Text embedding models yield 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 apply our framework to the videos' transcripts and the text of the quiz questions to quantify the content of each moment of video and each quiz question. We use 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 knows. 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 knows the to-be-learned information already, or how much they know 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.

A common approach to assessing a student's knowledge is to present them with a set of quiz questions, calculate the proportion they answer correctly, and provide them with feedback in the form of a simple numeric or letter grade. While such a grade can provide *some* indication of whether the student has mastered the to-be-learned material, any univariate measure of performance on a complex task sacrifices certain relevant information, risks conflating underlying factors, and so on. For example, consider the relative utility of the theoretical map described above that characterizes a student's knowledge in detail, versus a single annotation saying that the student answered 85% of their quiz questions correctly, or that they received a 'B'. Here, we show that the same quiz data required to compute proportion-correct scores or letter grades can instead be used to obtain far more detailed insights into what a student knew at the time they took the quiz.

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 [3, 7, 9, 10, 50]? Or weaving a lecture's atomic elements (e.g., its component words) into a structured network that describes how those individual elements are related [31]? 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 [28, 32, 47].

The difference between "understanding" and "memorizing," as framed by researchers in education, cognitive psychology, and cognitive neuroscience (e.g., 17, 20, 24, 32, 47), 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 clues as to what the document is about, just as memorizing a passage might provide some ability to answer simple questions about it. However, text embedding models (e.g., 4–6, 8, 11, 30, 39) 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 [33, 34]. According to these models, the deep conceptual meaning of a document may be captured by a feature vector in a high-dimensional representation space, wherein nearby vectors reflect conceptually related documents. A model that succeeds at capturing an analogue of "understanding" is able to assign nearby feature vectors to two conceptually related documents, *even when the specific words contained in those documents have very little overlap*.

Given these insights, what form might a 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 conceptual knowledge is acquired during learning. The central idea behind our framework is to use text embedding models to define the coordinate systems of two maps: a *knowledge map* that describes the extent to which each concept is currently known, and 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 answer sets of 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 [1, 14, 15, 18, 22, 45]. 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 that inform 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 [20, 32]. 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 visually capture knowledge (and changes in knowledge), we hope that such maps might lead to real-world tools for improving how we educate. Taken together, our work shows that existing course materials and evaluative tools like short multiple-choice quizzes may be leveraged to gain highly detailed insights into what students know and how they learn.

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. To begin to test this smoothness assumption, we sought to track participants' knowledge and how it changed over time in response to training. Two overarching goals guide our approach. First, we want to gain detailed insights into what learners know at different points in their training. For example, rather than simply reporting on the proportions of questions participants answer correctly (i.e., their overall performance), we seek estimates of their knowledge about a variety of specific concepts. Second, we want our approach to be potentially scalable to large numbers of diverse concepts, courses, and students. This requires that the conceptual content of interest be discovered *automatically*, rather than relying on manually produced ratings or labels.

We asked participants in our study to complete brief multiple-choice quizzes before, between, and after watching two lecture videos from the Khan Academy [27] platform (Fig. 1). 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 particular lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad

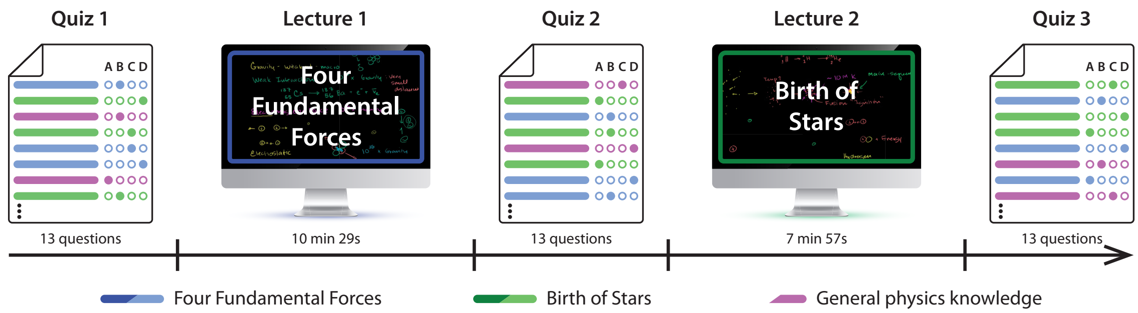

Figure 1: Experimental paradigm. Participants alternate between completing three 13-question multiple-choice quizzes and watching two Khan Academy lectures. Each quiz contains a mix of 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general physics knowledge. The specific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

audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training on participants' abilities to learn from the lectures. To this end, we selected two introductory videos that were intended to be viewed at the start of students' training in their respective content areas. Second, we wanted the two lectures to have some related content, so that we could test our approach's ability to distinguish similar conceptual content. To this end, we chose two videos from the same Khan Academy course domain, "Cosmology and Astronomy." Third, we sought to minimize dependencies and specific overlap between the videos. For example, we did not want participants' abilities to understand one video to (directly) influence their abilities to understand the other. To satisfy this last criterion, we chose videos from two different lecture series (Lectures 1 and 2 were from the "Scale of the Universe" and "Stars, Black Holes, and Galaxies" series, respectively).

