Text embedding models reveal high resolution insights into conceptual knowledge from short multiple choice quizzes

Paxton C. Fitzpatrick[1], Andrew C. Heusser[1, 2], and Jeremy R. Manning[1, *]

[1]Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

[2]Akili Interactive Labs

Boston, MA 02110, USA

[*]Corresponding author: Jeremy.R.Manning@Dartmouth.edu

Abstract

We develop a mathematical framework, based on natural language processing models, for tracking and characterizing the acquisition of conceptual knowledge. Our approach embeds each concept in a high dimensional representation space, where nearby coordinates reflect similar or related concepts. We test our approach using behavioral data from participants who answered small sets of multiple choice quiz questions interleaved between watching two course videos from the Khan Academy platform. We applied our framework to the videos' transcripts, and to text of the quiz questions, to quantify the content of each moment of video and each quiz question. We used these embeddings, along with participants' quiz responses, to track how the learners' knowledge changed after watching each video. Our findings show how a small set of quiz questions may be used to obtain rich and meaningful high resolution insights into what each learner knows, and how their knowledge changes over time as they learn.

Keywords: education, learning, knowledge, concepts, natural language processing

Introduction

Suppose that a teacher had access to a complete tangible "map" of everything a student knew. Defining what such a map might even look like, let alone how it might be constructed or filled in, is itself a non-trivial problem. But if a teacher *were* to gain access to such a map, how might it change their ability to teach that student? Perhaps they might start by checking how well the student knew the to-be-learned information already, or how much they knew about related concepts. For some students, they could potentially optimize their teaching efforts to maximize efficiency by focusing primarily on not-yet-known content. For other students (or other content areas), it might be more effective to optimize for direct connections between already known content and new material. Observing how the student's knowledge changed over time, in response to their teaching, could also help to guide the teacher towards the most effective strategy for that individual student.

Designing and building procedures and tools for mapping out knowledge touches on deep questions about what it means to learn. For example, how do we acquire conceptual knowledge? Memorizing course lectures or textbook chapters by rote can lead to the superficial *appearance* of understanding the underlying content, but achieving true conceptual understanding seems to require something deeper and richer. Does conceptual understanding entail connecting newly acquired information to the scaffolding of one's existing knowledge or experience [**? ? ? ? ?**]? Or weaving a lecture's atomic elements (e.g., its component words) into a structured network that describes how those individual elements are related? Conceptual understanding could also involve building a mental model that transcends the meanings of those individual atomic elements by reflecting the deeper meaning underlying the gestalt whole [**? ? ?**].

The difference between "understanding" and "memorizing," as framed by researchers in education, cognitive psychology, and cognitive neuroscience [e.g., **? ? ? ? ?**] has profound analogs in the fields of natural language processing and natural language understanding. For example, considering the raw contents of a document (e.g., its constituent symbols, letters, and words) might provide some information about what the document is about, just as memorizing a passage might provide some ability to answer simple questions about it [e.g., whether it contains words related to

furniture versus physics; **? ? ?**]. However, modern natural language processing models [e.g., **? ? ?**] also attempt to capture the deeper meaning *underlying* those atomic elements. These models consider not only the co-occurrences of those elements within and across documents, but also patterns in how those elements appear across different scales (e.g., sentences, paragraphs, chapters, etc.), the temporal and grammatical properties of the elements, and other high-level characteristics of how they are used [**? ?**]. According to these models, the deep conceptual meaning of a document may be captured by a feature vector in a high-dimensional representation space, where nearby vectors reflect conceptually related documents. A model that succeeds at capturing an analog of "understanding" is able to assign nearby feature vectors to two conceptually related documents, *even when the words contained in those documents have very little overlap*.

Given these insights, what form might the representation of the sum total of a person's knowledge take? First, we might require a means of systematically describing or representing the nearly infinite set of possible things a person could know. Second, we might want to account for potential associations between different concepts. For example, the concepts of "fish" and "water" might be associated in the sense that fish live in water. Third, knowledge may have a critical dependency structure, such that knowing about a particular concept might require first knowing about a set of other concepts. For example, understanding the concept of a fish swimming in water first requires understanding what fish and water *are*. Fourth, as we learn, our "current state of knowledge" should change accordingly. Learning new concepts should both update our characterizations of "what is known" and also unlock any now-satisfied dependencies of those newly learned concepts so that they are "tagged" as available for future learning.

Here we develop a framework for modeling how knowledge is acquired during learning. The central idea is to use text embedding models to define the coordinate systems of two maps: (a) a *knowledge map* that describes the extent to which each concept is currently known and (b) a *learning map* that describes changes in knowledge over time. Each location on these maps represents a single concept, and the maps' geometries are defined such that related concepts are located nearby in space. We use this framework to analyze and interpret behavioral data collected from an experiment that had participants watch and answer multiple-choice questions about a series of

recorded course lectures.

Our primary research goal is to advance our understanding of what it means to acquire deep, real-world conceptual knowledge. Traditional laboratory approaches to studying learning and memory (e.g., list learning studies) often draw little distinction between memorization and understanding. Instead, these studies typically focus on whether information is effectively encoded or retrieved, rather than whether the information is *understood*. Approaches to studying conceptual learning, such as category learning experiments, can begin to investigate the distinction between memorization and understanding, often by training participants to distinguish arbitrary or random features in otherwise meaningless categorized stimuli. However the objective of real-world training, or learning from life experiences more generally, is often to develop new knowledge that may be applied in *useful* ways in the future. In this sense, the gap between modern learning theories and modern pedagogical approaches and classroom learning strategies is enormous: most of our theories about *how* people learn are inspired by experimental paradigms and models that have only peripheral relevance to the kinds of learning that students and teachers actually seek [**?** **?**]. To help bridge this gap, our study uses course materials from real online courses to inform, fit, and test models of real-world conceptual learning. We also provide a demonstration of how our models can be used to construct "maps" of what students know, and how their knowledge changes with training. In addition to helping to visualize knowledge (and changes in knowledge), we hope that such maps might lead to real-world tools for improving how we educate.

