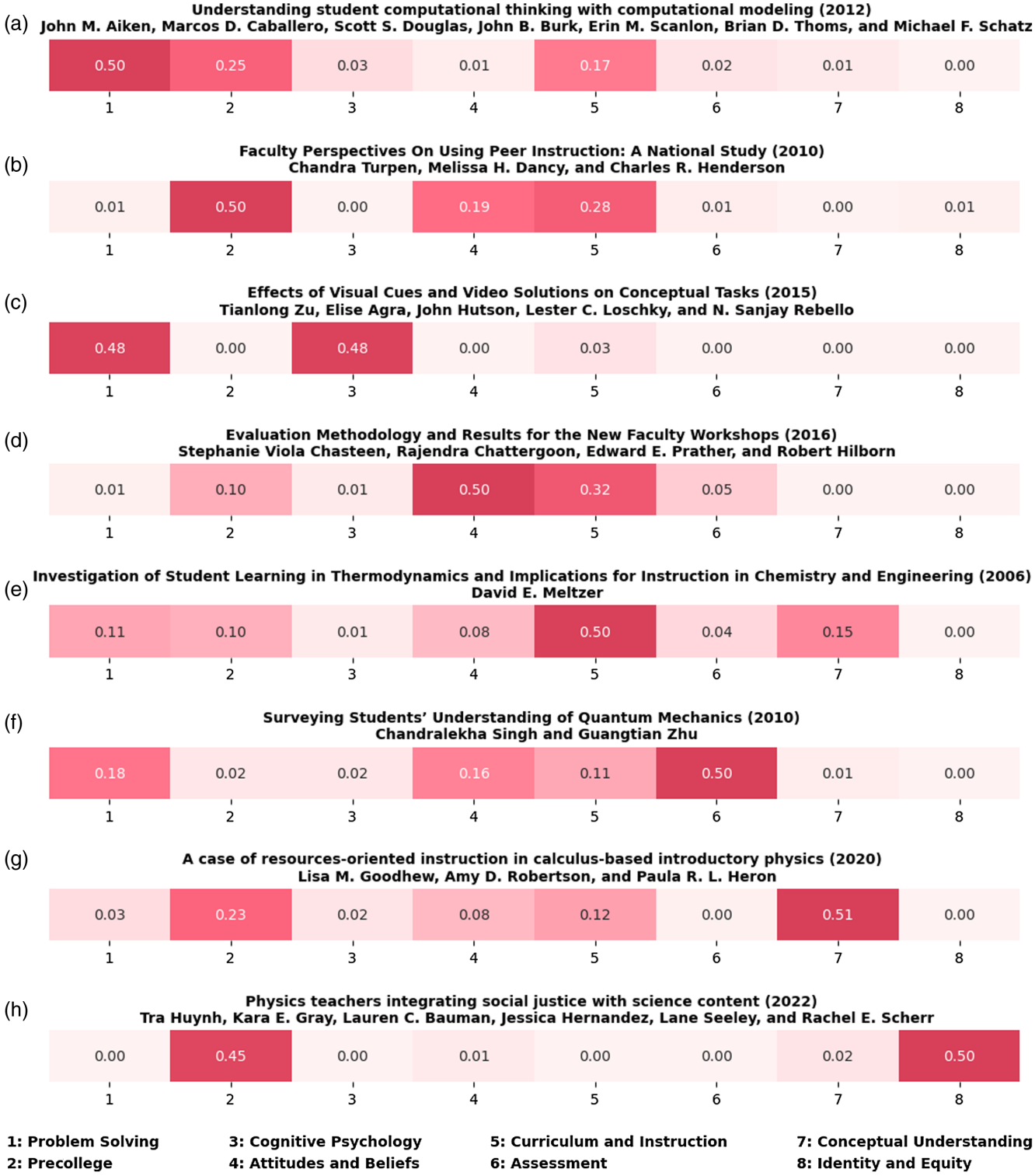


FIG. 11. Heatmap of embeddings model topic scores for eight example articles (a)–(h) scored according to eight researcher-defined topics. Articles were chosen because each includes approximately 50% prevalence of a researcher-defined topic. Darker colors indicate higher percent of individual topics.

school physics teachers making connections between their teaching and social justice.

