How has *Science Education* changed over the last 100 years? An analysis using natural language processing

Tor Ole B. Odden | Alessandro Marin | John L. Rudolph

1Department of Physics, Center for Computing in Science Education, University of Oslo, Oslo, Norway
2Department of Curriculum and Instruction, University of Wisconsin–Madison, Madison, Wisconsin, USA

Correspondence
Tor Ole B. Odden, Department of Physics, Center for Computing in Science Education, University of Oslo, Sem Sælands vei 24, 0371 Oslo, Norway.
Email: t.o.odden@fys.uio.no

Abstract
For well over a century, the journal *Science Education* has been publishing articles about the teaching and learning of science. These articles represent more than just a repository of past work: they have the potential to offer insights into both the history of science education as well as the dynamics of field-specific change. It can be difficult, however, for educators, researchers, reformers, and policymakers to grasp the nuances of over 100 years of scholarship given the overwhelming amount of textual material. To address this problem, we have used latent Dirichlet allocation, an automated machine-learning algorithm from the field of natural language processing, to perform an automated literature review and classification of the corpus of work in *Science Education*. Using this technique, we have classified research in the journal into 21 distinct topics, falling into three thematic groups: science content topics, teaching-focused topics, and student-focused topics. We have also quantified the rise and fall of these topics and groups over time, and used them to begin to extract insight into the development of the field, including the effects of national policy changes on topics of interest to the research community, the interrelationships
between different research topics, and the effects of intellectual cross-pollination. Based on this analysis, we argue that this technique shows great promise for even larger-scale analyses of educational literature and other textual data.

KEYWORDS
history of science education, literature review, machine learning, methods, natural language processing

1 | INTRODUCTION

It would be an understatement to say that the field of science education has dramatically changed over the last 100 years. Through wars and geopolitical events, reforms, epistemological, theoretical, and methodological revolutions, and constant work by an engaged community, science education today looks much different than it did in 1920. These shifts are not limited to the ways we do science education—they also impact the ways we think about, talk about, and conceptualize what it means to do science education. At the same time, certain themes echo and recur throughout the history of the field, as they are consistently re-discovered and re-emphasized by generations of science education researchers and reformers (DeBoer, 1991; Rudolph, 2019).

As science education researchers, practitioners, historians, and reformers ourselves, we are interested in understanding these themes and shifts. They can help us to answer important questions, such as

1. What are the key topics of interest throughout the history science education, and how and when did they emerge?
2. How long have educators been thinking about, working with, and discussing the topics we are concerned with today?
3. Have certain topics evolved, shifted, or seen renewed interest over the history of the field?

The historical traces of these topics—which would enable us to find answers to these questions—are obviously present in the science education literature. For example, the journal Science Education (originally published as General Science Quarterly) has been serving the research community continually since 1916. This long history, however, has produced a massive literature base, which presents a major challenge for answering these types of questions. Thousands of articles have been published during the last 100 years, and although we can do targeted literature reviews, these by nature focus on just a few specific keywords or concepts and do not encompass the breadth or history of what has been written. Simply stated, it would be difficult—perhaps even intractable—for a researcher or a team of researchers to try to read and digest this entire corpus of research.

In recent years, new techniques have been developed for automated analysis of text that hold promise in meeting this challenge. These techniques, which come from the field of natural language processing, allow us to use algorithms to do what humans cannot—process and analyze thousands of articles at a time to extract key themes and quantifiably track their development in the literature.

In this article, we report on this type of analysis as applied to one repository of the science education research literature base: research articles published in Science Education. We have used a machine-learning technique called latent Dirichlet allocation (LDA) to analyze nearly the entire literature corpus of the journal, from its first few years to the present day. This technique allows us to automatically and independently extract topics and themes from
the published articles and to gauge their prevalence in the literature as a function of time, thereby giving us insight into how various topics of interest have fared over the history of the journal. This analysis, in turn, allows us to begin to discern large-scale, aggregate trends in science education research that can be useful to a variety of audiences, including, among others, established researchers, beginning researchers, and those researchers seeking greater historical understanding of how science education research has changed over time.

The purpose of this paper is twofold. First, we present the LDA method, describing both the technique itself and our process for collecting, cleaning, and analyzing the text used in this study. Second, we present the topics derived by the analysis, which function as a kind of automated, meta-level literature review of the articles published in Science Education since 1920. Moreover, we group these topics into thematic clusters and use the built-in functionality of LDA to quantify the rise and fall of different topics over time. From these trends, we can begin to see shifts in the research priorities of the journal, determine the dominant research topics during different eras, and make visible interesting trends that merit further investigation.

2 | TOPIC ANALYSIS USING LDA

For this analysis, we are using a technique called Latent Dirichlet Allocation, or LDA (Blei et al., 2003; Hoffman et al., 2010). LDA comes out of the field of natural language processing, a subfield of machine learning that focuses on the recognition and analysis of language and text. LDA is an unsupervised machine learning technique, meaning that it aims to learn or discover patterns in the data with a minimal amount of human supervision. In other words, when using this technique, we do not label any of the data or specify possible outcomes in advance (as one would do, e.g., with algorithms for automated classification). Instead, we input certain parameters, run the algorithm without direct supervision, and then see and interpret the trends that were extracted.

More specifically, LDA is one of a number of different techniques for doing automated topic analysis (also known as topic modeling), wherein an algorithm aims to find latent sets of topics within a set of texts. Underlying this technique is the assumption that every analyzed text contains a discrete set of topics of which some or all of the writers were aware, but that these topics are not explicitly stated—rather, they need to be inferred by a reader (or algorithm) and are, in this sense, "latent."

Theoretically, LDA relies on a key assumption about language and writing: It assumes that when authors use a similar set of words in a similar way they are most likely writing about a similar topic. Thus, to find topics, the algorithm looks for groups or clusters of words that frequently occur together. For example, if an author frequently used the words "dog," "kennel," “bark,” "leash," and “park,” it indicates that their writing was likely at least partially on the topic of "dogs." Notably, however, the technique assumes that all documents are a percentage mix of topics. So, for example, one article might be 95% about dogs (as one would expect to find in a magazine like Dog Fancy), while another might be 50/50 dogs and outdoor life (as one might expect from a dogsledding magazine), and another might be <1% dogs (e.g., most articles in Science Education).

To infer these topics from the words used, LDA relies on another simplifying assumption: it assumes that each topic can be modeled as a statistical distribution of words. Stated differently, each “topic” can be thought of as a collection of words and associated probabilities, with certain words being more likely to be chosen and certain words less likely to be chosen when one is writing about that particular topic. So, for example, a biology-focused topic might assign a high probability to words like "biology" and “cell,” meaning that when an author writes about this topic they will tend to use those words frequently. In contrast, words that are less central to the content of the topic (e.g., physics-focused words like “acceleration” or “torque”) would have lower probability, meaning that they would be used less often when writing about this topic. However, importantly, LDA assumes that every topic does include every word, just with different probabilities—for example, biology articles could still use the word “torque,” especially when discussing the musculoskeletal system, but they would tend to use these words less frequently on average.
To extract these topics and their associated statistical distributions, LDA requires researchers to make a third simplifying assumption: within the analyzed texts, the order of words is disregarded, and the only factor that the algorithm takes into account is the number of times each word occurs within each document. This assumption, known as the “bag of words” approach, is common in the field of natural language processing (Zhang et al., 2010). Although little can be learned from a single document based purely on the number of word occurrences, when one has a large number of documents meaningful patterns begin to emerge.

Based on these assumptions, the LDA algorithm essentially tries to infer these statistical word distributions out of the documents it is trained on. That is, by looking at the ways words cluster or occur together across a large number of documents, it essentially tries to tease out patterns in word usage and turn these patterns into statistical word distributions. Once these distributions have been created, the algorithm can then use them to quantify how much of each topic is present within each document in the data set, based on how well the words in that document align with each topic. It is then the job of the researcher to extract meaning from these word distributions by examining the most highly-rated words—that is, recognizing that a topic which centers on the words “biology” and “cell” is likely focusing on the biological sciences—and by examining the documents that are most associated with each topic.

