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 [6, 11, 13, 15, 30, 64]? Or weaving a lecture's atomic elements (e.g., its component words) into a structured network that describes how those individual elements are related [40, 69]? 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 [37, 41, 61, 68].

The difference between "understanding" and "memorizing," as framed by researchers in education, cognitive psychology, and cognitive neuroscience [e.g., 23, 28, 33, 41, 61], 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., 7, 8, 10, 12, 16, 39, 50, 70] 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 (in many cases) 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 [42, 43]. To be clear, this is not to say that text embedding models themselves are capable of "understanding" deep conceptual meaning in any traditional sense. But rather, their ability to capture the underlying *structure* of text documents beyond their surface-level contents provides a computational framework through which those documents' deeper conceptual meanings may be quantified, explored, and understood. 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 limited overlap*. In this way, "concepts" are defined implicitly by the model's geometry [e.g., how the embedding coordinate of a given word or document relates to the coordinates of other text embeddings; 55].

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 (at least some subset of) 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, 20, 21, 24, 31, 58]. 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 [28, 41]. 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.

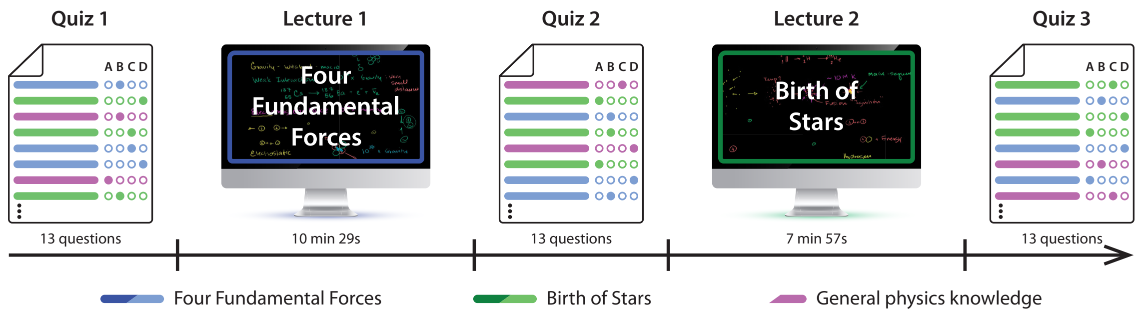

Figure 1: Experimental paradigm. Participants alternate between completing three 13-question multiple-choice quizzes and watching two Khan Academy lectures. Each quiz contains a mix of 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general physics knowledge. The specific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

We asked participants in our study to complete brief multiple-choice quizzes before, between, and after watching two lecture videos from the Khan Academy [36] 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 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

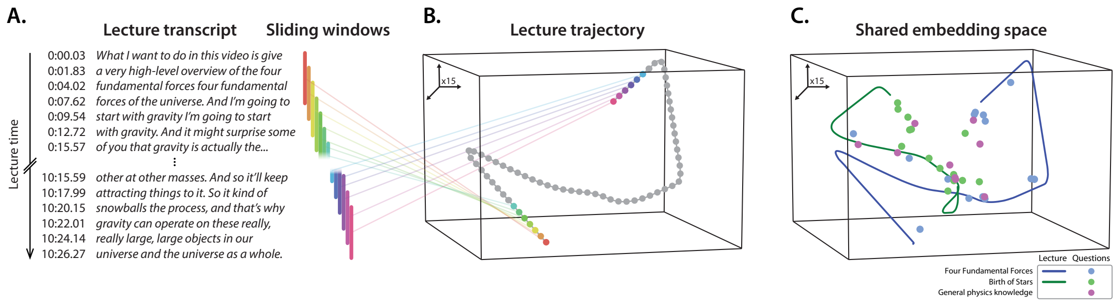

Figure 2: Modeling course content. A. Building a document pool from sliding windows of text. We decompose each lecture's transcript into a series of overlapping sliding windows. The full set of transcript snippets (across all windows) may be treated as a set of "documents" for training a text embedding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

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).