Figure 11 also shows, however, that some papers were given low scores across several categories. For example, article 11(e), *Investigation of Student Learning in Thermodynamics and Implications for Instruction in Chemistry and Engineering*, received scores of around 10%–15% on four different categories (attitudes and beliefs, conceptual understanding, precollege, and problem solving). This type of spread in topic scores is a clear challenge to the interpretability of this model; however, it should be noted that studies like this one may be difficult for even a human researcher to clearly classify into a single primary category, since they focus on topics like chemistry and engineering education which lie outside the standard range of PER topics and thus are not necessarily a clear fit with any single category. Such "edge cases" are common in all forms of qualitative data analysis, and so the fact that this model had difficulty classifying them is somewhat expected. We, as researchers, also have some control of the degree of topic "smear" using the tunable $\alpha$ parameter, since the number of low topic scores per article is inversely related to $\alpha$ (as can be seen by comparing the embeddings score distributions in Figs. 8 and 10, with $\alpha = 94$ and 200, respectively). However, this example shows that even very high $\alpha$ values will not necessarily eliminate such cases, especially for texts that fall outside the standard range of PER literature.

In summary, based on both the inspection of specific articles and the general distribution of topic scores, we argue that the model seems to be producing results with face validity.

### C. Trends in PERC proceedings literature derived from deductive embeddings analysis

Addressing our third research question, we now use this topic analysis to re-examine the development of the PERC proceedings literature over time.

Figure 12 shows plots of the cumulative prevalence over time of these eight researcher-defined topics. Here, *cumulative prevalence* is the sum, each year, of the topic contributions from each individual article. Thus, this measure can be thought of as the approximate number of whole papers published on that topic each year. Error bars have been created using the centroid resampling procedure described above, where the width corresponds to three standard deviations of the spread from 1000 centroid resampling runs.

From this plot, certain trends stand out. First, all topics have seen growth in the early years, in part because the number of published articles has increased from around 30 in 2001–2002 to over 110 around 2019. All topics also have seen decline after the COVID-19 pandemic, when the number of publications reduced by approximately 30%–40% from prepandemic highs, to 70–80 papers per year in the period of 2021–2023. However, some topics have seen significantly more growth than others. Chief among these is identity and equity, which grew from nearly 0 to one of the highest-prevalence topics in recent years.

Other topics have seen consistently high or low prevalence over time. Cognitive psychology, for example, has often hovered around five articles per year, and attitudes and beliefs has had around 10 articles per year at its peak, while curriculum and instruction and problem solving have both had consistently high prevalence over most of the PERC proceedings history.

Still other topics have seen significant shifts over time. Research on conceptual understanding appears to have had a major surge around 2012, followed by a major dip, which is also present to a lesser degree in research on precollege physics education and problem solving.

To put these shifts in context, Fig. 13 shows an alternate view of topic prevalence over time, in which all topics have been normalized to 100% each year. This view controls for changes in conference attendance and publication and allows one to see that all topics in the model have received some amount of consistent interest over time, albeit at different levels. Some (like cognitive psychology and attitudes and beliefs) have had low prevalence for most of the PERC publication history, suggesting that they might either be more niche topics or are more commonly used in combination with others, like cognitive psychological techniques for eye-tracking as a method for studying problem solving or student beliefs as an outcome of educational reforms. However, even when controlling for variations in article publication numbers, certain topics have seen major surges or increases: Assessment appears to have had a small surge in the early years of the PERC proceedings, and identity and equity has seen consistent yearly growth since around 2016, topping out at around 20% of the literature in 2022. In contrast, research on conceptual understanding appears to have waned significantly in recent years, although it is still represented in the literature.

What caused these changes? Some, like the growth in research on identity and equity and the decline of research on conceptual understanding, seem to represent long-term trends in the field. Such changes are expected—all research fields shift over time, including education research [25]. It is important to note, however, that these literature shifts have not occurred in a vacuum. The PERC proceedings is the product of a living community of researchers, whose work is dependent on support from universities and government foundations. The physics education research community has changed radically over the years of the PERC proceedings publication, going from a small field in the process of establishing itself in the early years to a recognized field with its own specific journals and