These simplifying assumptions introduce certain constraints into the use of LDA. Because the algorithm looks at statistical similarities, it performs better (becoming more stable and interpretable) the more documents one gives it. Typical data sets often include thousands or tens of thousands of documents. It also requires users to make certain analytic choices in the form of parameters given to the algorithm. Specifically, users must tell the algorithm how many topics it should look for, how mixed the topics will be (on average) in the documents, and how many passes through the data set it should make before stopping. As with other natural language processing techniques, data filtering is also important (Denny & Spirling, 2018). For example, “stop words” (words without a specific semantic meaning, such as “and,” “but,” and “the”) must be removed lest they take over topics and wash out more meaningful results.

As of writing, LDA has been developed and used for close to two decades, and has been applied to a variety of fields and applications. For example, researchers in political science have used it to extract the key themes and political priorities from large numbers of transcribed political speeches (Curran et al., 2018; Grimmer & Stewart, 2013) and to quantify the degree of polarization in political commentary (Balasubramanyan et al., 2012; Roberts et al., 2016). In the software industry, LDA has been used to analyze large software databases to try comprehend the functions of different parts of the code (Maskeri et al., 2008). In the field of journalism, LDA has been used to analyze the historical development of attitudes toward nuclear power, based on articles published in the New York times (Jacobi et al., 2016). LDA was also applied to create a recommendation engine for news articles in the New York Times (Spangher, 2015), which recommends new articles to readers based on their reading history. Within the academic sphere, LDA has been used to identify “hot topics” in scientific research (Griffiths & Steyvers, 2004) and physics education research (Odden, Marin, & Caballero, 2020), as well as map the recent historical development of the field of physics education research (Odden, Marin, & Caballero, 2020).

We see LDA as a promising method for analyzing sets of literature far too large for traditional literature reviews. Using LDA, we can perform a kind of replicable, automated literature review—replicable, in that anyone who runs our code will obtain the same results. The topics produced by this kind of analysis can then provide novel groupings and overviews of the literature. Furthermore, the fact that the algorithm can quantify how much of each topic is present in each paper allows us to plot and examine the rise and fall of topics over time. That is, we can see when topics started to emerge in the literature, when they crested, and when they faded away (if ever). We can also see which topics were most popular during different time periods, and when the current “hot topics” originally emerged.

1Note that this is actually an iterative process, where the algorithm first creates a preliminary word distribution for each topic, then uses it to estimate the topic percentage in each document, then uses that to update the word distribution for each topic, and so on until a certain threshold of “accuracy” is reached. Readers who are interested in a more detailed description of how the technique works are encouraged to consult Hoffman et al. (2010) and Odden, Marin, and Caballero (2020).
Given the details of this technique, our research questions are as follows:

1. What are the primary topics addressed in the science education research literature during the last century as published in the journal *Science Education*?
2. How has the prevalence of these topics changed over time, and what historical events might account for these changes?

## 3 | TEXT ANALYSIS VERSUS HISTORICAL RESEARCH

There are important differences between the textual analysis we propose here and more traditional historical research and analysis that should be acknowledged. To be clear, a text-analysis approach such as that presented here is in no way meant to serve as an alternative to substantive historical scholarship. Both, of course, seek to learn about the past through the examination of textual artifacts, and both approaches look to establish patterns in the data that serve as the raw material for subsequent analysis and interpretation. The difference lies in who does what and when. Although there are a variety of ways in which historians practice their craft, most begin with deep and careful reading of the secondary literature (i.e., the histories that have already been written related to the topic in question). From this reading a question of significance emerges that then guides the search for additional, relevant secondary scholarship and, most importantly, the primary source material (e.g., archival documents, published work from the time period in question, oral histories, etc.). These sources then enable the researcher to begin to understand, refine, and answer the question originally posed (Rudolph, 2015; see also Rudolph, 2008).

In contrast, the topic analysis described here begins with the assumption that changes have occurred in the textual corpus of a journal like *Science Education* over time and that these changes might be of interest to present-day researchers. The LDA algorithm processes the large repository of text, identifies topics, and then visually displays their frequencies over time. Thus, the goal of our method is to reveal patterns that might then prompt questions about their cause or shed light generally on the historical trends in science education research. Such patterns might be of interest to a variety of readers. Science education researchers embarking on a project in a new area might wish to identify past time periods when a similar topic was trending to see how that earlier work might be of value or can be built upon. Senior scholars in the field might find the pattern of past trends informative with respect to the overall direction of the field and use that knowledge to make recommendations for future productive work. Sociologists of knowledge might use the data to study the manner in which topic change and/or topic proliferation occurs in a field. And researchers interested in doing actual historical work might find that the trend patterns reveal interesting topics or specific time periods that merit more in-depth, contextualized study.

The key point we wish to make here is that LDA (or other text analysis techniques) should not be construed as an automated process for “doing history” or even as a technique designed to support only “historical” work; rather, it is simply an analytical method that provides a new way to sift through larger data sets of primary source material and present the results in a form that may lead to new questions and subsequent understandings of the past—for a wide range of audiences. Its contribution, perhaps obviously enough, is one of efficiency of processing and replicable analysis.

## 4 | METHODS

### 4.1 | Data collection and filtering

Based on our goal of analyzing the science education research literature over time, we needed a circumscribed literature base that covered an extended number of years. In this regard, articles from *Science Education* are an ideal data source since the journal has been publishing since 1916. Although the scope of the journal is significantly
differently today than during the early years (as is to be expected), it provides an excellent instance of an unbroken chain of publication through time.

The first step in the analysis was to collect copies of all articles published in *Science Education*. These were downloaded in PDF format in December of 2019 using Wiley's CrossRef API, in accordance with Wiley's Text and Data Mining Agreement (Wiley Online Library, 2019). We also downloaded metadata files for each article, which included information like publication year, author names, citations, and references. In the end, this resulted in nearly 11,000 PDFs. However, many of these were irrelevant to our research aims as they included a large number of short book reviews and advertisements published during the course of the journal's run. Additionally, the PDFs for volumes 5 and 6 of the journal (1920–1921) were unavailable, and the text from articles before 1920 was heavily corrupted. As a result, we chose to restrict our analysis to articles from 1922 to 2019.

We scraped the text from these PDFs into a single datafile and performed several data-cleaning steps on it. First, we removed all articles that either did not list an author in their metadata files or were listed as duplicates; these documents appeared to be advertisements or short book reviews, which we assumed would contribute little to our research goals. This substantially reduced the number of articles to be analyzed, leaving us with approximately 5577 authored articles in the final data set. Next, we removed duplicated text from each article. This is an important step since before 1969, articles in the journal were published in a magazine format in which the end of one article and the beginning of the next would share the same page. Processing that overlap text as belonging to both articles would introduce a bias in the model. For this reason, we developed an automated detection of article beginnings and endings that was able to remove any duplicate text. Next, we removed the reference sections of the articles, as the words in referenced paper titles tend to obscure interesting distinctions between articles, as well headings (in all caps) such as “ABSTRACT.” Next, we examined a subset of the processed text and reunited commonly occurring words that were separated due to line breaks, text recognition issues, or British versus American spellings (e.g. converting "per cent" to "percent"), removed all numbers, symbols, special characters, and punctuation, and lowercased all words. Next, we split each document into a list of individual words or "tokens," which were to be counted in the eventual "bag of words." From there, we removed all stop words, which are words without any semantic meaning on their own—"the," "in," "if," "and," "but," and so forth—and all single-letter words.

After this data cleaning was done, we performed the additional step of lemmatizing all words, reducing them to their more basic form. For example, the words test, tests, testing, and tested all share the same stem-word, "test" and all refer to the same phenomena. In their raw form, however, LDA would treat them as different words. By lemmatizing these words, we collapsed them into a single word ("test") so they would all be counted together. There is, however, a risk to lemmatization: in some cases, a word might carry a different meaning depending on its part of speech. For example, the word "test" has different meanings when used as a noun (e.g., a test) versus adjective (e.g., a testable hypothesis). So, to distinguish between such uses of the same word, we also included a part-of-speech tagger in our algorithm that would detect these differences and only aggregate and lemmatize words from the same part of speech (e.g., nouns vs. verbs).

We then performed an automated routine to detect and create bi-grams, sets of words that frequently co-occur and carry additional meaning together. These words were combined with an underscore, for example, "problem_solving" and "high_school."