Results

At its core, our main modeling approach is based around a simple assumption that we sought to test empirically: all else being equal, knowledge about a given concept is predictive of knowledge about similar or related concepts. From a geometric perspective, this assumption implies that knowledge is fundamentally "smooth." In other words, as one moves through a space representing an individual's knowledge (where similar concepts occupy nearby coordinates), their "level of knowledge" should change relatively gradually throughout that space. To begin to test this

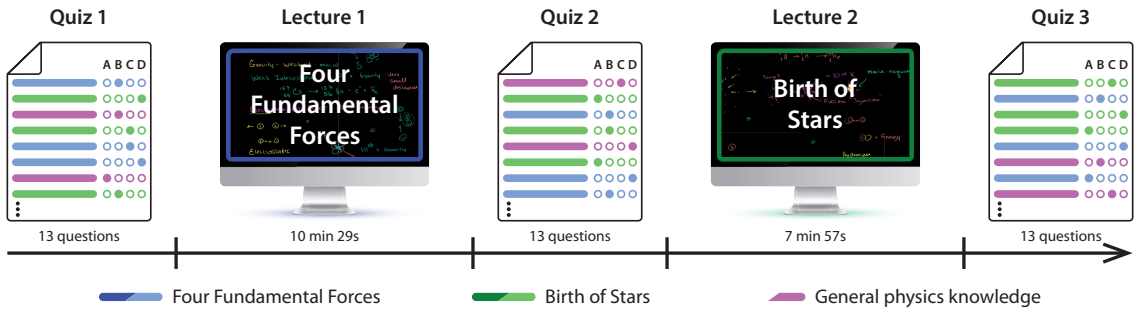

Figure 1: Experimental paradigm. Participants alternate between answering 13-question multiple-choice quizzes and watching two Khan Academy lectures. Each quiz contains a mix of 5 questions about lecture 1, 5 questions about lecture 2, and 3 questions about general physics knowledge. The specific questions reflected on each quiz, and the orders of each quiz's questions, were randomized across participants.

smoothness assumption, we sought to track participants' knowledge and how it changed over time in response to training.

We asked participants in our study to complete brief multiple-choice quizzes before, between, and after watching two lecture videos from the Khan Academy [**?**] platform (Fig. **??**). The first lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics: gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*, provided an overview of our current understanding of how stars form. We selected these lessons to be (a) accessible to a broad audience, i.e., requiring minimal prerequisite knowledge to understand; (b) conceptually related to each other, i.e., covering at least *some* similar or overlapping content; and (c) largely independent of each other, i.e., focused on sufficiently different material that understanding one did not require having seen the other. The two videos we selected are introductory lectures that both belong to Khan Academy's "Cosmology and Astronomy" course domain, but are taken from different lecture series ("Scale of the Universe" and "Stars, Black Holes, and Galaxies" for the first and second lectures, respectively).

We created a pool of multiple-choice quiz questions that would enable us to evaluate participants' knowledge about each individual lecture along with related content not specifically presented in either video (see Tab. S1). Participants answered questions randomly drawn from each content area (lecture 1, lecture 2, and general physics knowledge) on each of the three quizzes.

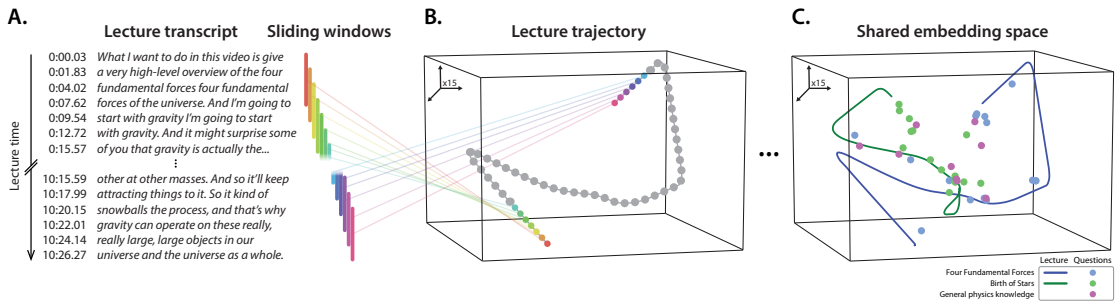

Figure 2: Modeling course content. A. Building a document pool from sliding windows of text. We decompose each lecture's transcript into a series of overlapping sliding windows. The set of transcript snippets (across all windows) may be treated as a set of "documents" for training a text embedding model. **B. Constructing lecture content *trajectories*.** After training our model on the sliding windows from both lectures, we transform each lecture into a "trajectory" through text embedding space by joining the embedding coordinates of successive sliding windows parsed from its transcript. **C. Embedding multiple lectures and questions in a shared space.** We apply the same model (trained on the two lectures' windows) to both lectures, along with the text of each question in our pool (Tab. S1), to project them into a shared text embedding space. This results in one trajectory per lecture and one coordinate for each question. Here we have projected the 15-dimensional embeddings onto their first 3 principal components for visualization.

Quiz 1 was intended to assess participants' "baseline" knowledge before training, quiz 2 assessed knowledge after watching the *Four Fundamental Forces* video (i.e., lecture 1), and quiz 3 assessed knowledge after watching the *Birth of Stars* video (i.e., lecture 2).

To study how participants' conceptual knowledge changed over the course of the experiment, we first sought to characterize the abstract concepts presented to them in each of the two lectures. We adapted an approach we developed in prior work [**?**] to extract the latent themes from the lectures' contents using a topic model [**?**]. Briefly, topic models take as input a collection of text documents and learn a set of "topics" (i.e., latent themes) from their contents. Once fit, a topic model can be used to transform arbitrary (potentially new) documents into sets of "topic proportions," describing the weighted blend of learned topics reflected in their texts. We parsed automatically generated transcripts of the two lectures into overlapping sliding windows, each containing the text of the lecture transcript from a particular time range. We treated the set of text (across all of these windows) as documents to fit our model (Fig. **??**A; see **??**). Transforming the text from every window with our model yielded a number-of-windows by number-of-topics (15) topic-proportions matrix that described the unique mixture of broad themes from both lectures

reflected in each window's content. Intuitively, each window's "topic vector" (i.e., column of the topic-proportions matrix) is analogous to a coordinate in a 15-dimensional space whose axes are topics discovered by the model. Within this space, each lecture's sequence of topic vectors (i.e., corresponding to sliding windows parsed from its transcript) forms a *trajectory* that captures how its conceptual content unfolds over time (Fig. **??**B). We resampled these trajectories to a resolution of one topic vector for each second of video (i.e., 1 Hz).

Next, we sought to characterize what information participants' performance on each of the three quizzes could provide about their conceptual knowledge at that point in time. Traditional approaches to evaluating students' performance on short-form knowledge assessments, such as those in our study, entail computing the proportion of correctly answered questions to assign a simple numeric score. While this may afford some measure of the *extent* of a learner's knowledge, such a score provides little information to either the teacher or the learner about the particular *contents* of their knowledge—in other words, raw "proportion-correct" measures may capture *how much* a student knows, but not *what* they know. For instance, suppose a participant in our study was highly knowledgeable about fundamental physical forces (i.e ., lecture 1 content) but unfamiliar with how stars are formed (i.e., lecture 2 content), while a second participant was unfamiliar with fundamental forces but had extensive prior knowledge about star formation. Since quiz 1 (completed before viewing either lecture) contained an equal number of questions about each lecture's content (see Fig. **??**), these two participants could easily achieve the same proportion-correct score despite their conceptual knowledge differing substantially. How might we distinguish these individuals?