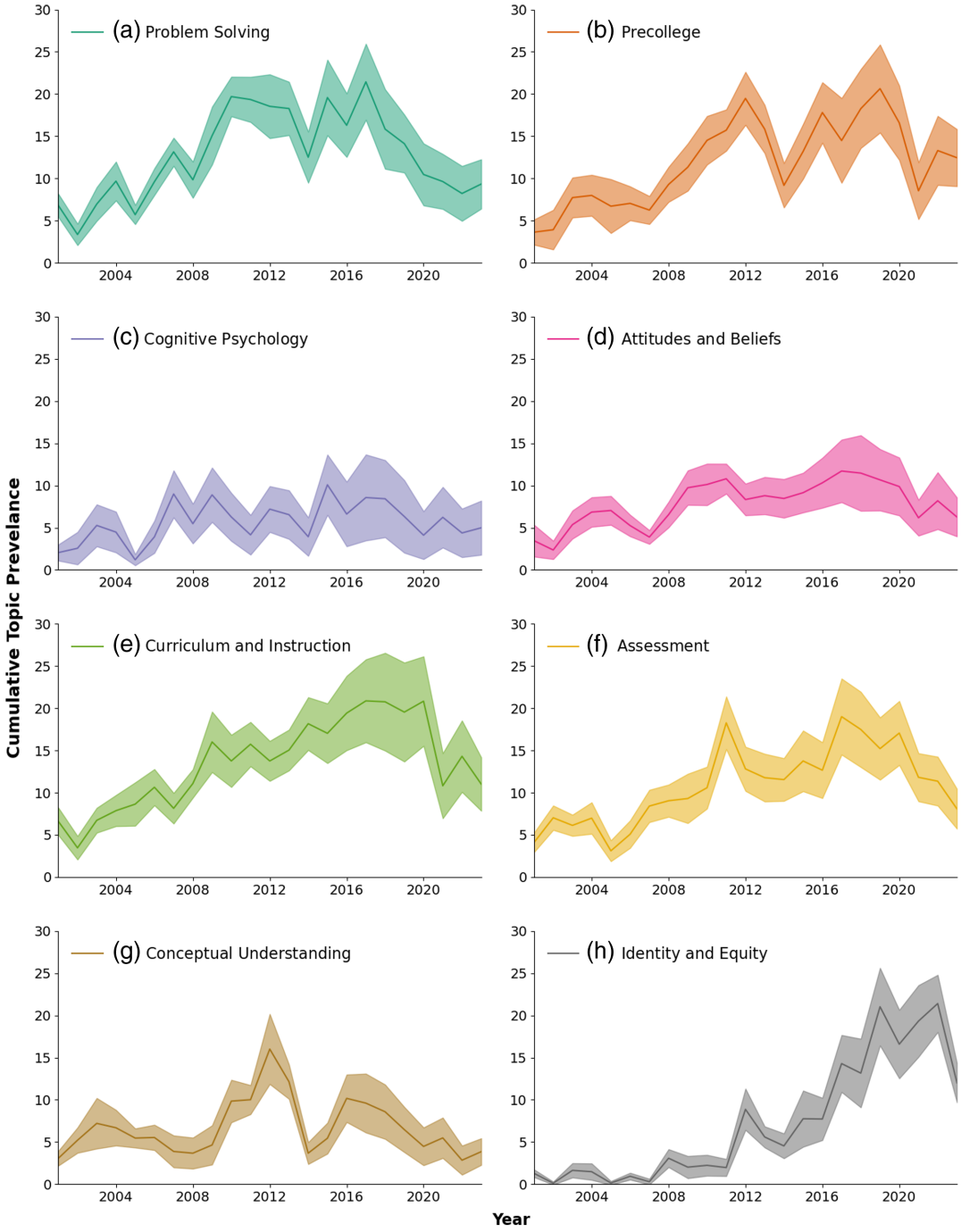


FIG. 12. Cumulative topic prevalence over time for researcher-defined topics. *Y* axis is equivalent to approximate number of complete articles published on each topic per year. Error bars are three standard deviations of topic prevalence spread based on centroid resampling.

conferences in the present day. Additionally, factors like editorial decisions, publication norms, and conference locations have also certainly affected publication numbers and topics. An interesting area of future research would be to check the shifts visible in Figs. 12 and 13 against historical records of funding initiatives and conference participant recollections.