Finally, we filtered the data set one last time, aiming to remove words that were either rarely or commonly used. The filtering of rare words is a technique that is frequently used in natural language processing to speed up computation time by reducing the amount of data used in the analysis (Denny & Spirling, 2018). Based on our prior experience using LDA for this type of analysis (Odden, Marin, & Caballero, 2020), we set this threshold such that any words used fewer than 15 times across the entire data set were removed. This involved an assumption on our part that these words did not carry any specialized meaning and so were fine to drop from the topic analysis; however, this assumption seemed to be justified, given that the list of words dropped on inspection turned out to be primarily individual author names, typos, or text that had been misrecognized in PDFs.
Filtering for common words is also essential to screen out journal words that are invariant over time, such as "science," "student," "education," and "learn." (It is no great discovery to find that the journal Science Education has focused on students learning science over the past 100 years.) This technique is also frequently used in LDA-based research, where researchers often remove words appearing in anywhere from 25% to 99% of documents in their data set (Cvitanic et al., 2016; Denny & Spirling, 2018; Grimmer & Stewart, 2013; Hopkins & King, 2010; Jacobi et al., 2016; Larsen & Thorsrud, 2019; Nardello et al., 2018; Syed & Spruit, 2018). We note that this step must be handled carefully, as the words removed are likely to be quite central to certain topics of interest.

To determine the optimal threshold, we used a combination of two methods. First, we iterated through different threshold values and inspected the list of removed words, seeing when the words went from being mostly rhetorical ("give," "need," "time," "work," etc.) or associated with education at the broadest level ("student," "teacher," "school," "college," "learn," etc.) to words associated with specific areas of research or methodologies ("child," "experiment," "measure," "represent," and "scientific"). Second, we performed some exploratory modeling at the different thresholds. Through this exploration, we found that looser thresholds (i.e. to say, more words left in) tended to produce topics that lacked specificity, whereas much harsher thresholds (i.e. to say, more removed) yielded topics that became difficult to interpret. 

In the end, we chose a threshold between these two extremes, with the goal that it would remove enough words to provide interesting results, but not so many words that it would make the topics difficult to interpret. Our chosen threshold was 50%, meaning that any word appearing in 50% or more of the articles in our corpus would be removed. This eliminated the following 133 words:

- ability
- able
- activity
- analysis
- appear
- approach
- area
- ask
- attempt
- base
- begin
- case
- certain
- change
- class
- classroom
- college
- come
- complete
- concept
- concern
- consider
- content
- course
- curriculum
- data
- design
- determine
- develop
- developed
- development
- difference
- different
- discussion
- education
- educational
- effect
- end
- evidence
- example
- experience
- fact
- follow
- form
- general
- give
- good
- great
- group
- help
- high
- high_school
- idea
- important
- include
- increase
- indicate
- individual
- information
- ing
- instruction
- involve
- know
- knowledge
- large
- lead
- learn
- level
- life
- like
- little
- major
- make
- material
- mean
- method
- nature
- necessary
- need
- new
- note
- number
- order
- particular
- place
- point
- possible
- practice
- present
- problem
- process
- program
- provide
- purpose
- question
- reason
- related
- report
- require
- research
- result
- school
- science
- second
- select
- set
- show
- similar
- situation
- state
- student
- study
- subject
- suggest
- support
- take
- teach
- teacher
- term
- test
- think
- time
- type
- understand
- university
- use
- value
- view
- way
- well
- work
- write
- year

When inspecting our final list, we note that some of the words that were removed might seem arbitrary when considering their surface meaning. It is important to remember that all of the words that were removed appeared in 50% or more of the documents across the entirety of the journal—roughly 2700 articles. So, for example, the words "high_school" and "college" were removed, while the word "elementary_school" was not. However, "high_school" appears in approximately 54% of the documents, and "college" in approximately 60%, while "elementary_school" appears in only 27%. We interpret these statistics to mean that "high_school" and "college" cover a large number of different areas of interest within the science education research literature—for example, high school biology, physics, and chemistry, high school attainment, and high school teacher preparation. Elementary_school, on the other hand, seems to appear more as a topic unto itself, as elementary school science is often a distinct subject within the curriculum, with specific content and teacher education models. Thus, we argue it makes sense that words like "elementary_school" would be included in our analysis, while words like "high_school" and "college" would be excluded.

2 For example, topics at the looser threshold tended to focus around the most general and common words in the journal: "student," "teacher," "education," and "class." An example topic, from a model where we only removed words that appeared in 90% of papers, is "teacher," "teach," "student," "classroom," "class," "learn," "school," "lesson," which, although recognizably about education, does not provide much insight into the literature. Topics produced at harsher thresholds (i.e., more words removed) quickly became granular to the point where they became difficult to interpret. However, in some cases this also led to topics that seemed to focus on specific influential authors. For example, a topic from a model where we removed all words appearing in more than 10% of papers was as follows: "piaget," "cognitive_structure," "proposition," "memory," "cluster," "conflict." This topic clearly focuses on Piagetian cognitive theory.
Theoretically, we also note that robust topics of research interest (i.e., those that are actually present in the literature) should in principle not be significantly affected by the removal of one or two common words. Because the LDA algorithm aims to identify clusters of co-occurring words, we expect that it should still identify a cluster even if a single word is removed from that cluster. For example, a topic that focuses on genetics, and features words like "genetics," "gene," "trait," "race," "genetic," "protein," "chromosome," and "inheritance" should still be identifiable even if the word "genetics" were to be removed since the remaining words will tend to co-occur with one another. In contrast, the inclusion of a single word (say "teacher" or "student") that appears across many areas of study can create problems, since it can cause the algorithm to lump together many different topics that happen to share an (overly general) word. Thus, we would argue that it is better to err on the side of removing words than leaving them in.

For illustration, in Figure 1, we show the top words and word counts before and after this filtering.

### 4.2 Extracting topics using LDA

With this data preparation and cleaning complete, we began the process of using LDA to extract the latent topics from the data set. Our primary tool for this was the open-source library called Gensim (Rehurek & Sojka, 2011), which is written in the Python programming language, along with the Natural Language Tool Kit (Bird et al., 2009), which was used for much of the data filtering.

In principle, this analysis is as easy as importing the data set and applying the LDA modeling function in Gensim to it. However, this quickly leads us to a problem: unsupervised topic modeling techniques like LDA are inherently probabilistic and unstable. In other words, each time you run the algorithm, it randomly initializes, and the resulting topics can change depending on this random state. For example, one run of the algorithm might produce a single topic that focuses on quantitative assessment of student outcomes, while in another run of the algorithm might produce two variations on that topic, where one focuses on attitudes and the other focuses on content knowledge. Individual topic models, therefore, are replicable, in that you will get the same results if you start from the same random state each time. However, there is no immediate way of determining which random state one should use.

Compounding this situation is the fact that it is not obvious, a priori, how many topics one should tell the model to look for. Intuitively, this question is like asking “how many different strands of research should we expect to find when we analyze the last 100 years of science education research literature?” There are benefits and drawbacks to different numbers of topics. A smaller number (e.g., 10) is more easily interpretable, but might hide or aggregate interesting distinctions in the literature (such as the differences between alternative assessment paradigms). Larger numbers of topics (e.g., 50) may capture these distinctions, but may produce results that are too unwieldy to interpret in a meaningful way.

One solution to this problem is to run models under different sets of conditions (number of topics, random initializations) and pick the single model that seems the most interesting or that tells a recognizable story—this is known as a “face validity” check. However, considering the huge number of possible combinations of topic numbers and initial conditions, this approach quickly bogs down in an avalanche of possibilities. So instead, we relied on an automated technique to help inform this choice, using a metric called topic coherence, which is commonly used to evaluate LDA models (Röder et al., 2015; Syed & Spruit, 2018). In essence, coherence is a measure of how well the topics produced by a model actually fit with chunks of text taken from the data set. The algorithm for calculating coherence crawls through each document, taking chunks of text roughly 100 words in length and checking how well they fit with different topics. If...
the fit is good, the topics are considered to be “coherent,” in that the words of the topic seem to occur in close proximity to one another. This results in a score from 0 to 1, with 1 being completely coherent (essentially impossible in practice) and 0 being completely random and incoherent. Typical published coherence values for analyses of disciplinary articles seem fall in the range of 0.4–0.6 (Odden, Marin, & Caballero, 2020; Syed & Spruit, 2018).