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 [29] to identify the latent themes in the lectures using a topic model [8]. 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 often used different words (Supp. Fig. 5) and phrasings. 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 the average topic weights from 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

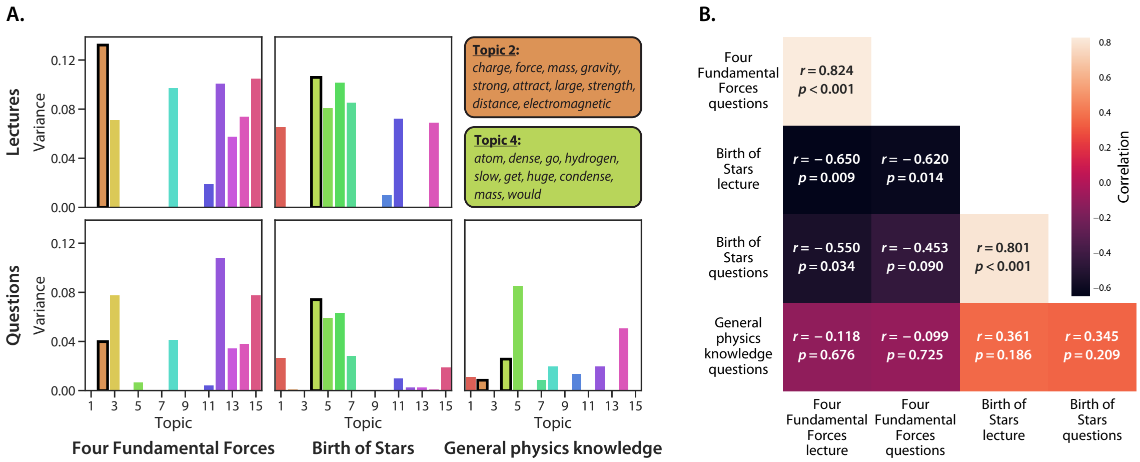

Figure 3: Lecture and question topic overlap. A. Topic weight variability. The bar plots display the variance of each topic's weight across lecture timepoints (top row) and questions (bottom row); colors denote topics. The top-weighted words from the most "expressive" (i.e., variable across observations) 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.

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" [22] 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 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 variabilities 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 variabilities 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 (Supp. Figs. 3, 4). We also qualitatively examined the best-matching intervals for each question by comparing the question's text to the transcribed text from the most-correlated parts of the lectures (Supp. Tab. 3). 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 administering some form of assessment (e.g., a quiz) and computing the proportion of correctly answered questions. But if two students receive identical scores on such an exam, might our