## V. DISCUSSION

### A. Reflections on the method

Our first two research questions were as follows: *How can embeddings be used to reproduce prior inductive qualitative research on scientific literature like the PERC proceedings?* and, *how can embeddings be used*

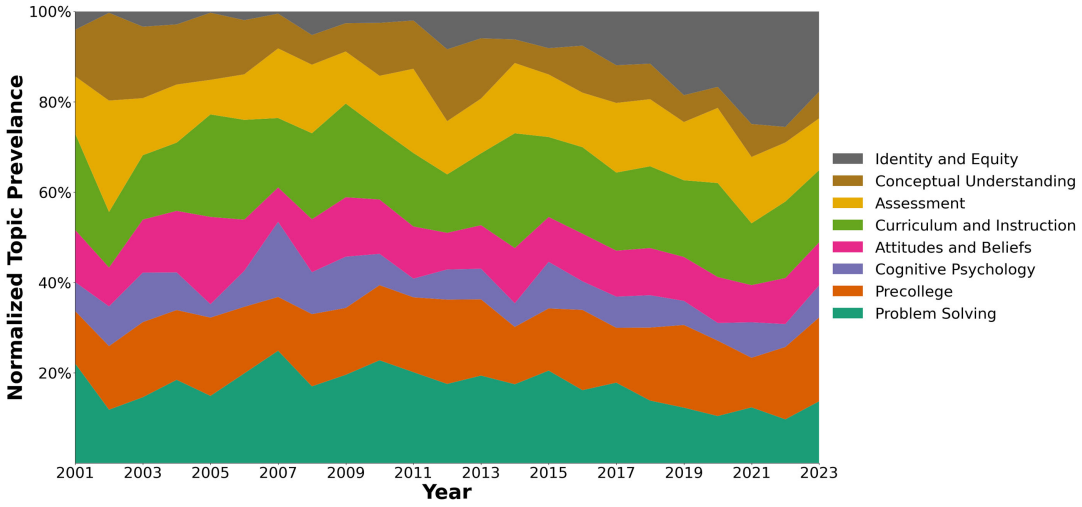


FIG. 13. Embeddings-based calculation of normalized prevalence as a function of time in PERC proceedings literature using researcher-defined topics.

*to perform deductive analyses of literature using researcher-defined categories?* To answer these questions, we have defined and applied a method that uses embeddings to both reproduce prior inductive analysis and perform a new deductive analysis of the literature using categories drawn from a traditional literature review. Using only 15 representative articles from each of the LDA-defined topics, we were able to recover most of the recognizable aggregate trends in the LDA model as well as approximately 70% of the specific topic scores. Furthermore, the article categorizations by both the LDA-based embeddings model and the embeddings model based on Docktor and Mestre's literature review [45] both have substantial face validity.

Returning to our third research question, *what trends do embeddings-based analyses reveal in the development of the literature as a function of time?* We can use these three analyses to make some arguments about the development of the PERC proceedings literature, and arguably the state of American PER, as a function of time. First, all three analyses of the PER literature suggest that research on identity, equity, and learning communities has become an increased focus of the field. A survey of recent PERC conference themes supports this conclusion: for instance, "Working together to Strengthen the PER Community of Practice" (2023), "Queering Physics Education" (2022), "Making Physics More Inclusive and Eliminating Exclusionary Practices in Physics" (2021), "Insights, Reflections, and Future Directions: Emergent Themes in the Evolving PER Community" (2020), and "Physics Outside of the Classroom: Teaching, Learning, and Cultural Engagement in Informal Physics Environments" (2019). However, the present analysis allows us to make some quantitative estimates on the approximate amount of attention this topic is getting from the field: approximately 20%–30%, depending on the model and whether communities of practice are combined with the topics of equity and identity. These trends appear to mirror trends in other related fields, like science education research, which has also seen a major shift toward research on equity, identity, and communities of practice in recent decades [25].

Given these results, embeddings analyses seem to offer significant advantages over other NLP approaches for qualitative data analysis. They require significantly less data cleaning and preprocessing than BoW approaches like LDA; as long as texts have been cleanly digitized, they can often be embedded and analyzed as is. They do not have the same issues with stochasticity as LDA-based methods [24,25]. They also only require a small amount of *a priori* categorization, in the form of choosing a few (10–15) representative articles for each topic centroid. In the present analysis, exploratory modeling suggested that more texts provide greater resolution, i.e., decreased JS divergence between LDA and embeddings; but there also seemed to be a point of diminishing returns, where doubling the number of articles only provides small decreases in JS divergence.[8] This suggests that researchers using this method may be able to reasonably define categories with only a few example texts, either taken from the dataset or constructed by the researcher.