We hypothesized that our text embedding model, which we had trained on transcripts of the two lectures to characterize their conceptual contents, should also allow us to capture the conceptual knowledge probed by each quiz question. If our model successfully represented information about the deeper conceptual content of the lectures (i.e., beyond surface-level details such as particular word choices) then we should be able to recover a correspondence between their embeddings and embeddings of other text that reflects related concepts (e.g., quiz questions that were ostensibly, by design, "about" one of the two lecture) despite differences in exact verbiage. Intuitively, while

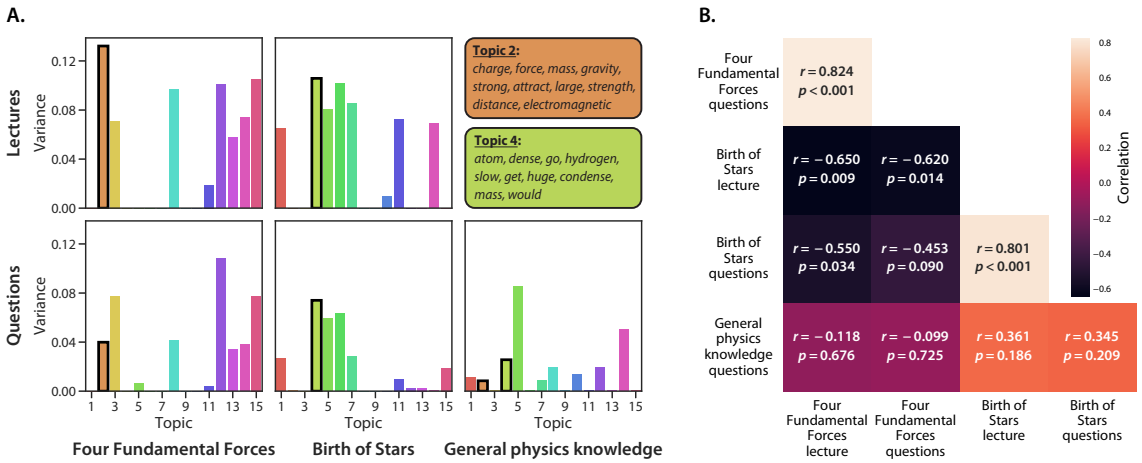

Figure 3: Lecture and question topic overlap. A. Topic weight variability. The bar plots display the variance of each topic's weight across lecture timepoints (top row) and questions (bottom row); colors denote topics. The top-weighted words from the most "expressive" (i.e., variable across observations) topic from each lecture are displayed in the upper right (orange: topic 2; yellow-green: topic 4). The top-weighted words from the full set of topics may be found in Table S2. **B. Relationships between topic weight variability.** Pairwise correlations between the distributions of topic weight variance for each lecture and question category. Each row and column corresponds to a bar plot in Panel A.

the content of each given lecture will naturally vary from moment to moment, the particular combination of dimensions along which its trajectory varies across moments correspond to topics that encode meaningful "information" [**?**] about that lecture's dynamic conceptual content. We therefore expected that quiz questions pertaining to those same concepts might express the same set of topics.

We used our model to transform the text of each question in our pool into the same embedding space as the two lectures' trajectories (Fig. **??**C). This yielded a single 15-dimensional coordinate (i.e., topic vector) for each question. We then computed the variance of each topic's weight across timepoints of each lecture, and across questions from each category (i.e., lecture 1-related, lecture 2-related, and general physics knowledge; Fig. **??**A). Visual inspection of Figure **??**A suggests a strong correspondence between the sets of topics expressed by each lecture and its related questions, with only limited overlap in the topics expressed by non-matched lecture-question set pairs. To quantify these apparent similarities and differences, we computed the correlation between each pair of topic-weight variance distributions (Fig. **??**B). We found that our model

represented the contents of lecture-related questions using a combination of topic dimensions (and relative variation along those topic dimensions) that was highly similar to that of their reference lecture (*Four Fundamental Forces* (*FFF*) questions vs. lecture: Pearson's $r(13) = 0.824$, $p < 0.001$, 95% confidence interval (CI) = $[0.696, 0.973]$; *Birth of Stars* (*BoS*) questions vs. lecture: $r(13) = 0.801$, $p < 0.001$, 95% CI = $[0.539, 0.958]$) but diverged significantly from that of the non-reference lecture (*FFF* questions vs. *BoS* lecture: $r(13) = -0.620$, $p = 0.014$, 95% CI = $[-0.871, -0.326]$; *BoS* questions vs. *FFF* lecture: $r(13) = -0.550$, $p = 0.034$, 95% CI = $[-0.803, -0.246]$). This indicated that our model captured the conceptual contents of the lectures and quiz questions sufficiently well to differentiate between questions relating to one lecture versus the other.

Although a single lecture may be organized around a single broad theme at a coarse scale, at a finer scale each moment of a lecture typically covers a narrower range of content. We wondered whether a text embedding model trained on the lectures' transcripts might capture some of this finer scale content. For example, if a particular question asks about the content from one small part of a lecture, we wondered whether our text embedding model could be used to automatically identify the "matching" moment(s) in the lecture. When we correlated each question's topic vector with the topic vectors for each second of the lectures, we found some evidence that each question is temporally specific (Fig. **??**). In particular, most questions' topic vectors were maximally correlated with a well-defined (and relatively narrow) range of timepoints from their corresponding lectures, and the correlations fell off sharply outside of that range. We also examined the best-matching intervals for each question qualitatively by comparing the text of the question to the text of the most-correlated parts of the lectures. Despite that the questions were excluded from the text embedding model's training set, in general we found (through manual inspection) a close correspondence between the conceptual content that each question covered and the content covered by the best-matching moments of the lectures. Two representative examples are shown at the bottom of Figure **??**.