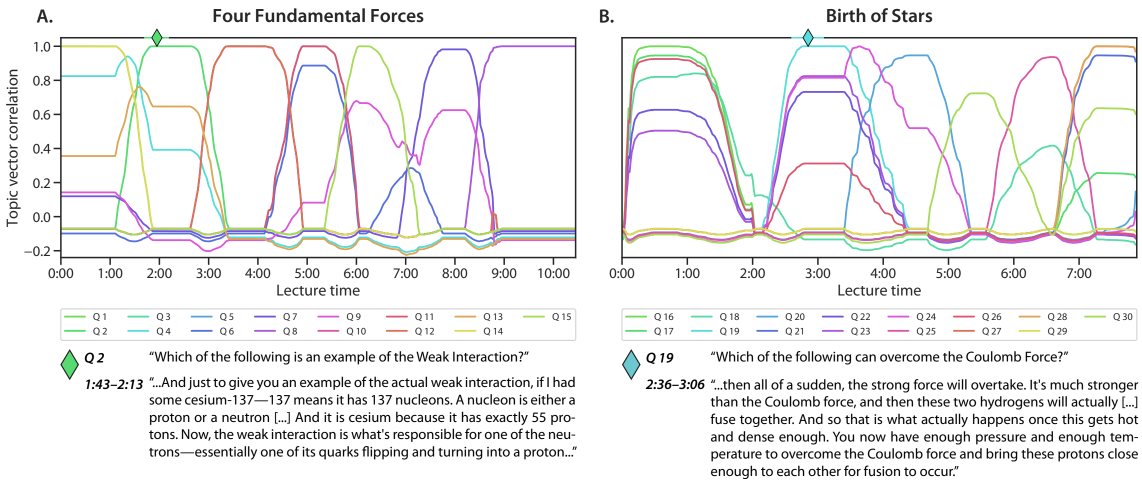

Figure 4: Which parts of each lecture are captured by each question? Each panel displays time series plots showing how each question's topic vector correlates with each video timepoint's topic vector (Panel **A.**: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel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

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. 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 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 time course describing how much "knowledge" that participant has about the content presented at 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 throughout 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 time courses in Figure 5A and C provide detailed *estimates* about participants' knowlege, 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 upon watching it, 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 *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

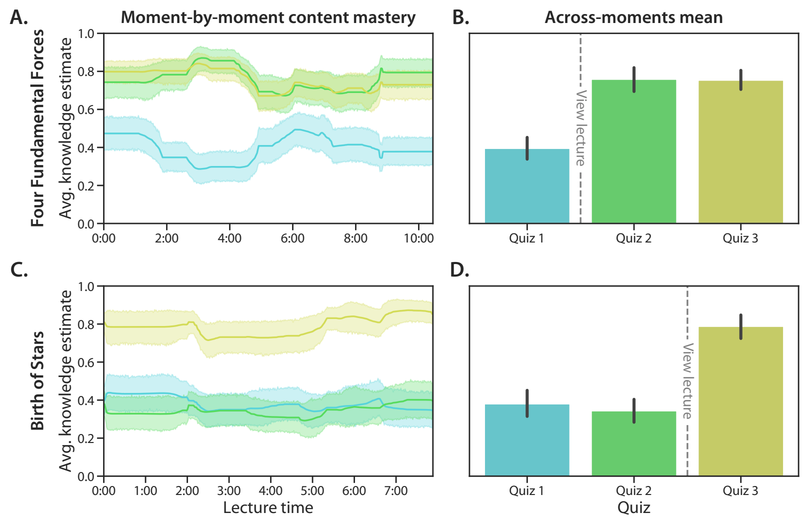

Figure 5: Estimating knowledge about the content presented at each moment of each lecture. A. Knowledge about the time-varying content of *Four Fundamental Forces*. Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (see *Estimating dynamic knowledge traces*), using responses from a single quiz (color). The traces are averaged across participants. **B. Average estimated knowledge about *Four Fundamental Forces*.** Each bar displays the across-timepoint average knowledge, estimated using the responses to one quiz's questions. **C. Knowledge about the time-varying content of *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 *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.

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 they are likely to answer that question correctly or incorrectly. We developed a statistical approach to test this claim. For each quiz question a participant answered, in turn, we used Equation 1 to estimate their knowledge at the given question's embedding space coordinate based on other questions that participant answered on the same quiz. We repeated this for all participants, and for each of the three quizzes. Then, separately for each quiz, we fit a generalized linear mixed model (GLMM) with a logistic link function to explain the probability of correctly answering a question as a function of estimated knowledge for its embedding coordinate, while accounting for random variation among participants and questions (see *Generalized linear mixed models*). To assess the predictive value of the knowledge estimates, we compared each GLMM to an analogous (i.e., nested) "null" model that did not consider estimated knowledge using parametric bootstrap likelihood-ratio tests.