Furthermore, the method behind this analysis was relatively straightforward: (i) embed texts; (ii) define centroids based on a selected sample of texts that exemplify desired features; (iii) calculate topic scores based on transformed distances, using scaling parameters to determine "mixedness" of the model; and (iv) evaluate results of the model analysis based on face validity of topic scores and general trends. Each of these components can easily be

[8]We provide an analysis of the effect of the number of centroid-defining articles on JS divergence in Supplementary Material of this article.

modified or replaced: for instance, the large language model used to create the embeddings can easily be swapped out or changed (and many such models are available, for free, on open-source repositories like HuggingFace); multiple distance measures can be used; transformation functions can be replaced, tuned, or modified; and data can easily be added to or removed from the model without any negative downstream effects. Finally, the texts used to define centroids can be easily, added, removed, or modified, and such texts can come from any source. This means that in literature reviews, researchers could use papers from outside the journal to define centroids, or even write and embed their own texts created specifically for centroid definition. For example, we could easily swap out the PERC proceedings articles used to define centroids for articles from other journals taken directly from the citations in Docktor and Mestre, or even use the topic vectors from this analysis to analyze other journals like *Physical Review Physics Education Research*.

Given this flexibility, we suggest that this technique holds great promise for helping researchers perform deductive qualitative analysis at scale on a wide variety of different types of data. Beyond literature reviews of research articles, one could easily imagine applying this same approach to code student responses on open-ended surveys or responses to conceptual physics questions, especially those that have already been partially analyzed by a researcher to define robust categories. One could also imagine segmenting interview transcripts and applying this type of analysis to different turns-of-talk or conversational segments in order to detect student ideas, attitudes, or misconceptions, similar to Sherin's [26] analysis. In this regard, this embeddings-based approach could act as a tool for both automating existing analyses and drawing new inferences from data. One could even imagine multiple cycles of analysis, in which edge cases like those described above are automatically detected (for example, by looking for topic "smear") and inspected by a human researcher, in order to modify or refine a coding scheme. This technique thus offers the opportunity to more flexibly integrate the capabilities of machine learning methods and human researchers, in line with calls by researchers like Kubsch *et al.* [46].

However, methods like these also raise important epistemological questions about the division between qualitative and quantitative methods in PER and their different standards of evaluation. We view this method as fundamentally a deductive, quantitative approach to analyzing qualitative data (text), appropriate for cases in which one has a substantial amount of text data that can be analyzed using researcher-defined categories. As such, it is reasonable to evaluate it according to standard criteria for the trustworthiness of qualitative research, such as credibility, dependability, transferability, and confirmability [8,47].

*Credibility* refers to the plausibility of the research results. In the present study, we have worked to demonstrate credibility through both check of face validity in classifications of individual articles and comparisons with results from LDA. Additionally, the results produce recognizable trends, such as the increasing prevalence of research on identity and equity in the field over the past 10 years.

*Dependability* refers to whether the research can be replicated in similar conditions. To demonstrate the dependability of our analysis, we have shared our code and data, to allow any interested reader to review the analysis, replicate it, and explore the results on their own [49]. Authors who have published in the PERC proceedings are encouraged to find their own papers and check the model's categorizations against their own judgment.

*Transferability* refers to the question of whether the results of an analysis can be applied to other groups outside the studied population. In the present literature review, we interpret this as asking whether the trends we have found may apply to other bases of literature, like *Physical Review Physics Education Research*. Although we have not yet confirmed it, we suspect the answer is yes, since many articles published in the PERC proceedings form the basis for longer papers in journals like PRPER. Such an analysis represents a clear next step for developing and applying this method.

Finally, the criterion of *confirmability* asks whether there is a clear link between the data and the findings. In this regard, we again point to both the face-validity checks (which show clear relationships between article content and categorizations) as well as the shared datasets and code, which clearly document the methodological link between the data and findings.

Although these methodological standards will require refinement in future studies, we see this as an important point for reflection by any researchers interested in using this method and encourage all researchers doing NLP-aided qualitative analysis to publicly share their datasets and analysis code.