The ability to quantify how much each question is "asking about" the content from each moment of the lectures could enable high-resolution insights into participants' knowledge. Traditional approaches to estimating how much a student "knows" about the content of a given lecture entail

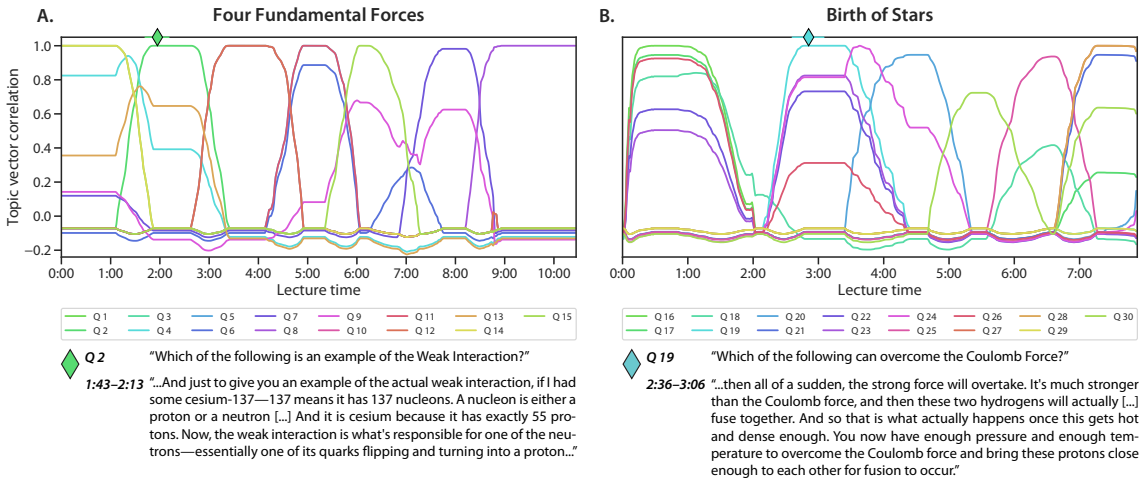

Figure 4: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question's topic vector correlates with each video timepoint's topic vector (Panel **A.**: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel **B.**: correlations for the *Birth of Stars* lecture and associated questions). The colors denote question identities. The diamonds in each panel denote the moment of peak correlation between the indicated question and the lecture trajectory. The associated questions' text and snippets of the lectures' transcripts from the surrounding 30 seconds, are displayed at the bottom of the figure.

computing the proportion of correctly answered questions. But if two students receive identical scores on an exam, might our modeling framework help us to gain more nuanced insights into the *specific* content that each student has mastered (or failed to master)? For example, a student who misses three questions that were all about the same concept (e.g., concept A) will have gotten the same *proportion* of questions correct as another student who missed three questions about three *different* concepts (e.g., A, B, and C). But if we wanted to fill in the "gaps" in the two students' understandings, we might do well to focus on concept A for the first student, but to also add in materials pertaining to concepts B and C for the second student.

We developed a simple formula (Eqn. **??**) for using a participant's responses to a small set of multiple-choice questions to estimate how much the participant "knows" about the concept reflected by any arbitrary coordinate, x, in text embedding space (e.g., the content reflected by any moment in a lecture they had watched; see **??**). Essentially, the estimated knowledge at the coordinate is given by the weighted average proportion of quiz questions the participant answered correctly, where the weights reflect how much each question is "about" the content at x. When we apply this approach to estimate the participant's knowledge about the content presented in each moment of each lecture, we can obtain a detailed timecourse describing how much "knowledge" the participant has about any part of the lecture. As shown in Figure **??**, we can also apply this approach separately for the questions from each quiz the participants took throughout the experiment. From just 13 questions per quiz, we obtain a high-resolution snapshot (at the time each quiz was taken) of what the participants knew about any moment's content, from either of the two lectures they watched (comprising a total of 1106 samples across the two lectures).

Of course, even though the timecourses in Figure **??**A and C provide detailed *estimates* about participants' knowlege, those estimates are only *useful* to the extent that they accurately reflect what participants actually know. As one sanity check, we anticipated that the knowledge estimates should show a content-specific "boost" in participants' knowledge after watching each lecture. In other words, if participants learn about each lecture's content when they watch each lecture, the knowledge estimates should reflect that. After watching the *Four Fundamental Forces* lecture, participants should show more knowledge for the content of that lecture than they had before,

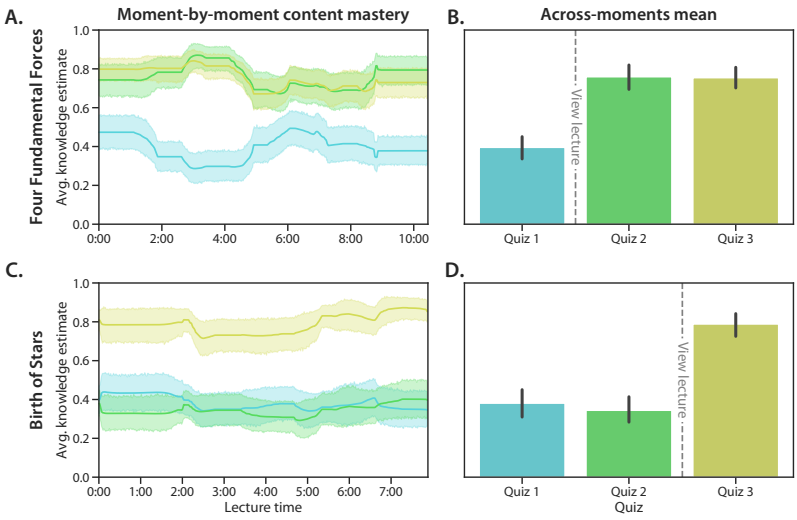

Figure 5: Estimating moment-by-moment knowledge acquisition. A. Moment-by-moment knowledge about the *Four Fundamental Forces*. Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (see **??**), using responses from one quiz (color). The traces are averaged across participants. **B. Average estimated knowledge about the Four Fundamental Forces.** Each bar displays the across-timepoint average knowledge, estimated using the responses to one quiz's questions. **C. Moment-by-moment knowledge about the *Birth of Stars*.** The panel is in the same format as Panel A, but here the knowledge estimates are for the moment-by-moment content of the *Birth of Stars* lecture. **D. Average estimated knowledge about the *Birth of Stars*.** The panel is in the same format as Panel B, but here the knowledge estimates are for the content of the *Birth of Stars* lecture. All panels: error ribbons and error bars denote 95% confidence intervals, estimated across participants.

and that knowledge should persist for the remainder of the experiment. Specifically, knowledge about that lecture's content should be relatively low when estimated using Quiz 1 responses, but should increase when estimated using Quiz 2 or 3 responses (Fig. **??**B). Indeed, we found that participants' estimated knowledge about the content of the *Four Fundamental Forces* was substantially higher on Quiz 2 versus Quiz 1 ($t(49) = 8.764, p < 0.001$) and on Quiz 3 versus Quiz 1 ($t(49) = 10.519, p < 0.001$). We found no reliable differences in estimated knowledge about that lecture's content on Quiz 2 versus 3 ($t(49) = 0.160, p = 0.874$). Similarly, we hypothesized (and subsequently confirmed) that participants should show more estimated knowledge about the content of the *Birth of Stars* lecture after (versus before) watching it (Fig. **??**D). Specifically, since participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a "boost" on Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ($t(49) = 1.013, p = 0.316$), but the estimated knowledge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and Quiz 3 versus 1 ($t(49) = 8.969, p < 0.001$).