We also wrote a set of multiple-choice quiz questions that we hoped would enable us to evaluate participants' knowledge about each individual lecture, along with related knowledge about physics concepts not specifically presented in either video (see Supp. Tab. 1 for the full list of questions in our stimulus pool). Participants answered questions randomly drawn from each content area (Lecture 1, Lecture 2, and general physics knowledge) on each of the three quizzes. 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).

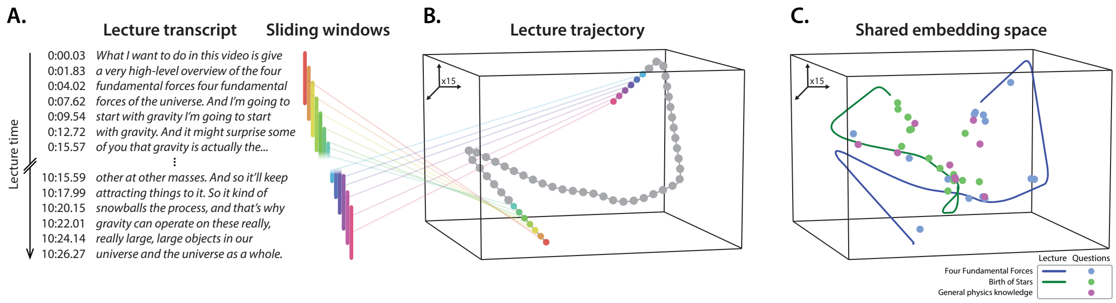

Figure 2: Modeling course content. A. Building a document pool from sliding windows of text. We decompose each lecture's transcript into a series of overlapping sliding windows. The full set of transcript snippets (across all windows) may be treated as a set of "documents" for training a text embedding model. **B. Constructing lecture content *trajectories*.** After training the 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 (Supp. Tab. 1), 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.

To study in detail how participants' conceptual knowledge changed over the course of the experiment, we first sought to model the conceptual content presented to them at each moment throughout each of the two lectures. We adapted an approach we developed in prior work [21] to identify the latent themes in the lectures using a topic model [5]. 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, where each window contained the text of the lecture transcript from a particular time span. We treated the set of text snippets (across all of these windows) as documents to fit the model (Fig. 2A; see *Constructing text embeddings of multiple lectures and questions*). Transforming the text from every sliding window with the model yielded a number-of-windows by number-of-topics (15) topic-proportions matrix describing the unique mixture of broad themes from both lectures reflected in each window's text. 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 its transcript's overlapping text snippets across sliding windows) forms a *trajectory* that captures how its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution of one topic vector for each second of video (i.e., 1 Hz).

We hypothesized that a topic model trained on transcripts of the two lectures should also capture the conceptual knowledge probed by each quiz question. If indeed the topic model could capture 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 each lecture and questions *about* each lecture. Importantly, such a correspondence could not solely arise from superficial text matching between lecture transcripts and questions, since the lectures and questions used different words. Simply comparing the average topic weights from each lecture and question set (averaging across time and questions, respectively) reveals a striking correspondence (Supp. Fig. 2). Specifically, the average topic weights from Lecture 1 are strongly correlated with the average topic weights from Lecture 1 questions ($r(13) = 0.809$, $p < 0.001$, 95% confidence interval (CI) = [0.633, 0.962]), and the average topic weights from Lecture 2 are strongly correlated with the average topic weights from Lecture 2 questions ($r(13) = 0.728$, $p = 0.002$, 95% CI = [0.456, 0.920]). At the same time, the average topic weights from the two lectures are *negatively* correlated with their non-matching question sets (Lecture 1 video vs. Lecture 2 questions: $r(13) = -0.547$, $p = 0.035$, 95% CI = [-0.812, -0.231]; Lecture 2 video vs. Lecture 1 questions: $r(13) = -0.612$, $p = 0.015$, 95% CI = [-0.874, -0.281]), indicating that the topic model also exhibits some degree of specificity. The full set of pairwise comparisons between average topic weights for the lectures and question sets is reported in Supplementary Figure 2.