We carried out three different versions of the analyses described above, wherein we considered different sources of information in our estimates of participants' knowledge for each quiz question. First, we estimated knowledge at each question's embedding coordinate using *all* other questions answered by the same participant on the same quiz ("All questions"; Fig. 6, top row). This test was intended to assess the overall predictive power of our approach. Second, we estimated knowledge for each question about a given lecture using only the other questions (from the same participant and quiz) about that *same* lecture ("Within-lecture"; Fig. 6, middle rows). This test was intended to assess the *specificity* of our approach by asking whether our predictions could distinguish between questions about different content covered by the same lecture. Third, we estimated knowledge for each question about one lecture using only the questions (from the same participant and

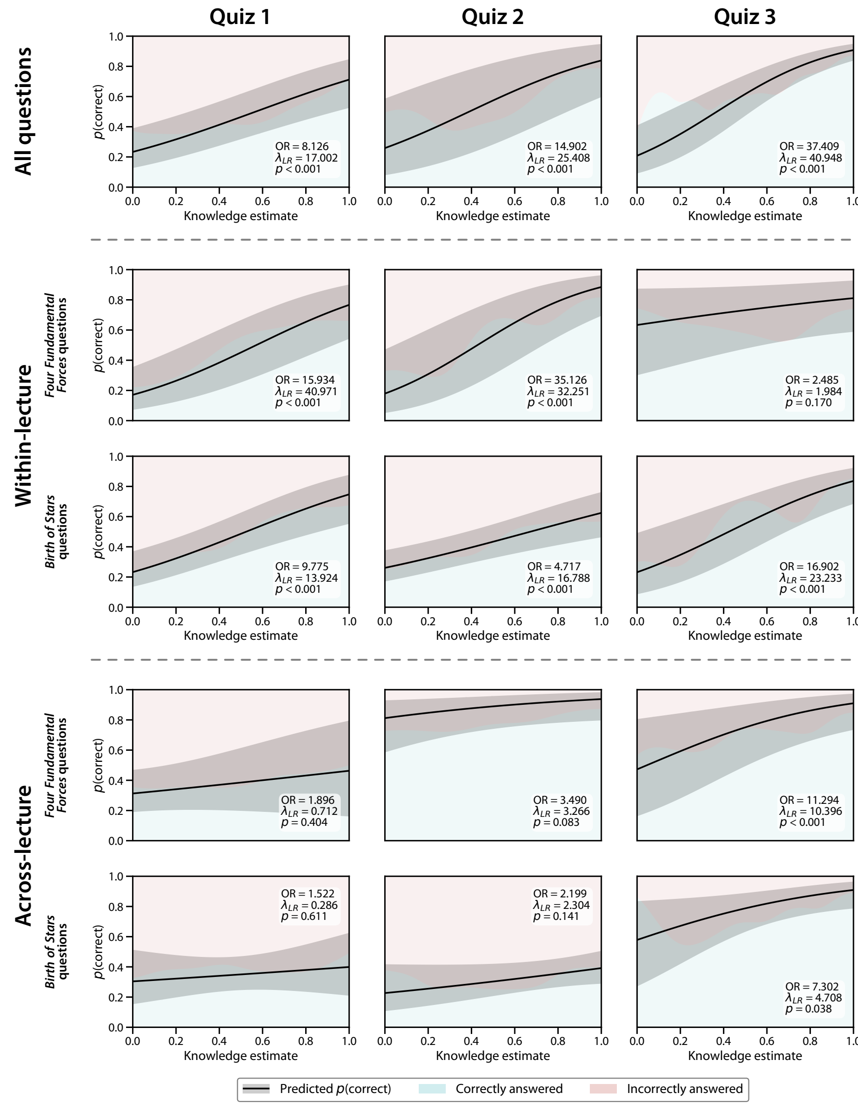

Figure 6: Predicting success on held-out questions using estimated knowledge. We used generalized linear mixed models (GLMMs) to model the likelihood of correctly answering a quiz question as a function of estimated knowledge for its embedding coordinate (see *Generalized linear mixed models*). Separately for each quiz (column), we examined this relationship based on three different sets of knowledge estimates: knowledge for each question based on all other questions the same participant answered on the same quiz ("All questions"; top row), knowledge for each question about one lecture based on all other questions (from the same participant and quiz) about the *same* lecture ("Within-lecture"; middle rows), and knowledge for each question about one lecture based on all questions (from the same participant and quiz) about the *other* lecture ("Across-lecture"; bottom rows). The backgrounds in each panel display kernel density estimates of the relative observed proportions of correctly (blue) versus incorrectly (red) answered questions, for each level of estimated knowledge along the x-axis. The black curves display the (population-level) GLMM-predicted probabilities of correctly answering a question as a function of estimated knowledge. Error ribbons denote 95% confidence intervals.