If we are able to accurately estimate a participant's knowledge about the content tested by a given question, the estimated knowledge should have some predictive information about whether the participant is likely to answer the question correctly or incorrectly. For each question in turn, for each participant, we used Equation **??** to estimate (using all *other* questions from the same quiz, from the same participant) the participant's knowledge at the held-out question's embedding coordinate. For each quiz, we grouped these estimates into two distributions: one for the estimated knowledge at the coordinates of each *correctly* answered question, and another for the estimated knowledge at the coordinates of each *incorrectly* answered question (Fig. **??**). We then used independent samples t-tests to compare the means of these distributions of estimated knowledge.

For the initial quizzes participants took (prior to watching either lecture), participants' estimated knowledge tended to be low overall, and relatively unstructured (Fig. **??**, left panel). When we held out individual questions and estimated their knowledge at the held-out questions' embedding

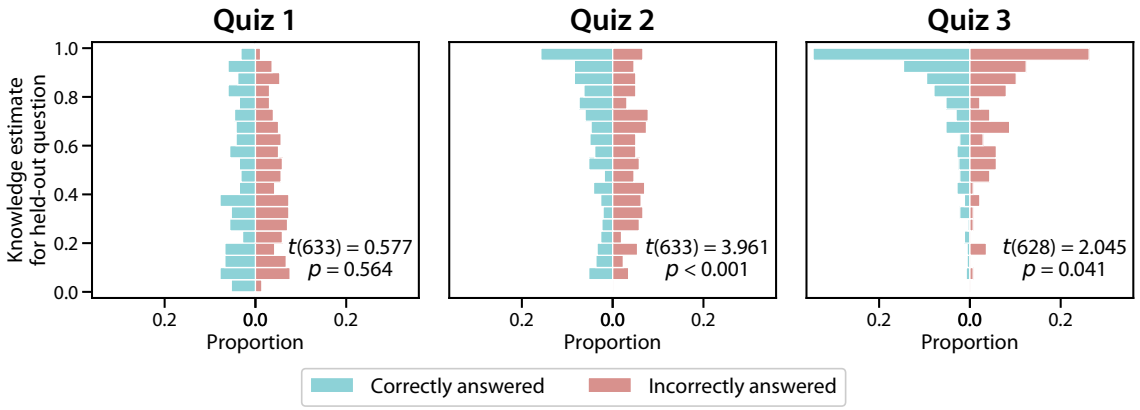

Figure 6: Estimating knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held out correctly (blue) or incorrectly (red) answered question. The *t*-tests reported in each panel are between the distributions of estimated knowledge at the coordinates of correctly versus incorrectly answered held-out questions.

coordinates, we found no reliable differences in the estimates when the held-out question had been correctly versus incorrectly answered ($t(633) = 0.577, p = 0.564$). After watching the first video, estimated knowledge for held-out correctly answered questions (from the second quiz; Fig. **??**, middle panel) exhibited a positive shift relative to held-out incorrectly answered questions ($t(633) = 3.961, p < 0.001$). After watching the second video, estimated knowledge (from the third quiz; Fig. **??**, right panel) for *all* questions exhibited a positive shift. However, the increase in estimated knowledge for held-out correctly answered questions was larger than for held-out incorrectly answered questions ($t(628) = 2.045, p = 0.041$).

Knowledge estimates need not be limited to the content of the lectures. As illustrated in Figure **??**, our general approach to estimating knowledge from a small number of quiz questions may be applied to *any* content, given its text embedding coordinate. To visualize how knowledge "spreads" through text embedding space to content beyond the lectures participants watched, we first fit a new topic model to the lectures' sliding windows with $k = 100$ topics. We hoped that increasing the number of topics from 15 to 100 might help us to generalize the knowledge predictions. (Aside from increasing the number of topics from 15 to 100, all other procedures and model parameters were carried over from the preceding analyses.) As in our other analyses, we

resampled each lecture's topic trajectory to 1 Hz and also projected each question into a shared text embedding space.

We projected the resulting 100-dimensional topic vectors (for each second of video and for each question) into a shared 2-dimensional space (see **??**). Next, we sampled points evenly from a 100×100 grid of coordinates that evenly tiled a rectangle enclosing the 2D projections of the videos and questions. We used Equation **??** to estimate participants' knowledge at each of these 10,000 sampled locations, and we averaged these estimates across participants to obtain an estimated average *knowledge map* (Fig. **??**). Intuitively, the knowledge map constructed from a given quiz's responses provides a visualization of how "much" participants know about any content expressible by the fitted text embedding model.

Several features of the resulting knowledge maps are worth noting. The average knowledge map estimated from Quiz 1 responses (Fig. **??**, leftmost map) shows that participants tended to have relatively little knowledge about any parts of the text embedding space (i.e., the shading is relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked increase in knowledge on the left side of the map (around roughly the same range of coordinates covered by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words, participants' estimated increase in knowledge is localized to conceptual content that is nearby (i.e., related to) the content from the lecture they watched prior to taking Quiz 2. This localization is non-trivial: the knowledge estimates are informed only by the embedded coordinates of the *quiz questions*, not by the embeddings of either lecture (Eqn. **??**). Finally, the knowledge map estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region surrounding the embedding of the *Birth of Stars* lecture participants watched prior to taking Quiz 3.