Another, more sensitive, way of summarizing the conceptual content of the lectures and questions is to look at *variability* in how topics are weighted over time and across different questions (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much "information" [16] the lecture (or question set) reflects about that topic. For example, suppose a given topic is weighted on heavily throughout a lecture. That topic might be characteristic of some aspect or property of the lecture *overall* (conceptual or otherwise), but unless the topic's weights

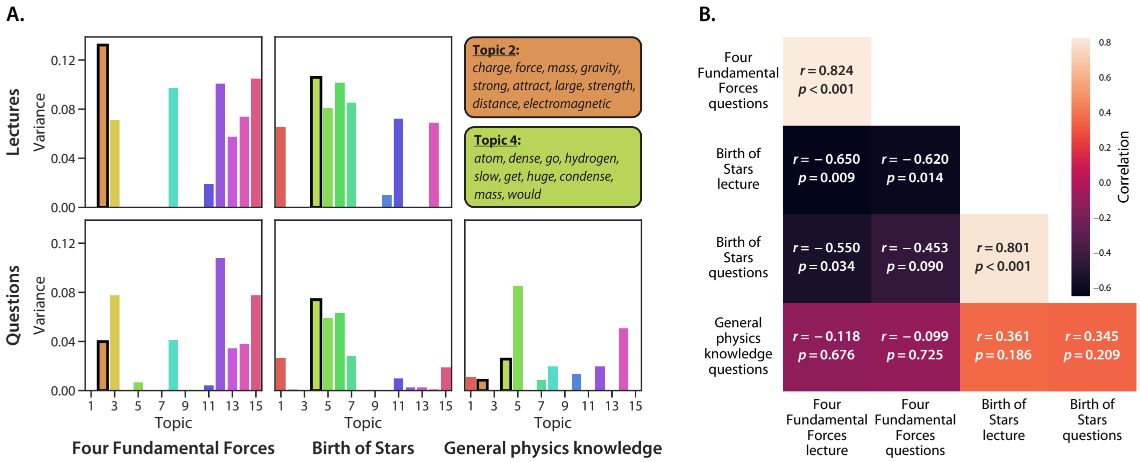

Figure 3: Lecture and question topic overlap. A. Topic weight variability. The bar plots display the variance of each topic's weight across lecture timepoints (top row) and questions (bottom row); colors denote topics. The top-weighted words from the most "expressive" (i.e., variable across observations)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

changed in meaningful ways over time, the topic would be a poor indicator of any *specific* conceptual content in the lecture. We therefore also compared the variances in topic weights (across time or questions) between the lectures and questions. The variability in topic expression (over time and across questions) was similar for the Lecture 1 video and questions ($r(13) = 0.824$, $p < 0.001$, 95% CI = [0.696, 0.973]) and the Lecture 2 video and questions ($r(13) = 0.801$, $p < 0.001$, 95% CI = [0.539, 0.958]). Simultaneously, as reported in Figure 3B, the variability in topic expression across *different* videos and lecture-specific questions (i.e., Lecture 1 video vs. Lecture 2 questions; Lecture 2 video vs. Lecture 1 questions) were negatively correlated, and neither video's topic variability was reliably correlated with the topic variability across general physics knowledge questions. Taken together, the analyses reported in Figure 3 and Supplementary Figure 2 indicate that a topic model fit to the videos' transcripts can also reveal correspondences (at a coarse scale) between the lectures and questions.

While an individual 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. Given the correspondence we found between the variability in topic expression across moments of each lecture and questions from its corresponding set (Fig. 3), we wondered whether the text embedding model might additionally capture these conceptual relationships at a finer scale. For example, if a particular question asks about the content from one small part of a lecture, we wondered whether the text embeddings could be used to automatically identify the "matching" moment(s) in the lecture. To explore this, we computed the correlation between each question's topic weights and the topic weights for each second of its corresponding lecture, and found that each question appeared to be temporally specific (Fig. 4). 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 qualitatively examined the best-matching intervals for each question by comparing the question's text 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 probed and the content covered by the best-matching moments of the lectures. Two representative examples are shown at the bottom of Figure 4.

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 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 help these two students fill in the "gaps" in their understandings, we might do well to focus specifically on concept A for the first student, but to also add in materials pertaining to concepts B and C for the second student. In other words, raw "proportion-correct" measures may capture *how much* a student knows, but not *what* they know.