quiz) about the *other* lecture ("Across-lecture"; Fig. 6, bottom rows). This test was intended to assess the *generalizability* of our approach by asking whether our predictions could extend across the content areas of the two lectures. When computing these knowledge estimates, we used a rebalancing procedure to ensure that (for a given participant and quiz) the knowledge estimates for correctly and incorrectly answered questions were computed from the same proportion of correctly answered questions (see *Generalized linear mixed models*).

When we fit a GLMM to estimates of participants' knowledge for each Quiz 1 question based on all other Quiz 1 questions, we found that higher estimated knowledge for a given question predicted a greater likelihood of answering it correctly (odds ratio (OR) = 8.126, 95% CI = [3.116, 20.123], likelihood-ratio test statistic (λ_{LR}) = 17.002, $p < 0.001$). This relationship held when we repeated this analysis for Quiz 2 (OR = 14.902, 95% CI = [4.976, 39.807], λ_{LR} = 25.408, $p < 0.001$) and again for Quiz 3 (OR = 37.409, 95% CI = [10.425, 107.145], λ_{LR} = 40.948, $p < 0.001$). Taken together, these results suggest that our knowledge estimates can reliably predict participants' performance on individual questions when aggregated across all quiz content.

We observed a similar set of results when we restricted our estimates of participants' knowledge for questions about each lecture to consider only their performance on other questions about the *same* lecture. Specifically, for Quiz 1, participants' knowledge of *Four Fundamental Forces*-related questions, estimated from their performance on other *Four Fundamental Forces*-related questions, was predictive of their ability to answer those questions correctly (OR = 15.934, 95% CI = [5.173, 38.005], λ_{LR} = 40.971, $p = 0.001$). The same was true of participants' estimated knowledge for *Birth of Stars*-related questions based on their performance on other *Birth of Stars*-related questions (OR = 9.775, 95% CI = [2.93, 25.08], λ_{LR} = 13.924, $p = 0.001$). These within-lecture knowledge estimates also predicted success on questions about both lectures when we computed them analogously for Quiz 2 (*Four Fundamental Forces*: OR = 35.126, 95% CI = [5.113, 123.868], λ_{LR} = 32.251, $p < 0.001$; *Birth of Stars*: OR = 4.717, 95% CI = [2.021, 9.844], λ_{LR} = 16.788, $p < 0.001$). For Quiz 3, we found that within-lecture knowledge estimates predicted participants' success on *Birth of Stars*-related questions (OR = 16.902, 95% CI = [3.353, 53.265], λ_{LR} = 23.233, $p < 0.001$) but not on *Four Fundamental Forces*-related questions (OR = 2.485, 95% CI = [0.724, 8.366], λ_{LR} =

1.984, p = 0.170). This may indicate that the within-lecture knowledge estimates are susceptible to ceiling effects in participants' quiz performance. On Quiz 3, after viewing both lectures, no participant answered more than three *Four Fundamental Forces*-related questions incorrectly, and all but five participants (out of 50) answered two or fewer questions incorrectly. (This was the only subset of questions about either lecture, across all three quizzes, for which this was true.) Because of this, when we held out one incorrectly answered *Four Fundamental Forces* question from a given participant's Quiz 3 responses and estimated their knowledge at its embedding coordinate using the remaining *Four Fundamental Forces* questions they answered, for 90% of participants, that estimate leveraged information about at most a single other question they were *not* able to correctly answer. This broad homogeneity in participants' success on questions used to estimate their knowledge may have hurt our ability to accurately characterize the specific (and by Quiz 3, relatively few) aspects of the lecture content they did *not* know about. Taken together, these results suggest that our knowledge estimates can reliably distinguish between questions about different content covered by a single lecture, provided there is sufficient diversity in participants' quiz responses to extract meaningful information about both what they know and what they do not know.