Another way of visualizing these content-specific increases in knowledge (apparently driven by watching each lecture) is displayed in Figure **??**B. Taking the point-by-point difference between the knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map* that describes the *change* in knowledge estimates from one quiz to the next. These learning maps highlight that the estimated knowledge increases we observed across maps were specific to the

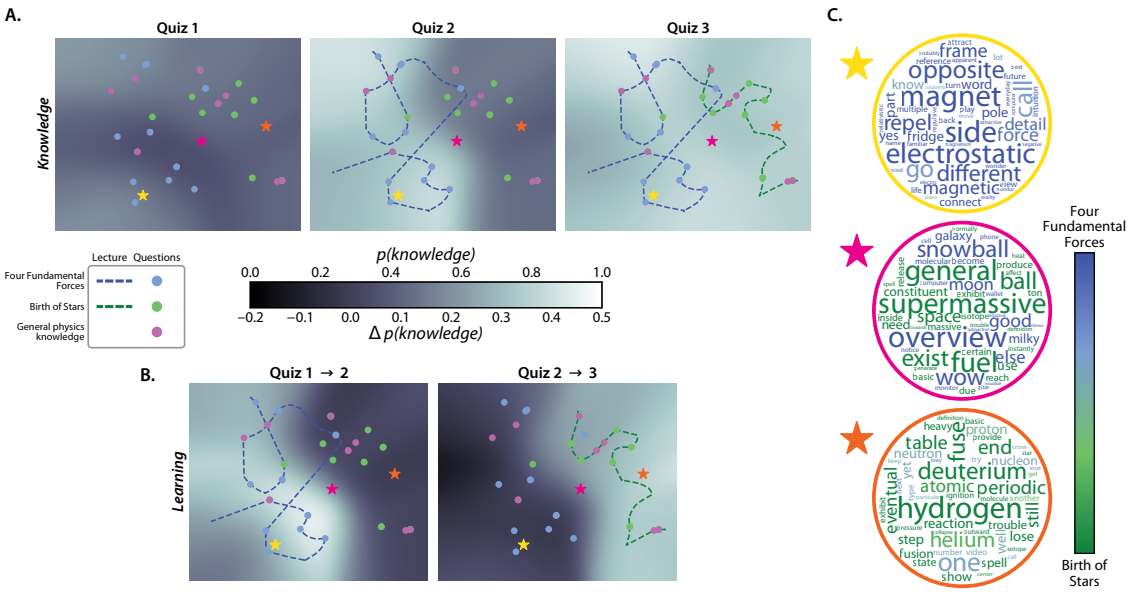

Figure 7: Mapping out the geometry of knowledge and learning. A. Average "knowledge maps" estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see **??**). The topic trajectories of each lecture and the coordinates of each question are indicated by dotted lines and dots. Each map reflects an average across all participants. For individual participants' maps, see Figures S1, S2, and S3. **B. Average "learning maps" estimated between each successive pair of quizzes.** The learning maps are in the same general format as the knowledge maps in Panel A, but each coordinate in the learning maps indicates the *difference* between the corresponding coordinates in the indicated *pair* of knowledge maps—i.e., how much the estimated knowledge "changed" across the two quizzes. Each map reflects an average across all participants. For individual participants' maps, see Figures S4 and S5. **C. Word clouds for sampled points in topic space.** Each word cloud displays the relative weights of each word (via their relative sizes) reflected by the blend of topics represented at the locations of the stars in the maps. The words' colors indicate how much each word is weighted, on average, across all timepoints' topic vectors in the *Four Fundamental Forces* (blue) and *Birth of Stars* (green) videos, respectively.

regions around the embeddings of each lecture in turn.

Because the 2D projection we used to construct the knowledge and learning maps is invertible, we may gain additional insights into the estimates by reconstructing the original high-dimensional topic vectors for any point(s) in the maps we are interested in. For example, this could serve as a useful tool for an instructor looking to better understand which content areas a student (or a group of students) knows well (or poorly). As a demonstration, we show the top-weighted words from the blends of topics reconstructed from three example locations on the maps (Fig. **??**C): one point near the *Four Fundamental Forces* embedding (yellow); a second point near the *Birth of Stars* embedding (orange), and a third point somewhere in between the two lectures' embeddings (pink). As shown in the word clouds in the Panel, the top-weighted words at the example coordinate near the *Four Fundamental Forces* embedding also tended to be weighted heavily by the topics expressed in that lecture. Similarly, the top-weighted words at the example coordinate near the *Birth of Stars* embedding tended to be weighted most heavily by the topics expressed in *that* lecture. And the top-weighted words at the example coordinate between the two lectures' embeddings show a roughly even mix of words most strongly associated with each lecture.

Discussion

Teaching, like effective writing and speaking, is fundamentally about empathy [**? ? ?**]. Great teachers consider students' interests [**? ?**], backgrounds [**? ? ?**], and working memory capacities [**?**], and flexibly optimize their teaching strategies within those constraints [**? ? ?**]. In the classroom, empathizing with students also means maintaining open lines of communication [**?**] by fostering an environment in which all students feel comfortable speaking up if they have an exciting new idea, or if they are having trouble understanding something [**? ?**]. In-person instruction also often entails dynamic student-teacher and student-student interactions. These in-person interactions can provide the instructor with valuable information about students' understanding of the course material, beyond what they can glean solely from exams or assignments [**? ? ?**]. In turn, this can allow the instructor to adapt their teaching approaches on-the-fly according to students' questions

and behaviors. But what does great teaching look like in asynchronous online courses, when the instructor typically prepares course lectures and materials without knowing who will ultimately be learning from them? Can the empathetic side of teaching be automated and scaled?

The notion of empathy also related to "theory of mind" of other individuals [**? ? ?**]. Considering others' unique perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and communicate [**? ? ?**]. The knowledge and learning maps we estimate in our study (Fig. **??**) hint at one potential form that an automated "empathetic" teacher might take. We imagine automated content delivery systems that adapt lessons on the fly according to continually updated estimates of what students know and how quickly they are learning different conceptual content [e.g., building on ideas such as **? ? ? ?** , and others].

Over the past several years, the global pandemic has forced many educators to teach remotely [**? ? ? ?**]. This change in world circumstances is happening alongside (and perhaps accelerating) geometric growth in the availability of high quality online courses on platforms such as Khan Academy [**?**], Coursera [**?**], EdX [**?**], and others [**?**]. Continued expansion of the global internet backbone and improvements in computing hardware have also facilitated improvements in video streaming, enabling videos to be easily downloaded and shared by large segments of the world's population. This exciting time for online course instruction provides an opportunity to re-evaluate how we, as a global community, educate ourselves and each other. For example, we can ask: what makes an effective course or training program? Which aspects of teaching might be optimized or automated? How can we provide How and why do learning needs and goals vary across people? How might we lower barriers to achieving a high quality education?

Alongside these questions, there is a growing desire to extend existing theories beyond the domain of lab testing rooms and into real classrooms [**?**]. In part, this has led to a recent resurgence of "naturalistic" or "observational" experimental paradigms that attempt to better reflect more ethologically valid phenomena that are more directly relevant to real-world situations and behaviors [**?**]. In turn, this has brought new challenges in data analysis and interpretation. A key step towards solving these challenges will be to build explicit models of real-world scenarios and how people behave in them (e.g., models of how people learn conceptual content from real-

world courses, as in our current study). A second key step will be to understand which sorts of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; **? ? ? ? ?**] might help to inform these models. A third major step will be to develop and employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic paradigms.