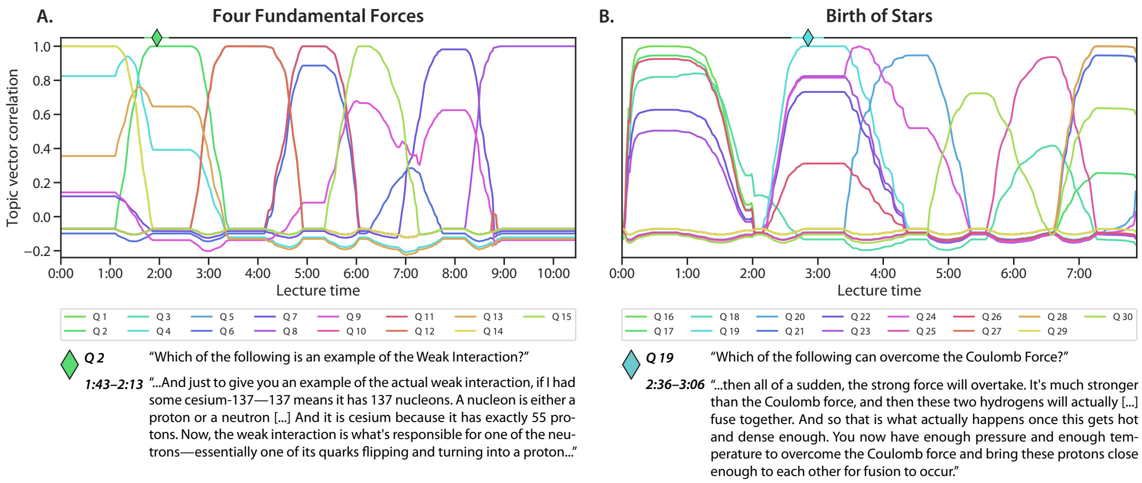

Figure 4: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question's topic vector correlates with each video timepoint's topic vector (Panel **A.**: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel **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.

We wondered whether our modeling framework might enable us to (formally and automatically) infer participants' knowledge at the scale of individual concepts (e.g., as captured by a single moment of a lecture).

We developed a simple formula (Eqn. 1) 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 *Estimating dynamic knowledge traces*). Essentially, the estimated knowledge at coordinate x 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 5A and C, we can apply this approach separately for the questions from each quiz participants took through-

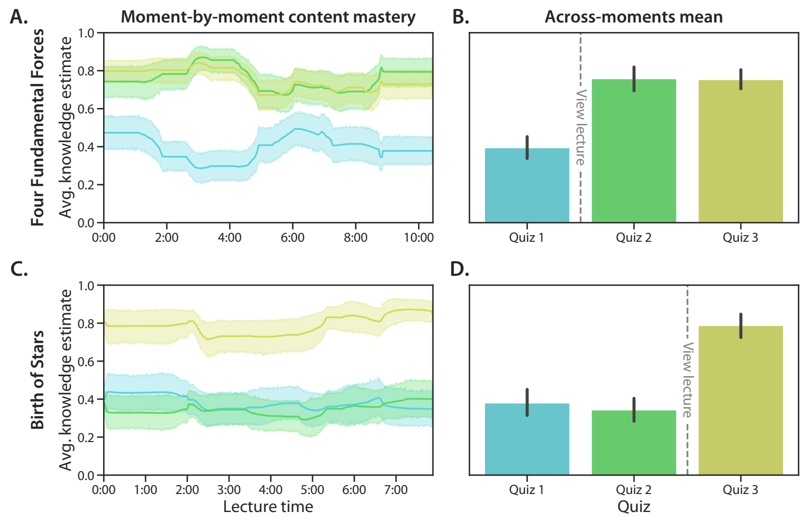

Figure 5: Estimating moment-by-moment knowledge acquisition. A. Moment-by-moment knowledge about the *Four Fundamental Forces*. Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (see *Estimating dynamic knowledge traces*), using responses from a single 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.

out the experiment. From just a few questions per quiz (see *Estimating dynamic knowledge traces*), 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 1,100 samples across the two lectures).