To apply this evaluation method, we repeatedly ran the LDA algorithm on our corpus with topic numbers ranging from 10 to 30. For each topic number we created 10 randomly initialized models, calculating the coherence

**FIGURE 1**  Plots of most common words (blue, left y axis) in the data set before filtering (a) and after filtering (b). Before filtering, the data set consisted of 200,752 words in total; after, it consists of 24,940 words. Each graph also shows the fraction of total documents containing that word (red, right y axis) [Color figure can be viewed at wileyonlinelibrary.com]
score for each and plotting these scores as a function of topic number. From this graph, we were able to use the elbow-method, in which we looked for a leveling-off point (or "elbow") of the graph, above which increasing the topic number would lead to diminishing returns in average topic coherence. In the end, we chose a topic number slightly above this point, which resulted in 23 topics.

Once we had chosen how many topics the algorithm should look for, we still had to decide which single model to choose (i.e., which random initialization to use). This is again an impactful decision, as the particular topics produced by the model can vary significantly depending on the initialization. In particular, we noted that of the 23 topics generated, 19 seemed fairly stable from run to run, while the remaining 4 tended to vary. To make this choice, we employed a technique from a previous study (Odden, Marin, & Caballero, 2020) in which we generated 400 different models and essentially "averaged" them using a clustering algorithm. We then chose the single model that was closest to this average. This model had a coherence score that was slightly higher than the average (0.52, as opposed to 0.5 on average). This finding served as an additional check that it was a reasonable model to use, although we also performed some additional checks on the model to make sure that its results seemed coherent and interpretable. These included a face-validity check, where we examined the topics to see if they made sense based on our understanding of the literature, and the generation of an intertopic distance map to examine the relative statistical distinctiveness of the topics.

4.3 Analysis and interpretation of model

Once this model was chosen, we used it to perform the following analyses. First, we selected and aggregated the top 10 most representative papers for each topic from across the entire literature base. These were the papers that the model had rated as containing the greatest percentage of the topic in question (in practice, 60%–98% was normal, depending on the topic). We used these papers, along with an inspection of the top words for each topic, to interpret the general meaning of each topic—that is, the kinds of research it seemed to encompass. Based on this meaning, we assigned each topic a label and description, which are shown in Tables 1–3.

Next, we plotted the average prevalence of each topic year by year. By prevalence, we mean the degree to which each topic was represented in the literature for that year. To calculate average prevalence, for each topic, we summed all of the percentage contributions from each paper published in a particular year. We then averaged this number, dividing it by the number of papers published that year, to control for variations in the number of articles published from year to year. This value thus allows us to see the degree to which different topics, on average, rose and fell in proportion to one another.

Because the data were noisy and had high year-to-year variation, we used a smoothing technique to dampen out the year-by-year shifts in the form of a 3-year rolling window that averages the prevalence values for each year with its two nearest neighbors. This has the effect of removing small, year-by-year jitters, while leaving larger-scale shifts intact and is aligned with our research goals.

Finally, we examined these topic descriptors and graphs of prevalence over time, and used them to construct narratives and hypotheses for how and why the different topics have risen and fallen. We report on these narratives and hypotheses in part B of our results.

---

5Because the details of this analysis become somewhat technical, we do not report them in detail here. Interested readers are encouraged to consult Odden, Marin, and Caballero (2020) for further details.

6The intertopic distance map was generated using the pyLDAvis python package (Sievert & Shirley, 2014), and shows a two-dimensional representation of the topics created through principal component analysis. This allows one to see if the topics are relatively evenly distributed or if they clump together with distinct outliers (which we have found often indicates a poor-quality model). We have included a PDF of our intertopic distance map in the online Supporting Information material.

7A copy of this list of top 10 papers for each topic is available in the online Supporting Information material.

8This quantity has also been called "attention" by Jacobi et al. (2016), referring to the percent attention to a particular topic by the literature as a whole per year.
For the purposes of transparency and replicability, we have posted our cleaned datafile and an annotated analysis script to an open-access data repository (Odden, Marin, & Rudolph, 2020). Readers interested in replicating our analysis or exploring other themes or trends in the data are encouraged to consult it.

5 | RESULTS

5.1 | Results Part A: Topics and general trends

We begin by presenting the topics derived by our model in Tables 1–3 and Figures 2–4. For each topic, we present the topic label that we have assigned it, the words themselves, our interpretation of the meaning of the topic, and a rough, qualitative description of when this topic was prevalent in the literature. The topic names and descriptions represent our interpretation of the topic’s meaning based on examination of both the top 20 words and the top 10 most representative papers. In cases where the prevalence values on papers tended to be low (never exceeding a threshold of 65%) we have noted this in our topic description, as it indicates that this topic is likely evenly spread across the literature base. Our commentary on the time period is based on examination of the prevalence graphs in Figures 2–4, as well as the publication dates of the top 10 papers for each topic.

Note that we have dropped two of the original topics produced by our model as they primarily focused on journal business and were mostly irrelevant to our research questions9: one was focused predominantly on book reviews and advertisements, while the other captured largely professional matters like meeting notes, award announcements, biographies, and obituaries. The remaining 21 topics have been numbered and gathered into three thematic clusters based on their content:

1. Topics that focus primarily on particular types of science content: Topics 1–5; Table 1 and Figure 2.
2. Topics that are primarily teaching-focused: teaching strategies, teacher preparation, curriculum design, educational philosophies, and science outreach: Topics 6–13; Table 2 and Figure 3.
3. Topics that are primarily student-focused: student interest, motivation, cognition, understanding, and identities: Topics 14–21; Table 3 and Figure 4.

For each table, we have organized the topics in chronological order based on their overall presence in the literature from the early years of the journal up to the present day.

The graphs in Figures 2–4 show the average prevalence values for each topic as a function of time. For each graph, we have applied the rolling 3-year window described above to smooth out small, year-to-year variations. In addition, the shaded regions represent a margin of error on the prevalence values from each topic, which were calculated using a jackknife resampling technique (Efron & Stein, 1981). With this technique, for each topic in each year we recalculated the prevalence values multiple times, leaving out a random subsample of 20% of the data each time. These calculations were done in batches of five, such that every data point was left out once per batch. After normalizing for the missing 20% of the data, we repeated this procedure 20 times to obtain $5 \times 20 = 100$ prevalence values. The shaded regions have as width three SDs of these 100 recalculated prevalence values. Thus, it represents a confidence interval within which prevalence values will almost certainly fall even if the algorithm were to misclassify a number of documents.

Using these topical groupings and graphs of prevalence over time, we make the following observations: First, the science content topics in Table 1 have had relatively uniform prevalence over time, with the exception of

---

9However, even though we have chosen to ignore these topics, it is encouraging to see that the algorithm was able to recognize and distinguish them from the remainder of the literature.
Science in Everyday Life (Topic 1). We interpret this to mean that most of these topics have garnered consistent levels of research interest throughout the life of the journal. This would make sense since subject-matter topics are typically used as the context for other studies but are not themselves necessarily the main topic of interest. The initial prevalence (and rapid decline) of Topic 1 seems to reflect a shift in publishing priorities: in the early years the