Ultimately, our work suggests a new line of questions regarding the future of education: which aspects of teaching can be optimized and/or automated? The social benefits of face-to-face instruction, such as social interactions, friendships, and emotional support, cannot (and perhaps should not) be fully replaced by an automated computer-based system. Nor can modern computer systems experience emotional empathy in the human sense of the word. On the other hand, perhaps it is possible to separate out the social aspects of classroom instruction from the purely learning-related aspects. Our study shows that text embedding models can uncover detailed insights into students' knowledge and how it changes over time during learning. We hope that these advances might help pave the way for new ways of teaching or delivering educational content that are tailored to individual students' learning needs and goals.

Materials and methods

Participants

We enrolled a total of 50 Dartmouth undergraduate students in our study. Participants received course credit for enrolling. We asked each participant to fill out a demographic survey that included questions about their age, gender, native spoken language, ethnicity, race, hearing, color vision, sleep, coffee consumption, level of alertness, and several aspects of their educational background and prior coursework.

Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09 years). A total of 15 participants reported their gender as male and 35 participants reported their gender as female. A total of 49 participants reported their native language as "English" and 1 reported having another native language. A total of 47 participants reported their ethnicity as

"Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants reported their races as White (32 participants), Asian (14 participants), Black or African American (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

A total of 49 participants reporting having normal hearing and 1 participant reported having some hearing impairment. A total of 49 participants reported having normal color vision and 1 participant reported being color blind. Participants reported having had, on the night prior to testing, 2–4 hours of sleep (1 participant), 4–6 hours of sleep (9 participants), 6–8 hours of sleep (35 participants), or 8+ hours of sleep (5 participants). They reported having consumed, on the same day and leading up to their testing session, 0 cups of coffee (38 participants), 1 cup of coffee (10 participants), 3 cups of coffee (1 participant), or 4+ cups of coffee (1 participant).

No participants reported that their focus was currently impaired (e.g., by drugs or alcohol). Participants reported their current level of alertness, and we converted their responses to numerical scores as follows: "very sluggish" (-2), "a little sluggish" (-1), "neutral" (0), "fairly alert" (1), and "very alert" (2). Across all participants, a range of alertness levels were reported (range: -2 – 1; mean: -0.10; standard deviation: 0.84).

Participants reported their undergraduate major(s) as "social sciences" (28 participants), "natural sciences" (16 participants), "professional" (e.g., pre-med or pre-law; 8 participants), "mathematics and engineering" (7 participants), "humanities" (4 participants), or "undecided" (3 participants). Note that some participants selected multiple categories for their undergraduate major. We also asked participants about the courses they had taken. In total, 45 participants reported having taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan Academy courses. Of those who reported having watched at least one Khan Academy course, 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We also asked participants about the specific courses they had watched, categorized under different subject areas. In the "Mathematics" area, participants reported having watched videos on AP Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-

culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants), Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other videos not listed in our survey (5 participants). In the "Science and engineering" area, participants reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 participants); Physics, AP Physics I, or AP Physics II (15 participants); Biology, AP Biology; or High school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed in our survey (19 participants). We also asked participants whether they had specifically seen the videos used in our experiment. Of the 45 participants who reported having having taken at least one Khan Academy course in the past, 44 participants reported that they had not watched the *Four Fundamental Forces* video, and 1 participant reported that they were not sure whether they had watched it. All participants reported that they had not watched the *Birth of Stars* video. When we asked participants about non-Khan Academy online courses, they reported having watched or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 participants), Computing (2 participants), and other categories not listed in our survey (18 participants). Finally, we asked participants about in-person courses they had taken in different subject areas. They reported taking courses in Mathematics (39 participants), Science and engineering (38 participants), Arts and humanities (35 participants), Test preparation (27 participants), Economics and finance (26 participants), Computing (15 participants), College and careers (7 participants), or other courses not listed in our survey (6 participants).

Experiment

We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces* (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force; duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed; duration: 7 minutes and 57 seconds). We then hand-created 39 multiple-choice questions: 15 about

the conceptual content of *Four Fundamental Forces* (i.e., lecture 1), 15 about the conceptual content of *Birth of Stars* (i.e., lecture 2), and 9 questions that tested for general conceptual knowledge about basic physics (covering material that was not presented in either video). The full set of questions and answer choices may be found in Table S1.

Over the course of the experiment, participants completed three 13-question multiple-choice quizzes: the first before viewing lecture 1, the second between lectures 1 and 2, and the third after viewing lecture 2 (Fig. **??**). The questions appearing on each quiz, for each participant, were randomly chosen from the full set of 39, with the constraints that (a) each quiz contain 5 questions about lecture 1, 5 questions about lecture 2, and 3 questions about general physics knowledge, and (b) each question appear exactly once for each participant. The order of questions on each quiz, and the order of answer options for each question, were also randomized. Our experimental protocol was approved by the Committee for the Protection of Human Subjects at Dartmouth College. We used the experiment to develop and test our computational framework for estimating knowledge and learning.

Analysis

Constructing text embeddings of multiple lectures and questions

We adapted an approach we developed in prior work [**?**] to embed each moment of the two lectures and each question in our pool in a common representational space. Briefly, our approach uses a topic model (Latent Dirichlet Allocation; **?**), trained on a set of documents, to discover a set of k "topics" or "themes." Formally, each topic is defined as a set of weights over each word in the model's vocabulary (i.e., the union of all unique words, across all documents, excluding "stop words."). Conceptually, each topic is intended to give larger weights to words that are conceptually related or that tend to co-occur in the same documents. After fitting a topic model, each document in the training set, or any *new* document that contains at least some of the words in the model's vocabulary, may be represented as a k-dimensional vector describing how much the document (most probably) reflects each topic. (Unless, otherwise noted, we used $k = 15$ topics.)

As illustrated in Figure **??**A, we start by building up a corpus of documents using overlapping sliding windows that span each video's transcript. Khan Academy provides professionally created, manual transcriptions of all videos for closed captioning. However, such transcripts would not be readily available in all contexts to which our framework could potentially be applied. Khan Academy videos are hosted on the YouTube platform, which additionally provides automated captions. We opted to use these automated transcripts (which, in prior work, we have found are sufficiently near-human quality yield reliable data in behavioral studies; **?**) when developing our framework in order to make it more easily extensible and adaptable by others in the future.