While the timecourses in Figure 5A and C provide detailed *estimates* about participants' knowledge, these estimates are of course only *useful* to the extent that they accurately reflect what participants actually know. As one sanity check, we anticipated that the knowledge estimates should reflect 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 capture that. After watching the *Four Fundamental Forces* lecture, participants should exhibit more knowledge for the content of that lecture than they had before, 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. 5B). 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 greater estimated knowledge about the content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 5D). 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, our estimates of their knowledge should carry some predictive information about whether the participant is likely to answer that question correctly or incorrectly. We developed a statistical approach to test this claim. For each question, in turn, we used Equation 1 to estimate each participant's knowledge at the given question's embedding space coordinate, using all *other* questions that participant answered on the same quiz. For each quiz, we grouped these estimates into two distributions: one for the estimated knowledge at the coordinates of *correctly* answered questions, and another for the estimated knowledge at the coordinates of *incorrectly* answered questions (Fig. 6). 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. 6, left panel). When we held out individual questions and estimated their knowledge at the held-out questions' embedding

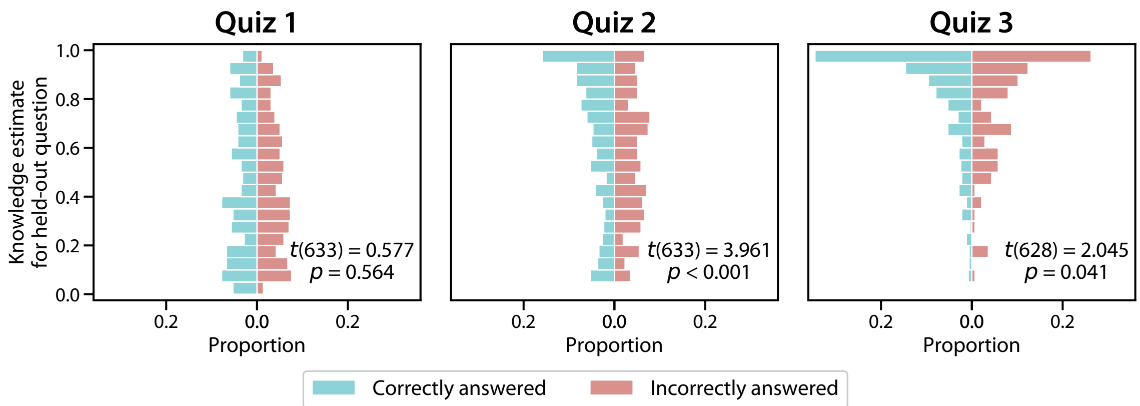

Figure 6: Estimating knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held-out correctly (blue) or incorrectly (red) answered question. The t-tests reported in each panel 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 ($t(633) = 0.577$, $p = 0.564$). This reflects a floor effect: when knowledge is low everywhere, there is little signal to differentiate between what is known versus unknown. After watching the first lecture, estimated knowledge for held-out correctly answered questions (from the second quiz; Fig. 6, middle panel) exhibited a positive shift relative to held-out incorrectly answered questions ($t(633) = 3.961$, $p < 0.001$). This second quiz provides the maximally sensitive test for our knowledge predictions, since (if knowledge is estimated accurately) participants' Quiz 2 responses should demonstrate specific knowledge about Lecture 1 content, but knowledge about Lecture 2 and general physics concepts should be roughly unchanged from before they watched Lecture 1. After watching the second lecture, estimated knowledge (from the third quiz; Fig. 6, right panel) for *all* questions exhibited a positive shift. However, the estimated knowledge for held-out correctly answered questions remained greater than that for held-out incorrectly answered questions ($t(628) = 2.045$, $p = 0.041$). This third contrast reflects a ceiling effect: when knowledge is relatively high everywhere, the signal differentiating what is known versus unknown is relatively weak. Taken together, this set of analyses demonstrates that our knowledge prediction framework is most informative when participants exhibit variability in their knowledge of the content captured by the text embedding model.

Knowledge estimates need not be limited to the content of the lectures. As illustrated in Figure 7, our general approach to estimating knowledge from a small number of quiz questions may be extended 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 (up to) $k = 100$ topics. Conceptually, increasing the number of topics used by the model functions to increase the "resolution" of the embedding space, providing a greater ability to estimate knowledge for content that is highly similar to (but not precisely the same as) that contained in the two lectures. 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 projected each question into a shared text embedding space.

We projected the resulting 100-dimensional topic vectors (for each second of video and each quiz question) onto a shared 2-dimensional plane (see *Creating knowledge and learning map visualizations*). Next, we sampled points from a 100×100 grid of coordinates that evenly tiled a rectangle enclosing the 2D projections of the videos and questions. We used Equation 4 to estimate participants' knowledge at each of these 10,000 sampled locations, and averaged these estimates across participants to obtain an estimated average *knowledge map* (Fig. 7A). Intuitively, the knowledge map constructed from a given quiz's responses provides a visualization of how "much" participants knew about any content expressible by the fitted text embedding model at the point in time when they completed that quiz.