| Topic name and number | Top 20 words | Topic interpretation | Time period |
|-----------------------|--------------|----------------------|-------------|
| 1. Science in Everyday Life | man, plant, world, day, thing, food, home, people, men, water, small, long, say, animal, control, soil, go, air, old, machine | Topic 1 seems to focus on early attempts at science outreach through written descriptions of natural and mechanical phenomena: for example, the life cycles of trees and insects, the development of the match, and the laying of undersea cables | Heavy concentration of high-prevalence papers in the 1920s and 1930s. Prevalence graph shows strong decay in interest during early years of the journal |
| 2. Earth and Space Science | earth, energy, evolution, water, rock, moon, sun, topic, astronomy, geology, planet, space, weather, fossil, tree, ecosystem, natural, cause, globe, sample | Topic 2 seems to be about Earth and space science (which is explicitly named in some high-prevalence papers) with a specific focus on geology, evolution, climate, and space. No individual article exceeds 62% prevalence in this topic | High-prevalence papers are spread across the period of 1950–present. Prevalence graph shows uniform, low levels of interest over time |
| 3. Physical Science | force, water, energy, substance, heat, matter, weight, chemical, atom, air, object, answer, gas, particle, temperature, mass, volume, element, property, electron | Topic 3 seems to focus on basic physical science (especially physics and chemistry) and related phenomena such as gas, chemicals, heat, force, and electricity | High-prevalence papers are spread out over the entire publication history, but concentrate from 1950–1980. Prevalence graph shows uniform, low levels of interest over time, with small bumps in the early years and 1990s |
| 4. Biological Science | biology, biological, cell, genetics, plant, organism, gene, human, animal, food, population, trait, specie, evolution, genetic, body, disease, race, protein, cause | Topic 4 seems to be about the teaching and learning of biological science, with secondary focuses on biology students, textbooks, and teacher-training. No individual paper exceeds 60% prevalence in this topic | High-prevalence papers are spread fairly evenly across the entire publication history. Prevalence graph also shows uniform interest over time |
| 5. Light and Observation | light, experiment, water, observation, color, explanation, go, observe, answer, look, plant, unit, try, notice, investigation, figure, table, day, object, explain | Topic 5 seems to be about children's and students' observations of light and shadow, although lower-probability words also cover phenomena like plants and water. The topic also covers young students learning science through observation and inquiry | High-prevalence papers are spread across the entire publication history. Prevalence graph also shows uniform interest over time |

Science in Everyday Life (Topic 1). We interpret this to mean that most of these topics have garnered consistent levels of research interest throughout the life of the journal. This would make sense since subject-matter topics are typically used as the context for other studies but are not themselves necessarily the main topic of interest. The initial prevalence (and rapid decline) of Topic 1 seems to reflect a shift in publishing priorities: in the early years the
journal published numerous “science stories” in the form of dialogues, plays, or descriptions of scientific phenomena in everyday life that could be used as models of classroom instruction or lesson inspiration (e.g., Barbell, 1927; Tower, 1928; Wathen, 1927). As time went on, though, this genre of writing became less and less common.

Second, within the category of teaching-focused topics, most of the years before 1970 were dominated by discussion of general techniques and ideas for how to teach science and prepare science teachers (Topic 6). This topic is quite broad, and seems to cover discussions of how science teaching should respond to new developments in educational technology such as film (Blanc, 1953; Bullington, 1948), advances in more specialized science teacher preparation (Brown, 1948), and discussions of the purpose of science education within the wider educational landscape (Davis, 1924; Meister, 1948). We also see an early emphasis on assessments and standards for teacher training (Topic 7), often in the form of descriptions of science teacher preparation from programs and schools across the United States (e.g., Pella, 1958; Watson, 1941). After 1970, these two topics see a major downturn, while Topics 9–12 (Inquiry-based curriculum reform; Science-Technology-Society, Socioscientific and Cultural Issues; History and Philosophy of Science; and Teacher professional development) emerge and remain prevalent through the present day.

Third, within the category of student-focused topics, we see three waves of research: the first is an early focus on quantitative measures of student performance (Topic 15), which seems to represent the initial efforts of science education researchers to apply experimental and statistical methodologies from the sciences to the study of teaching and learning. In the 1970s, we see the rise of the next wave, which consists of two topics that rise and fall in sync with one another: science education research related to educational psychology (Topic 16), and quantitative studies of latent constructs such as student interest and motivation (Topic 17). Finally, in the 1980s and 1990s, we see the rise of research topics focused on students’ mental models, conceptual change, and fluency with representations of science knowledge (Topic 18), student use of argumentation and other scientific practices (Topic 19), and student identity and discourse (Topic 21). Interestingly, the topic that focuses on race, gender, and the STEM pipeline (Topic 20) straddles both the second and third waves of student-focused research, with noticeable upticks in both the mid 1980s and after 2010.

To provide an overview of how the journal as a whole has shifted, we plotted all the topics together in a stacked area plot, shown in Figure 5. In this plot, each shaded line represents one topic, with the width of the line corresponding to the prevalence of the topic in the literature during a particular year. To dampen out the year-to-year fluctuations, we used a smoothing technique to create a more continuous representation of the data.
| Topic name and number | Top 20 words | Topic interpretation | Time period |
|-----------------------|-------------|----------------------|-------------|
| 6. Science Teacher Preparation, Program Development, and Technology | pupil, principle, objective, unit, scientific, field, list, elementary, plan, interest, technique, subject_matter, grade, experiment, procedure, attitude, education_vol, topic, demonstration, elementary_school | Topic 6 seems to focus on early discussions (pre-1970s) of ways to broadly improve science education through new developments in educational technology and teacher training, along with broader discussion of the purpose and organization of science instruction | Most high-prevalence papers appear before 1955. Prevalence graph shows high interest in this topic before the 1970s |
| 7. Science Teacher Training and Content Preparation | biology, percent, training, physical, graduate, field, mathematics, survey, preparation, institution, total, hour, grade, taught, offer, table, senior, secondary_school, average, degree | Topic 7 seems to focus more narrowly on early assessment and descriptions of science teacher content preparation, as well as programmatic aspects of school science offerings | Strong concentration of high-prevalence papers from the 1940s to the 1960s. Prevalence graph shows moderate interest in the topic until the 1990s |
| 8. Physics Education and Laboratory Instruction | physic, chemistry, laboratory, experiment, lecture, examination, lab, topic, grade, instructor, textbook, traditional, practical, demonstration, emphasis, chemical, taught, experimental, syllabus, secondary_school | Topic 8 seems to be primarily about physics education, with a secondary focus on laboratory instruction. Other subtopics include student interest and motivation related to laboratory work. No individual articles exceed 52% prevalence in this topic | Strong concentration of high-prevalence papers in the 1970s. Prevalence graph shows uniform, low-level interest over time |
| 9. Inquiry-Based Instruction and Research Experiences | inquiry, project, standard, team, scientist, goal, skill, faculty, gain, evaluation, researcher, discipline, participant, outcome, professional, focus, undergraduate, effort, scientific, conduct | Topic 9 seems to be about inquiry-based curricula and standards. Secondary focus on training professional scientists through research experience for undergraduate students. No individual articles exceed 60% prevalence in this topic | High prevalence-papers cluster from around 2004 to present. Prevalence graph reflects this, but shown an additional wave of interest during the period of 1970–1985 |
| 10. Science Technology-Society, Socioscientific and Cultural Issues | social, technology, community, issue, society, people, culture, educator, public, cultural, human, world, country, environmental, policy, engineering, future, national, environment, local | Topic 10 seems to be about the subfield of STS (Science–Technology–Society), with a secondary focus on environmental and cultural issues | High-prevalence papers concentrate around 1980–2010. Prevalence graph shows steadily growing interest until around |
variations and reveal overall trends, we again applied a 3-year rolling window to each of the topic graphs. We additionally renormalized each of the trends to account for the smoothing and two neglected topics, uniformly rescaling the prevalence values for all graphs so that they add up to 100%. We then grouped the topics such that the science content topics are at the top, the teaching-focused topics are in the middle, and the student-focused topics are at the bottom. Using this grouping, it is possible to see both how the individual topics emerged and evolved relative to the rest of the literature in the journal and how each of these three groups of topics developed over time.