We fetched these automated transcripts using the `youtube-transcript-api` Python package (**Jeremy can you add a citation for this?** https://github.com/jdepoix/youtube-transcript-api). The transcripts consisted of one timestamped line of text for every few seconds (mean: 2.34s; standard deviation: 0.83s) of spoken content in the video (i.e., corresponding to each individual caption that would appear on-screen if viewing the lecture via YouTube, and when those lines would appear). We defined a sliding window length of (up to) $w = 30$ transcript lines, and assigned each window a timestamp corresponding to the midpoint between its first and last lines' timestamps. These sliding windows ramped up and down in length at the very beginning and end of the transcript, respectively. In other words, the first sliding window covered only the first line from the transcript; the second sliding window covered the first two lines; and so on. This insured that each line of the transcript appeared in the same number (w) of sliding windows. After performing various standard text preprocessing (e.g., normalizing case, lemmatizing, removing punctuation and stop-words), we treated the text from each sliding window as a single "document," and we combined these documents across the two videos' windows to create a single training corpus for the topic model. The top words from each of the 15 discovered topics may be found in Table S2.

After fitting a topic model to each videos' transcripts, we could use the trained model to transform arbitrary (potentially new) documents into k-dimensional topic vectors. A convenient property of these topic vectors is that documents that reflect similar blends of topics (i.e., documents that reflect similar themes, according to the model) will yield similar coordinates (in terms of Euclidean distance, correlation, or other geometric measures). In general, the similarity between

different documents' topic vectors may be used to characterize the similarity in conceptual content between the documents.

We transformed each sliding window's text into a topic vector, and then used linear interpolation (independently for each topic dimension) to resample the resulting timeseries to one vector per second. We also used the fitted model to obtain topic vectors for each question in our pool (Tab. S1). Taken together, we obtained a *trajectory* for each video, describing its path through topic space, and a single coordinate for each question (Fig. **??**C). Embedding both videos and all of the questions using a common model enables us to compare the content from different moments of videos, compare the content across videos, and estimate potential associations between specific questions and specific moments of video.

Estimating dynamic knowledge traces

We used the following equation to estimate each participant's knowledge about timepoint t of a given lecture, $\hat{k}(t)$:

$$\hat{k}\left(f(t,L)\right) = \frac{\sum_{i \in \text{correct}} \text{ncorr}\left(f(t,L), f(i,Q)\right)}{\sum_{j=1}^{N} \text{ncorr}\left(f(t,L), f(j,Q)\right)}, \tag{1}$$

where

$$\text{ncorr}(x,y) = \frac{\text{corr}(x,y) - \text{mincorr}}{\text{maxcorr} - \text{mincorr}}, \tag{2}$$

and where mincorr and maxcorr are the minimum and maximum correlations between any lecture timepoint and question, taken over all timepoints and questions across both lectures and all five question used to estimate the knowledge trace **Note: make sure this is correct in results section; not sure i caught all instances**. We also define $f(s,\Omega)$ as the s^{th} topic vector from the set of topic vectors Ω. Here t indexes the set of lecture topic vectors, L, and i and j index the topic vectors of questions used to estimate the knowledge trace, Q. Note that "correct" denotes the set of indices of the questions the participant answered correctly on the given quiz.

Intuitively, $\text{ncorr}(x,y)$ is the correlation between two topic vectors (e.g., the topic vector from one timepoint in a lecture, x, and the topic vector for one question, y), normalized by the minimum and

maximum correlations (across all timepoints t and questions Q) to range between 0 and 1, inclusive. Equation **??** then computes the weighted average proportion of correctly answered questions about the content presented at timepoint t, where the weights are given by the normalized correlations between timepoint t's topic vector and the topic vectors for each question. The normalization step (i.e., using ncorr instead of the raw correlations) insures that every question contributes some non-zero amount to the knowledge estimate.

Creating knowledge and learning map visualizations

An important feature of our approach is that, given a trained text embedding model and participants' quiz performance on each question, we can estimate their knowledge about *any* content expressible by the embedding model—not solely the content explicitly probed by the quiz questions or even appearing in the lectures. To visualize these estimates (Figs. **??**, S1, S2, S3, S4, and S5), we used Uniform Manifold Approximation and Projection (UMAP; **?**) to construct a 2D projection of the text embedding space. Sampling the original 100-dimensional space at high resolution to obtain an adequate set of topic vectors spanning the embedding space would be computationally intractable. However, sampling a 2D grid is much more feasible. We defined a rectangle enclosing the 2D projections of the lectures' and quizzes' embeddings, and we sampled points from a regular 100×100 grid of coordinates that evenly tiled the enclosing rectangle. We sought to estimate participants' knowledge (and learning, i.e., changes in knowledge) at each of the resulting 10,000 coordinates.

To generate our estimates, we placed a set of 39 radial basis functions (RBFs) throughout the embedding space, centered on the 2D projections for each question (i.e., we included one RBF for each question). At coordinate x, the value of an RBF centered on a question's coordinate μ, is given by:

$$\text{RBF}(x, \mu, \lambda) = \exp\left\{-\frac{\|x - \mu\|^2}{\lambda}\right\}. \tag{3}$$

The λ term in the RBF equation controls the "smoothness" of the function, where larger values of λ result in smoother maps. In our implementation we used $\lambda = 50$. Next, we estimated the

"knowledge" at each coordinate, x, using:

$$\hat{k}(x) = \frac{\sum_{i \in \text{correct}} \text{RBF}(x, q_i, \lambda)}{\sum_{j=1}^{N} \text{RBF}(x, q_j, \lambda)}. \tag{4}$$

Intuitively, Equation **??** computes the weighted proportion of correctly answered questions, where the weights are given by how nearby (in the 2D space) each question is to the x. We also defined *learning maps* as the coordinate-by-coordinate differences between any pair of knowledge maps. Intuitively, learning maps reflect the *change* in knowledge across two maps.

Author contributions

Conceptualization: PCF, ACH, and JRM. Methodology: PCF, ACH, and JRM. Software: PCF. Validation: PCF. Formal analysis: PCF. Resources: PCF, ACH, and JRM. Data curation: PCF. Writing (original draft): JRM. Writing (review and editing): PCF, ACH, and JRM. Visualization: PCF and JRM. Supervision: JRM. Project administration: PCF. Funding acquisition: JRM.

Data and code availability

All of the data analyzed in this manuscript, along with all of the code for running our experiment and carrying out the analyses may be found at https://github.com/ContextLab/efficient-learning-khan.

Acknowledgements

We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of this study, and assistance with some of the data collection efforts from Will Baxley, Max Bluestone, Daniel Carstensen, Kunal Jha, Caroline Lee, Lucy Owen, Xinming Xu, and Kirsten Ziman. Our work was supported in part by NSF CAREER Award Number 2145172 to JRM. The content is solely the responsibility of the authors and does not necessarily represent the official views of our

supporting organizations. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.