Several features of the resulting knowledge maps are worth noting. The average knowledge map estimated from Quiz 1 responses (Fig. 7A, 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 traversed 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.,

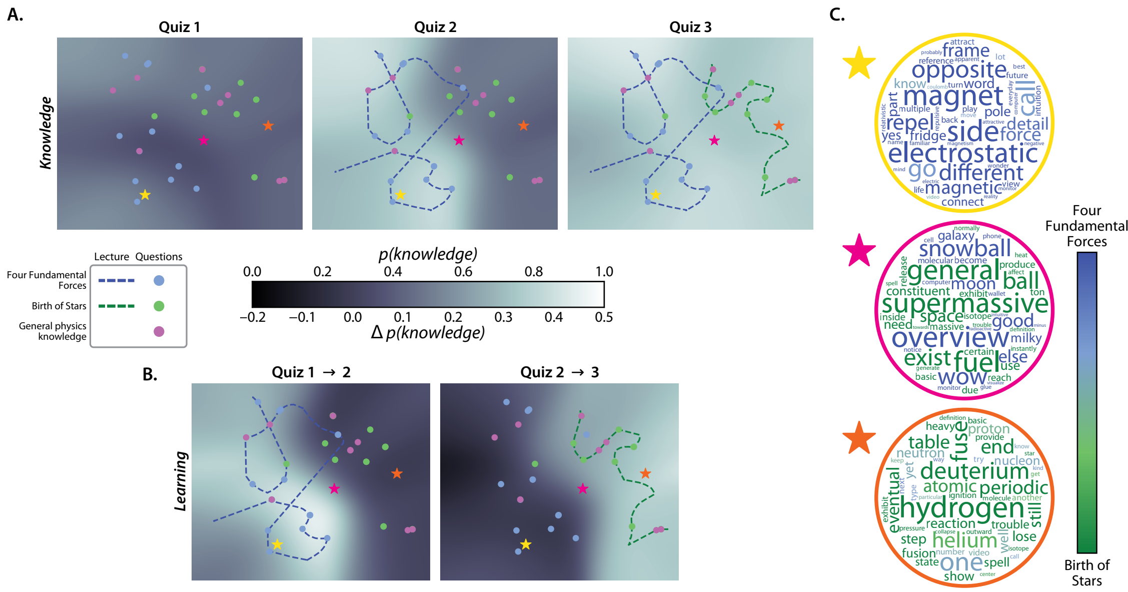

Figure 7: Mapping out the geometry of knowledge and learning. A. Average "knowledge maps" estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*). The topic trajectories of the two lectures are indicated by dotted lines (blue: Lecture 1; green: Lecture 2), and the coordinates of each question are indicated by dots (light blue: Lecture 1-related; light green: Lecture 2-related; purple: general physics knowledge). Each map reflects an average across all participants. For individual participants' maps, see Supplementary Figures 3, 4, and 5. **B. Average "learning maps" estimated between each successive pair of quizzes.** The learning maps follow the same general format as the knowledge maps in Panel A, but here the shading at each coordinate indicates the *difference* between the corresponding coordinates in the indicated *pair* of knowledge maps—i.e., how much the estimated knowledge "changed" between the two quizzes. Each map reflects an average across all participants. For individual participants' maps, see Supplementary Figures 6 and 7. **C. Word clouds for sampled points in topic space.** Each word cloud displays the weighted blend of words underlying the topic proportions represented at the corresponding colored star's location on the maps. In each word cloud, the words' relative sizes correspond to their relative weights at the starred location, and their colors indicate their relative weights in the *Four Fundamental Forces* (blue) versus *Birth of Stars* (green) lectures, on average, across all timepoints' topic vectors.

related to) the content from the lecture they watched prior to taking Quiz 2. This localization is non-trivial: these knowledge estimates are informed only by the embedded coordinates of the *quiz questions*, not by the embeddings of either lecture (see Eqn. 4). 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 immediately prior to taking Quiz 3.

Another way of visualizing these content-specific increases in knowledge after participants viewed each lecture is displayed in Figure 7B. 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 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 these maps' meaning by reconstructing the original high-dimensional topic vector for any location on the map 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. 7C): one point near the *Four Fundamental Forces* embedding (yellow), a second point near the *Birth of Stars* embedding (orange), and a third point 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 tended to be weighted more 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 more 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

We developed a computational framework that uses short multiple-choice quizzes to gain nuanced insights into what learners know and how their knowledge changes with training. First, we show that our approach can automatically match the conceptual knowledge probed by individual quiz questions to the corresponding moments in lecture videos when those concepts were presented (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment "knowledge traces" that reflect the degree of knowledge participants have about each video's time-varying content, and capture temporally specific increases in knowledge after viewing each lecture (Fig. 5). We also show that these knowledge estimates can generalize to held-out questions (Fig. 6). Finally, we use our framework to construct visual maps that provide snapshot estimates of how much participants know about any concept within the scope of our text embedding model, and how much their knowledge of those concepts changes with training (Fig. 7).