Finally, when we estimated participants' knowledge for each question about one lecture using their performance on questions (from the same quiz) about the *other* lecture, we observed a somewhat different pattern of results. Here we found that before viewing either lecture (i.e., on Quiz 1), participants' abilities to answer *Four Fundamental Forces*-related questions could not be predicted from their responses to *Birth of Stars*-related questions (OR = 1.896, 95% CI = [0.419, 9.088], λ_{LR} = 0.712, p = 0.404), nor could their abilities to answer *Birth of Stars*-related questions be predicted from their responses to *Four Fundamental Forces*-related questions (OR = 1.522, 95% CI = [0.332, 6.835], λ_{LR} = 0.286, p = 0.611). We similarly found that participants' success on questions about either lecture could not be predicted given their responses to questions about the other lecture after viewing *Four Fundamental Forces* but before viewing *Birth of Stars* (i.e., on Quiz 2; *Four Fundamental Forces* questions given *Birth of Stars* questions: OR = 3.49, 95% CI = [0.739, 12.849], λ_{LR} = 3.266, p = 0.083; *Birth of Stars* questions given *Four Fundamental Forces* questions: OR = 2.199, 95% CI = [0.711, 5.623], λ_{LR} = 2.304, p = 0.141). Only after viewing *both* lectures (i.e., on Quiz 3) did these across-lecture knowledge estimates reliably predict participants' success on individual quiz questions (*Four Fundamental Forces* questions given *Birth of Stars* questions: OR = 11.294, 95% CI = [1.375, 47.744], λ_{LR} = 10.396, $p < 0.001$; *Birth of Stars* questions given *Four Fundamental Forces* questions: OR = 7.302, 95% CI = [1.077, 44.879], λ_{LR} = 4.708, p = 0.038). Taken together, these results suggest that our knowledge estimates can be used to predict participants' success across content areas once they have received some training on both the content about which their knowledge is estimated and the content used to construct these estimates.

That the knowledge predictions derived from the text embedding space reliably distinguish between held-out correctly versus incorrectly answered questions (Fig. 6) suggests that spatial relationships within this space can help explain what participants know. But how far does this explanatory power extend? For example, suppose we know that a participant correctly answered a question at embedding coordinate x. As we move farther away from x in the embedding space, how does the likelihood that the participant knows about the content at a given location "fall off" with distance? Conversely, suppose the participant instead answered that same question *in*correctly. Again, as we move farther away from x in the embedding space, how does the likelihood that the participant does *not* know about a coordinate's content change with distance? We reasoned that, assuming our embedding space is capturing something about how individuals actually organize their knowledge, a participant's ability to answer questions embedded very close to x should tend to be similar to their ability to answer the question embedded *at* x. Whereas at another extreme, once we reach some sufficiently large distance from x, our ability to infer whether or not a participant will correctly answer a question based on their ability to answer the question at x should be no better than guessing based on their *overall* proportion of correctly answered questions. In other words, beyond the maximum distance at which the participant's ability to answer the question at x is informative of their ability to answer a second question at location y, then guessing the outcome at y based on x should be no more successful than guessing based on a measure that does not consider embedding space distance.

With these ideas in mind, we asked: conditioned on answering a question correctly, what