### B. Limitations and future work

We have shown a proof-of-concept that embeddings can be used for deductive qualitative research. As discussed, other researchers have shown that embeddings can also be used for inductive qualitative research, such as computational grounded theory [31]. However, a great deal of work remains to understand how best to leverage embeddings in qualitative data analysis. For example, this study was done with an out-of-the-box, general-purpose large language model; such models can be domain adapted or fine-tuned to specific contexts, which allows them to better capture (and distinguish) the meaning of texts from those contexts. One could easily imagine domain adapting a model using literature from studies of physics teaching and learning (such as articles from *The Physics Teacher*, *Physics*

*Education*, or *The American Journal of Physics*) to provide better resolution on PER-specific language and ideas.

Regarding the method we have proposed, another important area of consideration is the $\alpha$ value. When one has external data (for example, texts that have been categorized by hand), one can tune this value based on that data, as we did in the LDA reproduction analysis. However, absent such data, one must choose an $\alpha$ value. Low values of $\alpha$ will tend to give distributions that have fairly similar scores across all topics, while very high $\alpha$ values will produce categorizations with nearly 100% of the score assigned to a single topic for each paper. So, setting this value amounts to asking the question "how often do I want my model to classify a particular paper as only one topic vs give it a mixed topic score?" Future work will need to critically engage with this question, especially in cases where researchers are analyzing larger bodies of literature or more heterogeneous textual datasets.

Further, there is a need to critically evaluate the way embeddings capture the meaning of texts. For instance, is a single vector enough to capture meaning from a text, or might one gain better insight by representing different parts of a text using multiple vectors? What are the benefits and trade-offs of different vector dimensionalities for different types and lengths of text?

We also see a need to explore the space of different distance measures and transformations. The transformation functions used in this analysis were fairly straightforward, primarily meant to invert and magnify minute differences in distance within a high-dimensional meaning space. They were chosen primarily due to their simplicity and familiarity, since they are common to many areas of physics and mathematics. However, it is not clear at this point what effect the particulars of the scaling function have on specific categorizations, beyond the general alignment with LDA results. There are certainly other more sophisticated transformations that may give better results for specific applications, especially when combined with different distance measures, and we view this as an important area of future work. In this regard, we can count ourselves lucky to be part of a field that has significant experience with using vectors to analyze the behavior of physical systems.

In the future, in order to critically evaluate this method, these types of analyses will need to be benchmarked against existing qualitative analyses, using standard statistical methods for evaluating inter-rater reliability. One downside of the present analysis is the fact that the embeddings analysis was benchmarked against another NLP model, and neither model was benchmarked against any kind of "ground truth" (in the form of human-performed qualitative analysis). This, we feel, was an appropriate approach for a proof-of-concept, and it offers interesting insights into the differences between LDA and embeddings for topic analysis. But it leaves open the question of how such models would perform relative to a human coder. This suggests that a natural next step would be to apply this technique to large, text-based datasets that have already been qualitatively coded, like those from Wilson *et al.* [19] or Sabo *et al.* [6], and compare the results to human codes. Given the recent focus on data availability in the research community, we hope that more researchers will be willing to publicly share their analyzed datasets alongside publications to aid in this type of analysis.

Finally, because this method is fundamentally a form of qualitative analysis, it inherits many of the limitations inherent to those family of methods. That is to say, the analytic choices the researcher makes (for example, which categories to include or which example texts should be used to define topic centroids) will have a significant effect on the results. Furthermore, the results of this type of analysis require some amount of further qualitative interpretation, like inspecting topic scores of articles to determine face validity and adjusting model parameters when results do not make sense.

There are certainly benefits to this NLP approach to qualitative analysis, in that it is scalable to large datasets and reproducible. However, there are also new potential sources of bias and issues with transparency: for example, although the general method for turning texts into vectors is well understood at a technical level, the actual details of the process will vary based on the particular large-language model used and researchers will need to carefully evaluate whether the models they are using are capturing the particular semantic meanings they care about in the texts. It may be that addressing these issues will require some kind of combination of traditionally qualitative and quantitative analysis methods.

## VI. CONCLUSION

Qualitative research has long been a time-consuming task, wherein the researcher themselves essentially acts as the analysis instrument [50]. Computational tools have been useful for managing datasets and handling the mechanics of the process, but their limited ability to extract meaning from text greatly constrained what they were able to do. Now, we as a field have reached a point where we can use AI to quantitatively analyze and compare meaning in text. This technological development will continue; already, companies are connecting large language models to image models, so soon it may even be possible to analyze meaning across multiple representational modes. Thus, the future of qualitative research will likely look radically different from what we have known so far. It is time that we, as a field, begin to explore the use of these new tools. We look forward to seeing what is possible and hope that these tools will soon help us to gain a deeper and broader understanding of physics teaching and learning.