| Topic name and number | Top 20 words                                                                 | Topic interpretation                                                                 | Time period                                                                                     |
|-----------------------|------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| 11. History and Philosophy of Science | scientist, theory, scientific, world, belief, believe, claim, thought, law, position, hypothesis, sense, argument, statement, issue, thing, religion, history, truth, philosophy | Topic 11 seems to be about the history and philosophy of science, with a special focus on science and religion | 2000, after which it begins to decrease |
| 12. Teacher Professional Development and Science Teacher Education | lesson, belief, preservice_teacher, interview, elementary, participant, instructor, strategy, taught, focus, role, mentor, effective, professional_development, goal, topic, model, observation, opportunity, plan | Topic 12 seems to focus on training teachers, including both pre-service and in-service teachers. Secondary focus on teacher beliefs and interview methodologies | High-prevalence papers concentrate in the period of 1970–1995. Prevalence graph also shows a wave of interest around this time, which begins to decline after the 1990s |
| 13. Science Learning in Museums and Other Informal Settings | exhibit, museum, visitor, family, visit, center, parent, informal, pro_le, adult, engagement, interaction, behavior, trip, interest, game, exhibition, topic, et_al, motivation | Topic 13 focuses heavily on science museums and informal science centers, and the ways in which individuals and families experience them | High prevalence papers appear from the mid-1990s to the mid-2000s. Prevalence graph shows the emergence of the topic in the 1990s but low interest across the field as a whole |
From this graph, we can begin to discern larger-scale trends in this literature base. Roughly speaking, we see that the journal seems to have undergone several shifts: from a focus on science content topics during the early years, to teaching-focused topics in the middle years, to an emphasis on student-focused topics in the most recent years. More specifically, the early years of the journal were dominated by science content topics, especially Science in Everyday Life (Topic 1). By the mid-1930s, however, these topics had given way to an increased focus on teaching topics, especially Science Teacher Preparation, Program Development, and Technology (Topic 6). Then, in the 1970s, Science Education underwent a rapid transition (one might even call it a phase change) as student-focused topics rose to prominence, collectively taking up roughly half of the attention of the journal by the mid-1970s. Teaching-focused topics regained some of their prominence in the 1990s, but the emergence and rapid rise of sociocultural, student-focused topics such as Argumentation and Scientific Practices (Topic 19) and Identities and Discourse Analysis (Topic 21) since the late 1990s has again shifted the balance away from teaching and toward students.

From both the plots of individual topics in Figures 2–4 and the plot of all topics in Figure 5, we can see that over time, certain topics have risen in prominence relative to the rest of the literature, visibly dominating the prevalence graph in Figure 5 during certain eras. These include the following:

- Science in Everyday Life (Topic 1): dominant until the early 1930s.
- Science Teacher Preparation, Program Development, and Technology (Topic 6): dominant until the early 1970s.
- Educational Psychology and Problem-Solving (Topic 16): dominant until the mid 1980s.
- History and Philosophy of Science (Topic 11): dominant throughout the 1990s.
- Argumentation and Scientific Practices (Topic 19): dominant since the mid-2000s.
- Identities and Discourse Analysis (Topic 21): also dominant since the mid-2000s.

We have isolated these topics and plotted them together in Figure 6. Aggregated in this way, it again becomes clear that Science Education (the journal, and arguably also the field) has seen radical shifts in its research focus over the years. That is to say, the most prominent topics during the first half century (Topics 1 and 6) mostly faded away by the mid-1970s, while those that saw the greatest prevalence in the last quarter century were almost unknown before the 1990s. The exception to this trend is History and Philosophy of Science (Topic 11), which was relatively dominant in the 1990s but seems to have also been somewhat prevalent throughout both the early and late history of the journal. We suspect that this is related to the consistent attention to the nature of scientific thought, attitudes, and methods throughout the history of the journal, for example, Ebel (1938), Schaffner (1964), and Alchich (2004).
| Topic name and number | Top 20 words | Topic interpretation | Time period |
|-----------------------|--------------|----------------------|-------------|
| 14. Children's Ideas  | child, age, grade, object, parent, interview, elementary_school, year_old, response, adult, interest, young_child, thing, environment, old, piaget, elementary, young, animal, conservation | Topic 14 seems to focus on the development of scientific concepts and ideas in young children, influenced by the theories of conceptual change and Piagetian stages | High-prevalence papers are distributed across the publication history of the journal. Prevalence graph shows that the topic has seen uniform low-level low level over time, with a small uptick in interest around the mid-1950s, and another in the late 1970s |
| 15. Quantitative Measures of Student Performance | score, achievement, measure, significant, treatment, total, grade, table, low, comparison, control_group, hypothesis, average, control, gain, experimental, factor, significant_difference, sample, experimental_group | Topic 15 seems to focus on quantitative measures of student performance in science courses, with a secondary focus on statistical measures and applications of scientific experimental design to educational research | High-prevalence papers cluster in the period of 1957–1970. Prevalence graph shows high interest (with several small spikes) before the 1980s, after which it rapidly decreases |
| 16. Educational Psychology and Problem-Solving | task, problem_solve, skill, performance, behavior, strategy, cognitive, structure, learner, specific, solution, instructional, procedure, variable, unit, sequence, solve, control, pattern, rule | Topic 16 seems to focus on research related to educational psychology within science education. Specific focus on problem-solving, cognition, and student reasoning strategies and heuristics | High-prevalence papers strongly concentrate in the 1970s and early 1980s. Prevalence graph shows a significant (and distinctive) wave of interest from the mid-1960s to the late 1990s |
| 17. Quantitative Methods, and Assessment of Student Interest and Motivation | item, response, attitude, measure, scale, instrument, score, factor, category, statement, sample, questionnaire, table, assessment, construct, low, dimension, behavior, respondent, range | Topic 17 seems to focus on development and deployment of quantitative methods for studying a variety of latent student constructs, with a special focus on assessments of student interest and motivation for learning science | Most high-prevalence papers cluster in the period of 1970–1990. Prevalence graph shows a significant wave of interest from the mid-1960s to the late 1990s, with another smaller wave in recent years |
| 18. Cognitive Models and Representations | model, explanation, conception, representation, phenomenon, figure, conceptual, represent, explain, construct, structure, context, analogy, relationship, meaning, interview, rst, et_al, object, metaphor | Topic 18 seems to focus on students' cognitive models, conceptual change, and understanding of representations of scientific knowledge such as visual depictions and graphs. Secondary focus on differences in understanding between experts and novices | High-prevalence papers appear in period of 1990–present, with the majority appearing after 2000. Prevalence graph shows that the topic has seen a nearly linear growth in interest since its emergence in the 1970s |

(Continues)
This summary of the data addresses our first research question "What are the primary topics addressed in the science education research literature during the last century as published in the journal Science Education?" It also begins to address our second research question, “How has the prevalence of these topics changed over time, and what historical events might account for these changes?” To further answer this question, we now turn to a more detailed analysis of these trends, including a discussion of the possible inter-relationships between different topics and explanations for some of the shifts seen above.

5.2 Results Part B: Specific trends and topic relationships

One interesting aspect of this analysis is its potential to uncover large-scale trends and shifts in the literature over time that would be difficult to see using more traditional methods such as targeted literature reviews. For example, the effects of science education policy changes, geopolitical events, or changes in editorial priority likely unfold over the course of years, if not decades. Similarly, shifts in theoretical framing, research interest, and research
methodology also operate on equally long timescales. There are numerous such shifts visible in the graphs shown above. How might these be explained?

In this section, we delineate and unpack some interesting trends that we have noticed, along with our hypotheses for how and why they may have emerged. Fully developing robust historical explanations would require a degree of analysis that is beyond the scope of this article. However, this represents a promising area of future work.

With regard to the shifts in topical prevalence over time visible in Figure 5, we suspect that they emerge in response to two kinds of factors. First, certain topics seem to spike in response to changes in science education policy as advanced in various national reports. For example, gender, race, and STEM pipeline (Topic 20) has two visible waves of interest as seen in Figure 4: one in the 1980s and another since around 2005. As this topic is related to the recruitment of scientists (via the so-called STEM pipeline), we suspect the early wave may have been a response to the report A Nation at Risk (The National Commission on Excellence in Education, 1983), which called for increasing rigor in academic subjects such as science in a desire to foster technological innovation in hopes of greater economic development. The second wave may have been related to the Carnegie-funded The Opportunity Equation (Commission on Mathematics and Science Education, 2009) which, as part of a series of reports at the time (see, e.g, Committee on Prospering in the Global Economy of the 21st Century, 2007), similarly called for a renewed emphasis on science education for meeting economic challenges.

A second factor that seems to result in observable topical shifts comes from rapid developments in science education-adjacent fields or from the development of new fields entirely. The rapid growth of the field of cognitive psychology in the 1960s and 1970s (Cohen-Cole, 2014), for example, appears to have stimulated research in science education that drew upon some of that field’s key concepts and methodological techniques. This perhaps explains the rise of the topic on Children’s Ideas (Topic 14) with its emphasis on Piagetian stages and conceptual change, and of Educational Psychology and Problem Solving (Topic 16). Clearly, one would expect such intellectual cross-pollination at the intersection of work on student cognition and a content-focused educational field like science education.