Over the past several years, the global pandemic has forced many educators to suddenly adapt to teaching remotely [26, 40, 49, 52]. This change in world circumstances is happening alongside (and perhaps accelerating) geometric growth in the availability of high-quality online courses from platforms such as Khan Academy [27], Coursera [53], EdX [29], and others [46]. Continued expansion of the global internet backbone and improvements in computing hardware have also facilitated improvements in video streaming, enabling videos to be easily shared and viewed by increasingly 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 defines an effective course or training program? Which aspects of teaching might be optimized and/or augmented by automated tools? How and why do learning needs and goals vary across people? How might we lower barriers of access to 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 [25]. 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 [41]. 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; 2, 13, 38, 42, 43) 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.

Beyond specifically predicting what people *know*, the fundamental ideas we develop here also relate to the notion of "theory of mind" of other individuals [19, 23, 37]. Considering others' unique perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and communicate [44, 48, 51]. One could imagine future extensions of our work (e.g., analogous to the knowledge and learning maps shown in Fig. 7), that attempt to characterize how well-aligned different people's knowledge bases or backgrounds are. In turn, this might be used to model how knowledge (or other forms of communicable information) flows not just between teachers and students, but between friends having a conversation, individuals on a first date, participants at a business meeting, doctors and patients, experts and non-experts, political allies or adversaries, and more. For example, the extent to which two people's knowledge maps "match" or "align" in a given region of text embedding space might serve as a predictor of how effectively they will be able to communicate about the corresponding conceptual content.

Ultimately, our work suggests a rich new line of questions about the geometric "form" of knowledge, how knowledge changes over time, and how we might map out the full space of what an individual knows. Our finding that detailed estimates about knowledge may be obtained from short quizzes shows one way that traditional approaches to evaluation in education may be extended. We hope that these advances might help pave the way for new approaches to 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 optional course credit for enrolling. We asked each participant to complete 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(s). 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 Calculus 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 (18 participants); Biology, AP Biology; or High school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed in our survey (5 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 (17 participants). Finally, we asked participants about in-person courses they had taken in different subject areas. They reported taking courses in Mathematics (38 participants), Science and engineering (37 participants), Arts and humanities (34 participants), Test preparation (27 participants), Economics and finance (26 participants), Computing (14 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 Supplementary Table 1.

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 (see Fig. 1). 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 exactly 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 orders of questions on each quiz, and the orders 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 this 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 [21] 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; 5), trained on a set of documents, to discover a set of (up to) k "topics" or "themes." Formally, each topic is defined as a distribution 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 semantically related or 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. To select an appropriate k for our model, we identified the minimum number of topics that yielded at least one "unused" topic (i.e., in which all words in the vocabulary were assigned uniform weights) after training. This indicated that the number of topics was sufficient to capture the set of latent themes present in the two lectures (from which we constructed our document corpus, as described below). We found this value to be $k = 15$ topics. The distribution of weights over words in the vocabulary for each discovered topic is shown in Supplementary Figure 1, and each topic's top-weighted words may be found in Supplementary Table 2.

As illustrated in Figure 2A, 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 to be of sufficiently near-human quality to yield reliable data in behavioral studies; 54) when developing our framework in order to make it more directly extensible and adaptable by others in the future.

We fetched these automated transcripts using the `youtube-transcript-api` Python package [12]. The transcripts consisted of one timestamped line of text for every few seconds (mean: 2.34 s; standard deviation: 0.83 s) 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 the timestamps for its first and last lines. These sliding windows ramped up and down in length at the beginning and end of each transcript, respectively. In other words, each transcript's first sliding window covered only its first line, the second sliding window covered the first two lines, and so on. This ensured that each line from the transcripts 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 combined these documents across the two videos' windows to create a single training corpus for the topic model.