The code and data used in this analysis are openly available on GitHub [51] and Zenodo [49].

## ACKNOWLEDGMENTS

We gratefully acknowledge the contribution of Alessandro Marin in developing the methods that led to this work. We would also like to thank Danny Caballero, Ben Zwickl, David Hammer, Noah Finkelstein, Valerie Otero, Kirsty Dunnett, Emily Bolger, and Cassandra Lem for their thoughts and feedback on this work. This work was funded by the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education (NOKUT), which supports the University of Oslo's Center for Computing in Science Education and Center for Interdisciplinary Education.

[1] D. Hammer, The necessarily, wonderfully unsettled state of methodology in PER: A reflection, in *The International Handbook of Physics Education Research: Special Topics* (AIP Publishing LLC, Melville, NY, 2023).

[2] C. E. Wieman, The similarities between research in education and research in the hard sciences, Educ. Res. **43**, 12 (2014).

[3] D. Hammer and A. Elby, Tapping epistemological resources for learning physics, J. Learn. Sci. **12**, 53 (2003).

[4] E. Kuo, M. M. Hull, A. Gupta, and A. Elby, How students blend conceptual and formal mathematical reasoning in solving physics problems, Sci. Educ. **97**, 32 (2013).

[5] E. Kuo, M. M. Hull, A. Elby, and A. Gupta, Assessing mathematical sensemaking in physics through calculation-concept crossover, Phys. Rev. Phys. Educ. Res. **16**, 020109 (2020).

[6] H. C. Sabo, L. M. Goodhew, and A. D. Robertson, University student conceptual resources for understanding energy, Phys. Rev. Phys. Educ. Res. **12**, 010126 (2016).

[7] V. Braun and V. Clarke, Using thematic analysis in psychology, Qual. Res. Psychol. **3**, 77 (2006).

[8] V. K. Otero, D. B. Harlow, and D. E. Meltzer, Qualitative methods in physics education research, in *The International Handbook of Physics Education Research: Special Topics* (AIP Publishing LLC, Melville, NY, 2023).

[9] A. A. Disessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).

[10] R. Staley, Physics as a human endeavor, in *The International Handbook of Physics Education Research: Special Topics* (AIP Publishing LLC, Melville, NY, 2023).

[11] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[12] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, arXiv:1301.3781.

[14] O. Press, N. A. Smith, and M. Lewis, Train short, test long: attention with linear biases enables input length extrapolation, arXiv:2108.12409.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA* (Curran Associates, Inc., Red Hook, NY, 2017), Vol. 30.

[16] J. M. Aiken, R. De Bin, M. Hjorth-Jensen, and M. D. Caballero, Predicting time to graduation at a large enrollment American university, PLoS One **15**, e0242334 (2020).

[17] P. Wulff, Machine learning in science education—Realized potentials, expected developments, and fundamental challenges, in *Proceedings of the Lehren Und Forschen in Einer Digital Geprägten Welt—GDPC Tagungsband 2023* (2023), https://gdcp-ev.de/wp-content/uploads/securepdfs/2023/05/02PV_Wulff.pdf.

[18] X. Zhai, Y. Yin, J. W. Pellegrino, K. C. Haudek, and L. Shi, Applying machine learning in science assessment: A systematic review, Stud. Sci. Educ. **56**, 111 (2020).

[19] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. J. Lewandowski, Classification of open-ended responses to a research-based assessment using natural language processing, Phys. Rev. Phys. Educ. Res. **18**, 010141 (2022).

[20] P. Wulff, D. Buschhüter, A. Westphal, A. Nowak, L. Becker, H. Robalino, M. Stede, and A. Borowski, Computer-based classification of preservice physics teachers' written reflections, J. Sci. Educ. Technol. **30**, 1 (2021).

[21] P. Wulff, L. Mientus, A. Nowak, and A. Borowski, Utilizing a pretrained language model (BERT) to classify preservice physics teachers' written reflections, Int. J. Artif. Intell. Educ. **33**, 439 (2023).

[22] R. K. Fussell, A. Mazrui, and N. G. Holmes, Machine learning for automated content analysis: Characteristics of training data impact reliability, presented at PER Conf. 2022, Grand Rapids, MI, 10.1119/perc.2022.pr.Fussell.