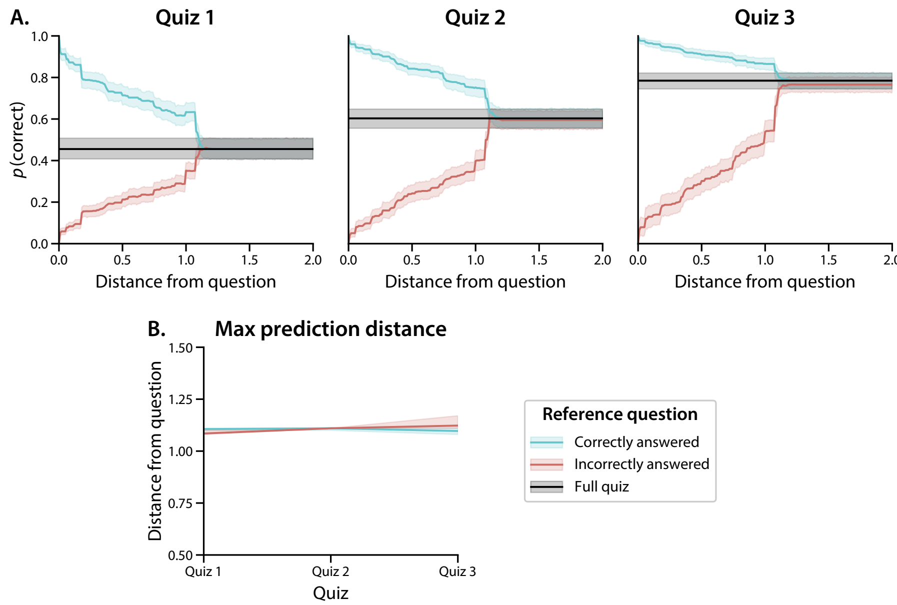

Figure 7: Knowledge falls off gradually in text embedding space. A. Performance versus distance. For each participant, for each correctly answered question (blue) or incorrectly answered question (red), we computed the proportion of correctly answered questions within a given distance of that question's embedding coordinate. We used these proportions as a proxy for participants' knowledge about the content within that region of the embedding space. We repeated this analysis for all questions and participants, and separately for each quiz (column). The black lines denote the average proportion correct across *all* questions included in the analysis at the given distance. **B. Maximum distance for which performance is reliably different from the average.** We used a bootstrap procedure (see *Estimating the "smoothness" of knowledge*) to estimate the point at which the blue and red lines in Panel A reliably diverged from the black line. We repeated this analysis separately for correctly and incorrectly answered questions from each quiz. **All panels.** Error ribbons denote bootstrap-estimated 95% confidence intervals.

proportion of all questions (within some radius, r, of that question's embedding coordinate) were answered correctly? We plotted this proportion as a function of r. Similarly, we could ask, conditioned on answering a question incorrectly, how the proportion of correct responses changed with r. As shown in Figure 7, we found that quiz performance falls off smoothly with distance, and the "rate" of the falloff does not appear to change across the different quizzes, as measured by the distance at which performance becomes statistically indistinguishable from a simple proportion-correct score (see *Estimating the "smoothness" of knowledge*). This suggests that, at least within the region of text embedding space covered by the questions our participants answered (and as characterized using our topic model), the rate at which knowledge changes with distance is relatively constant, even as participants' overall level of knowledge varies across quizzes or regions of the embedding space.

Knowledge estimates need not be limited to the content of the lectures. As illustrated in Figure 8, 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 $k = 100$ 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. We note that we used these 2D maps solely for visualization; all relevant comparisons, distance computations, and statistical tests we report above were carried out in the original 15-dimensional space, using the 15-topic model. 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. 8A). 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. 8A, leftmost map) shows that participants tended to have relatively little knowledge about any parts of the text embedding space (i.e., the shading is