After fitting a topic model to the two 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 correlation, cosine similarity, Kullback-Leibler divergence, Euclidean distance, 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 (see Supp. Tab. 1). Taken together, we obtained a *trajectory* for each video, describing its path through topic space, and a single coordinate for each question (Fig. 2C). 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 in the given lecture, and all five questions *about* that lecture appearing on the given quiz. 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 1 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-negative 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 (Fig. 7, Supp. Figs. 3, 4, 5, 6, and 7), we used Uniform Manifold Approximation and Projection (UMAP; 35, 36) 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 trivial.

At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing the cross-entropy between the pairwise (clustered) distances between the observations in their original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise distances in the original high-dimensional space were defined as 1 minus the correlation between each pair of coordinates, and pairwise distances in the low-dimensional embedding space were defined as the Euclidean distance between each pair of coordinates.

In our application, all of the coordinates we embedded were topic vectors, whose elements are always non-negative and sum to one. Although UMAP is an invertible transformation at the embedding locations of the original data, other locations in the embedding space will not necessarily follow the same implicit "rules" as the original high-dimensional data. For example, inverting an arbitrary coordinate in the embedding space might result in negative-valued vectors, which are incompatible with the topic modeling framework. To protect against this issue, we log-transformed the topic vectors prior to embedding them in the 2D space. When we inverted the embedded vectors (e.g., to estimate topic vectors or word clouds, as in Fig. 7C), we passed the inverted (log-transformed) values through the exponential function to obtain a vector of non-negative values, and normalized them to sum to one.

After embedding both lectures' topic trajectories and the topic vectors of every question, we defined a rectangle enclosing the 2D projections of the lectures' and quizzes' embeddings. We then sampled points from a regular 100×100 grid of coordinates that evenly tiled this 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:

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

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:

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

Intuitively, Equation 4 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 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.

References

[1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56:149–178.

[2] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.

[3] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in children: distinguishing response flexibility from conceptual flexibility; the protracted development of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.

[4] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing Machinery.

[5] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,

R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.

[7] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.

[8] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *arXiv*, 1803.11175.

[9] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.

[10] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004). Evidence for a new conceptualization of semantic representation in the left and right cerebral hemispheres. *Cortex*, 40(3):467–478.

[11] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

[12] Depoix, J. (2018). YouTube transcript API. `https://github.com/jdepoix/youtube-transcript-api`.

[13] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J., Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.

[14] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.

[15] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, 115:155–174.

[16] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222(602):309–368.

[17] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge. *School Science and Mathematics*, 100(6):310–318.

[18] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the "weather prediction" task? individual variability in strategies for probabilistic category learning. *Learning and Memory*, 9:408–418.

[19] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of Cognition and Development*, 13(1):19–37.

[20] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual learning, pages 212–221. Sage Publications.

[21] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behavioral and neural signatures of transforming naturalistic experiences into episodic memories. *Nature Human Behavior*, 5:905–919.

[22] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–4008.

[23] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating empathy and theory of mind. In *Social Behavior from Rodents to Humans*, pages 193–206. Springer.

[24] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*. Columbia University Press.

[25] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*, 326(7382):213–216.

[26] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021). Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International Journal of Environmental Research and Public Health*, 18(5):2672.

[27] Khan, S. (2004). *The Khan Academy*. Salman Khan.

[28] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.

[29] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses. *The Chronicle of Higher Education*, 21:1–5.

[30] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

[31] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467–022–31965–2.

[32] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of Educational Studies*, 53(2):129–147.

[33] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook of Human Memory*. Oxford University Press.

[34] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum "memory wave" function? *Psychological Review*, 128(4):711–725.

[35] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 1802(03426).

[36] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

[37] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of mind. In *The Wiley-Blackwell handbook of childhood cognitive development*. Wiley-Blackwell.

[38] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson, U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467–021–22202–3.

[39] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*, 1301.3781.

[40] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications from a national survey of language educators. *System*, 97:102431.

[41] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.

[42] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022). Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective Neuroscience*, 17(4):367–376.

[43] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*, 7:43916.

[44] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.

[45] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning. *Biological Cybernetics*, 45(1):35–41.

[46] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in higher education: unmasking power and raising questions about the movement's democratic potential. *Educational Theory*, 63(1):87–110.

[47] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter Student conceptions and conceptual learning in science. Routledge.

[48] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-based empathy training improves the communication skills of neonatal nurses. *Clinical Simulation in Nursing*, 22:32–42.

[49] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching during COVID-19. *Children and Youth Services Review*, 119:105578.

[50] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in Mathematics Education*, 35(5):305–329.

[51] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal Medicine*, 21:524–530.

[52] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.

[53] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from free courses. *The Chronicle of Higher Education*, 19(7):1–4.

[54] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods*, 50:2597–2605.