[23] R. K. Fussell, E. M. Stump, and N. G. Holmes, A method to assess trustworthiness of machine coding at scale, arXiv:2310.02335.

[24] 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).

[25] 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).

[26] B. Sherin, A computational study of commonsense science: An exploration in the automated analysis of clinical interview data, J. Learn. Sci. **22**, 600 (2013).

[27] P. Wulff, D. Buschhüter, A. Westphal, L. Mientus, A. Nowak, and A. Borowski, Bridging the gap between qualitative and quantitative assessment in science educa-

tion research with machine learning—A case for pretrained language models-based clustering, J. Sci. Educ. Technol. **31**, 490 (2022).

[28] J. M. Geiger, L. M. Goodhew, and T. O. B. Odden, Developing a natural language processing approach for analyzing student ideas in calculus-based introductory physics, presented at PER Conf. 2022, Grand Rapids, MI, 10.1119/perc.2022.pr.Geiger.

[29] L. K. Nelson, Computational grounded theory: A methodological framework, Sociol. Methods Res. **49**, 3 (2020).

[30] 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).

[31] P. Tschisgale, P. Wulff, and M. Kubsch, Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory, Phys. Rev. Phys. Educ. Res. **19**, 020123 (2023).

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. **3**, 993 (2003).

[33] M. D. Hoffman, D. M. Blei, and F. Bach, Online learning for latent Dirichlet allocation, in *Proceedings of the 23rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada*, edited by J. D. A. Lafferty, C. K. I. Williams, J. Shawe-Taylor, and R. S. Zemel (Curran Associates Inc., Red Hook, NY, 2010), pp. 856–864.

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

[35] J. Grimmer and B. M. Stewart, Text as data: The promise and pitfalls of automatic content analysis methods for political texts, Polit. Anal. **21**, 267 (2013).

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

[37] N. Reimers and I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, arXiv:1908.10084.

[38] E. Terreau, A. Gourru, and J. Velcin, Writing style author embedding evaluation, in *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems* (2021), pp. 84–93, 10.18653/v1/2021.eval4nlp-1.9.

[39] M. Günther *et al.*, Jina embeddings 2: 8192-token general-purpose text embeddings for long documents, arXiv:2310.19923.

[40] T. Schopf, D. Braun, and F. Matthes, Evaluating unsupervised text classification: Zero-shot and similarity-based approaches, in *Proceedings of the 6th International Conference on Natural Language Processing and Information Retrieval* (Association for Computing Machinery, New York, NY, 2023), pp. 6–15, 10.1145/3582768.3582795.

[41] R. E. Beaty and D. R. Johnson, Automating creativity assessment with SemDis: An open platform for computing semantic distance, Behav. Res. Methods **53**, 757 (2021).

[42] M. Grootendorst, 9 Distance measures in data science, https://www.maartengrootendorst.com/blog/distances/.

[43] OpenAI, OpenAI embeddings: Frequently asked questions, https://platform.openai.com/docs/guides/embeddings/frequently-asked-questions.

[44] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.20.020151 for (1) analysis of correlations between LDA and embeddings topic scores. (2) Complete list of PERC Proceedings articles used to define centroids for analysis based on literature review by Docktor and Mestre (2014). (3) Analysis of Jensen-Shannon divergence between LDA model and embeddings model classifications as a function of the number of articles in the centroid.

[45] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).

[46] 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).

[47] T. Stenfors, A. Kajamaa, and D. Bennett, How to … assess the quality of qualitative research, Clin. Teach. **17**, 596 (2020).

[49] T. O. B. Odden, H. Tyseng, J. T. Mjaaland, and M. F. Kreutzer, Physics education research conference proceedings literature review using text embeddings, https://zenodo.org/records/10702781 (2024).

[50] A. E. Pezalla, J. Pettigrew, and M. Miller-Day, Researching the researcher-as-instrument: An exercise in interviewer self-reflexivity, Qual. Res. **12**, 165 (2012), https://pmc.ncbi.nlm.nih.gov/articles/PMC4539962/.

[51] H. Tyseng, M. F. Kreutzer, J. T. Mjaaland, and T. O. B. Odden, Deductive analysis of PERC proceedings articles at scale using text embeddings, https://github.com/markusuio/G.E.V.I.R (2024).