We also see several topics that, based on our knowledge of the science education research literature, seem related to one another. When we superimpose plots of their prevalence over time, we can begin to see other explanations, and construct possible narratives, for how the field has developed. We present three of these narratives here.
FIGURE 5  Stacked area plot of all topics, grouped according to their thematic clusters: science content topics (red, top), teaching-focused topics (green, middle), student-focused topics (blue, bottom) [Color figure can be viewed at wileyonlinelibrary.com]
5.2.1 | Shifting standards and goals for science teacher preparation (Topics 7 and 12)

Two of the topics from the teaching-focused cluster seem to have an especially strong relationship to one another: Science Teacher Training and Content Preparation (Topic 7), and Teacher Professional Development and Science Teacher Education (Topic 12). Plotting these three topics together (Figure 7), we can see that the journal had an early focus on Topic 7, which is primarily concerned with descriptions of science-teacher content preparation, assessment of teacher training and knowledge, and programmatic aspects of science courses. Although Topic 7 rises and falls several times over the years, it remains fairly prevalent until the late 1960s, when it begins to drop away. However, right at that point, Teacher Professional Development (Topic 12) spikes in prevalence. In contrast to Topic 7, Topic 12 is primary concerned
with teachers’ beliefs, practices, and mentorship and encompasses both pre-service teacher preparation and in-service teacher professional development. By superimposing the two, we can see that there is a period of overlap, from the early 1970s to the mid 1990s, after which Topic 7 and Topic 12 dramatically diverge. We suspect the Topic 12 spike in the late 1990s may have been precipitated by the 1996 publication of the National Science Education Standards (National Research Council, 1996) that strongly focused on the professional development of science teachers.

5.2.2 | The “practice turn” (Topics 11, 18, and 19)

Several of the topics revealed by our analysis focus on describing and applying scientific practices in education. Specifically, History and Philosophy of Science (Topic 11) concerns itself in large part with unpacking the scientific endeavor and applying those findings to science education. This topic saw a major uptick in the mid 1990s, rising simultaneously with two other, related topics: Topic 18, which focuses on students’ cognitive models and their use of different representations of science knowledge, and Topic 19, which focuses on student engagement in authentic scientific practices such as argumentation. We have plotted these three topics together in Figure 8.

We hypothesize that these trends are related. In looking at the trend lines, it seems likely that the research on History and Philosophy of Science laid the groundwork for the emergence of the other two topics by delineating and articulating the particular practices used by scientists. These results were then employed in research on student learning about and engagement with those practices.

5.2.3 | The shift from cognitivist to sociocultural frameworks (Topics 16, 17, 18, and 21)

We can also see theoretical and methodological shifts related to cognitivist and sociocultural framework topical trends. Specifically, one can see a shift in the field from a primarily quantitative and cognitivist focus to a more qualitative, sociocultural focus by examining the development over time of Educational Psychology and Problem-Solving (Topic 16);
Quantitative Methods, and Assessment of Student Interest and Motivation (Topic 17); Cognitive Models and Representations (Topic 18); Identities and Discourse Analysis (Topic 21). We have plotted these topics together in Figure 9.

As can be seen from the figure, the 1970s heralded a wave of interest in research on student psychology and cognition (Topics 16 and 17), which was tied to a focus on quantitative research methodology. This wave appears to have lasted until the early 1990s. However, as those topics began to fade in prevalence, we see the rise of Topic 18, which (due to its incorporation of both cognitive models and practice-based conceptualizations of science learning) represents a hybrid between cognitivism and sociocultural/practice-based views of teaching and learning. Simultaneously, we see the steep rise of Topic 21, which is an explicitly sociocultural topic that focuses on student identities and discourse analysis across different communities. This rise seems to represent a shift in priorities in the journal toward more qualitative methods, such as discourse analysis, and a greater emphasis on sociocultural views of teaching and learning.

6 | DISCUSSION

What can we take away from these results? First, we argue that this study provides a methodological contribution. Techniques like LDA, we propose, can be used to analyze the literature at a scale that is challenging to do otherwise. That is, we have demonstrated the ability to effectively perform a replicable, automated, thematic literature review of all articles published in *Science Education*, something that would be difficult for any individual or even team to do by hand, as it requires digesting and categorizing a huge amount of literature. While many readers (especially longstanding members of the science education community) may recognize the most recent trends from personal experience, and readers familiar with science education history may recognize the longer-term trends, we have been able to go one step further in our analysis to quantify these trends as a function of time.

Second, through visual mapping, this analysis reveals trends and shifts over time across the various topics—including interesting microlevel trends and connections that would otherwise be obscured by the large numbers of documents and expansive timescales involved. For example, our model shows two separate waves of research on teacher preparation: one that focused primarily on assessments and standards for teacher training (which was
dominant during the first 50 years of the journal) and another that focused on teacher professional development, pre-service teacher education, and their effects on teacher beliefs, which held sway during the second 50 years. Given the long timescales involved, it would be difficult for an average reader to distinguish and track such trends, much less quantify them as a function of time. Similarly, our analysis suggests that in the 1970s there were two simultaneous waves of student-focused research: one that focused on problem solving and heuristics and another centered on student interest and motivation. Given their occurrence at the same time, it might be hard for readers to distinguish between them, but the algorithm is able to differentiate them based on the differences in language used across the articles.

We note that LDA is ideally suited for this kind of analysis as it assumes that each document is a percentage mix of topics. This means that LDA is able to tease out the emergence of new topics, even when the topic has low overall prevalence and is distributed across many different articles in a year. This kind of sensitivity, inherent to the statistical nature of the technique, would be difficult for a human reader to replicate.

A third takeaway from this analysis is that our field seems to have benefitted from periods of intellectual “cross-pollination” with adjacent fields. For example, the topics on Quantitative Methods, and Assessment of Student Interest and Motivation (Topic 17) and Educational Psychology and Problem-Solving (Topic 16) both seem to have emerged from ideas originating in the fields of educational psychology and general psychology. The topics on Science–Technology–Society, Socioscientific and Cultural Issues (Topic 10) and History and Philosophy of Science (Topic 11) also appear to have come out of adjacent but different fields. And the more recent sociocultural focus has come, in part, out of the emergence of critical theories and theories of sociocultural learning brought over from sociology and the learning sciences.

These results suggest that future waves of research might also emerge in response to similar cases of cross-pollination. However, an additional takeaway is that researchers looking to have a strong impact on the field would do well to familiarize themselves with literature outside of science education as they pursue their work. Selectively borrowing concepts and ideas from other fields seems to be a productive way of improving research in our field and can also be a way for individual researchers to distinguish themselves and possibly contribute to large-scale shifts in topical focus. More work remains to be done, however, to better determine and track the between-field interactions that are suggested here.

Overall, the usefulness of this analysis, we propose, lies primarily in what it can tell us about how our field has developed historically. Generally speaking, we can see a large-scale movement in the literature: according to this analysis Science Education started with a focus on quantitative studies of student outcomes and teacher preparation paired with general discussion of science education philosophy, technology, and techniques, before about 1970. This gave way to a focus on more latent variables like student cognition, interest, and motivation, along with increased interest in inquiry-based teaching, and a focus on science education history, philosophy, and socioscientific issues, from the 1970s to the 1990s. From there, we see a shift toward scientific practices, with a growing interest in student discourse and argumentation; student conceptual understanding and use of representations; student identities; as well as teacher professional development, pre-service teacher education, and their effects on beliefs, interests, and motivations.

This type of mapping of the development of the journal may be especially useful for researchers just entering the field of science education. As previously discussed, one difficulty with a research field as mature as science education is that there is an overwhelming amount of previously published literature, much of which may be relevant to current scholarship. When entering a field like this, there is an ever-present danger of recapitulating known results—“reinventing the wheel,” in other words. Analyses such as the one presented above may help new scholars to situate their research foci within the greater literature base and determine whether they are addressing an enduring theme, and if so, when in the history of the journal they might find other scholars addressing similar subjects. One of the true hallmarks of real scholarship is the careful situation of new work within the existing research of a field. For example, based on the results presented above, researchers interested in applying psychological constructs to science education may find it productive to
explore the literature published between 1970 and 1990. Conversely, this analysis also points out that certain topics genuinely are new in the literature, indicating that scholars studying these topics could safely restrict their focus to the last two or three decades of published literature. Additionally, this proof-of-concept suggests that journals such as *Science Education* might benefit from an LDA-based recommendation engine, similar to that developed for the New York Times (Spangher, 2015), which would help scholars discover work related to their interests.

Based on the trends for the most recent years, we note that certain contemporary topics appear to be experiencing a massive wave of interest: for example, Gender, Race, and the STEM Pipeline (Topic 20) and Identities and Discourse Analysis (Topic 21) both have experienced a dramatic increase in the last few years. With this kind of intellectual momentum (i.e., the fact that trends take some time to get going, but then keep going for years or decades), it seems reasonable to think that these topics will remain of keen interest to researchers for the next few years at least, if not longer—especially given the current sociopolitical situation in the United States.

Lastly, although not the primary focus of this article, we also wish to note the potential usefulness of this technique for analyzing additional articles and bodies of literature beyond those used to train the model. Once a model is trained, LDA provides functionality to evaluate other documents and quantify the prevalence of topics within them. So far, we have only performed limited tests of this functionality, but the results are encouraging. For example, our model characterized the article “Invented Science: A Framework for Discussing a Persistent Problem of Practice” (Russ & Berland, 2018), published in the *Journal of the Learning Sciences*, as primarily a mixture of Cognitive Models and Representations (24%), Argumentation and Scientific Practices (18%), History and Philosophy of Science (16%), and Identities and Discourse Analysis (11%). Since this article presents a framework for science learning that emphasizes individual knowledge construction and is illustrated with excerpts of student discourse, these results seem to make sense. Similarly, the article “Science for What Public? Addressing Equity in American Science Museums and Science Centers” (Feinstein & Meshoulam, 2014), published in the *Journal of Research in Science Teaching*, was evaluated as a fairly even mixture of Science–Technology–Society, Socioscientific and Cultural Issues (27%), Identities and Discourse Analysis (24%), Science Learning in Museums and Other Informal Settings (15%), and Inquiry-Based Instruction and Research Experiences (13%). Considering this article focuses on the role of science museums and science centers in society, the results also seem to have face validity.

7 | CONCLUSION

We have used LDA to perform an automated, quantitative literature review and categorization for all authored articles from *Science Education* over approximately a century’s worth of publication history. Using this analysis, we have been able to detect and describe several enduring threads of research in the journal, as well as various waves of interest and evidence that our field has productively borrowed from other fields.

Although we feel these results are compelling, we wish to be clear about the limitations of this analysis. First, these results need to be understood as a model of this body of literature and interpreted as such—they are not objectively “true,” but rather emerge from a combination of the literature base used, the particular mathematical model of text and algorithm employed in the analysis, and the data cleaning, filtering, and modeling decisions made by the authors. Thus, although this technique can be used to augment our understanding of the literature (Grimmer & Stewart, 2013), we are not suggesting that it replaces the need for human interpretation.

In accordance with this understanding, it should also be acknowledged that these trends represent the analysis and predictions of one kind of topic model, from one specific journal. Although it may show particular topics as increasing (or declining), this says nothing about the value of studying those topics. We wish to emphasize that the algorithm looks at the literature writ large and is silent about the merits of individual articles, studies, or research foci.
With these caveats in mind, we believe that this technique shows promise for future exploration of the science education research field. A natural extension of this study would be to apply a similar analysis to a data set composed of articles from multiple science education research-focused journals to account for different publication histories, editorial priorities, geographical settings, and so on, and simply to capture a wider variety of research. Such a study might contribute to an overall meta-analysis of the field of science education as a whole.

This technique can also be naturally used to analyze the historical trends within other fields of scholarship. For example, we have performed an initial study in this realm by analyzing conference proceedings from the field of Physics Education Research (Odden, Marin, & Caballero, 2020). There is a rich potential line of work applying similar techniques to other subfields and trying to tease out the relationships within them. And, beyond the research-focused literature, there are interesting possibilities in analyzing work from practitioner journals to see how writing about the teaching and learning of science (rather than research on that teaching and learning) has changed over time.

Finally, it should be noted that the technique is equally applicable to other forms of text—qualitative research data such as student written responses, for example. In that context, this kind of automatic classification could be used to augment more ordinary forms of thematic qualitative analysis. Although we have yet to try LDA on this type of data, it seems like a promising area of future work. Some scholars have already made initial ventures in this direction (see, e.g., Sherin, 2013), and LDA shows particular promise for the kind of large-scale automatic classification that many researchers are interested in.

ACKNOWLEDGMENTS
This study was funded by Norges Forskningsråd Project No. 288125 “International Partnership for Computing in Science Education” and by the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education (DIKU) which supports the Center for Computing in Science Education. We thank Anders Malthe-Sørenssen, Christine Lindstrøm, Marieke Kuijjer, Danny Caballero, and John M. Aiken for their support and feedback on this project.

CONFLICT OF INTERESTS
The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.4094974.

ORCID
Tor Ole B. Odden https://orcid.org/0000-0003-1635-9491
Alessandro Marin https://orcid.org/0000-0002-5524-3451
John L. Rudolph https://orcid.org/0000-0001-9103-9653

REFERENCES
Allchin, D. (2004). Should the sociology of science be rated X? Science Education, 88(6), 934–946. https://doi.org/10.1002/sce.20026
Balasubramanyan, R., Cohen, W. W., Pierce, D., & Redlawsk, D. P. (2012). Modeling polarizing topics: When do different political communities respond differently to the same news. ICWSM 2012—Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (pp. 18–25). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4525/4962
Barbell, C. W. (1927). Laying the world’s fastest ocean cable. General Science Quarterly, 11(2), 118–124.
Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: Analyzing text with the natural language toolkit. O’Reilly Media, Inc.
Rehurek, R., & Sojka, P. (2011). Gensim—python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016). Navigating the local modes of big data: The case of topic models. In R. M. Alvarez (Ed.), Computational Social Science: Discovery and Prediction (pp. 51–97). Cambridge University Press. https://doi.org/10.1017/CBO9781131625734.004

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. WSDM 2015—Proceedings of the 8th ACM International Conference on Web Search and Data Mining (pp. 399–408). https://doi.org/10.1145/2684822.2685324

Rudolph, J. L. (2008). Historical writing on science education: A view of the landscape. Studies in Science Education, 44(1), 63–82.

Rudolph, J. L. (2015). Scholarly norms for writing history. American Journal of Physics, 83(11), 911. https://doi.org/10.1119/1.4930894

Rudolph, J. L. (2019). How we teach science: What’s changed, and why it matters. Harvard University Press.

Russ, R. S., & Berland, L. K. (2018). Invented science: A framework for discussing a persistent problem of practice. Journal of the Learning Sciences, 28, 279–301. https://doi.org/10.1080/10508406.2018.1517354

Schaffner, K. F. (1964). Science: Metaphysic, method, and law. Science Education, 48(5), 436–441.

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. Journal of the Learning Sciences, 22(4), 600–638. https://doi.org/10.1080/10508406.2013.836654

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces (pp. 63–70). https://doi.org/10.3115/v1/w14-3110

Spangher, A. (2015). Building the Next New York Times recommendation engine. https://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine/

Syed, S., & Spruit, M. (2018). Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. In 2017 IEEE International conference on data science and advanced analytics (DSAA) (pp. 165–174). https://doi.org/10.1109/DSAA.2017.61

The National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. The Elementary School Journal, 84. https://doi.org/10.1086/461348

Tower, C. V. (1928). A war of the centuries. General Science Quarterly, 13(2), 101–111. https://doi.org/10.1002/sce.3730130208

Wathen, A. W. (1927). The cabin of sound. General Science Quarterly, 12(1), 347–349. https://doi.org/10.1002/sce.373012010857

Watson, D. R. (1941). The training and experience of instructors in survey courses in physical science. Science Education, 25(2), 80–84. https://doi.org/10.1002/sce.3730250206

Wiley Online Library. (2019). Text and Data Mining. https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics, 1(1), 43–52. https://doi.org/10.1007/s13042-010-0001-0

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Odden, T. O. B., Marin, A., & Rudolph, J. L. (2021). How has Science Education changed over the last 100 years? An analysis using natural language processing. Science Education, 105, 653–680. https://doi.org/10.1002/sce.21623