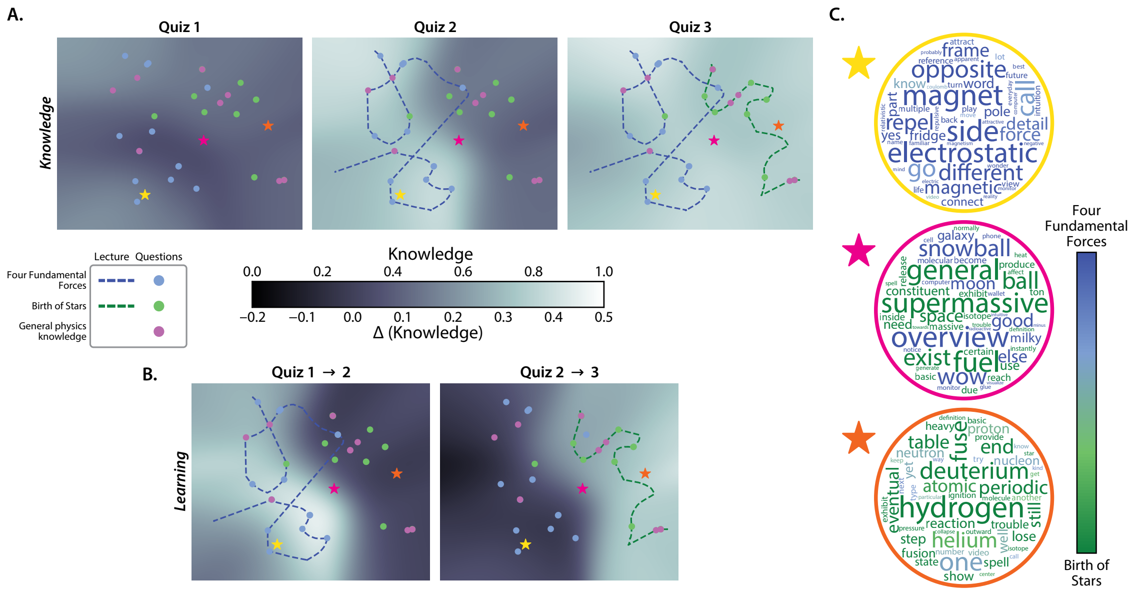

Figure 8: Mapping out the geometry of knowledge and learning. A. Average "knowledge maps" estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*). The 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 7, 8, and 9. **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 10 and 11. **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 *Four Fundamental Forces* (blue) versus *Birth of Stars* (green) lectures, on average, across all timepoints' topic vectors.

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., 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 8B. 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' meanings 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. 8C): 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. 8).

We view our work as making several contributions to the study of how people acquire conceptual knowledge. First, from a methodological standpoint, our modeling framework provides a systematic means of mapping out and characterizing knowledge in maps that have infinite (arbitrarily many) numbers of coordinates, and of "filling out" those maps using relatively small numbers of multiple choice quiz questions. Our experimental finding that we can use these maps to predict responses to held-out questions has several psychological implications as well. For example, concepts that are assigned to nearby coordinates by the text embedding model also appear to be "known to a similar extent" (as reflected by participants' responses to held-out questions; Fig. 6). This suggests that participants also *conceptualize* similarly the content reflected by nearby embedding coordinates. How participants' knowledge falls off with spatial distance is captured by the knowledge maps we infer from their quiz responses (e.g., Figs. 7, 8). In other words, our study shows that knowledge about a given concept implies knowledge about related concepts, and we also show how estimated knowledge falls off with distance in text embedding space.

In our study, we characterize the "coordinates" of participants' knowledge using a relatively simple "bag of words" text embedding model [LDA; 8]. More sophisticated text embedding models, such as transformer-based models [18, 54, 67, 70] can learn complex grammatical and semantic relationships between words, higher-order syntactic structures, stylistic features, and more. We considered using transformer-based models in our study, but we found that the text embeddings derived from these models were surprisingly uninformative with respect to differentiating or otherwise characterizing the conceptual content of the lectures and questions we used. We suspect that this reflects a broader challenge in constructing models that are high-resolution within a given domain (e.g., the domain of physics lectures and questions) *and* sufficiently broad so as to enable them to cover a wide range of domains. For example, we found that the embeddings derived even from much larger and more modern models like BERT [18], GPT [70], LLaMa [67], and others that are trained on enormous text corpora, end up yielding poor resolution within the content space spanned by individual course videos (Supp. Fig. 6). Whereas the LDA embeddings of the lectures and questions are "near" each other (i.e., the convex hull enclosing the two lectures' trajectories is highly overlapping with the convex hull enclosing the questions' embeddings), the BERT embeddings of the lectures and questions are instead largely distinct (top row of Supp. Fig. 6). The LDA embeddings of the questions for each lecture and the corresponding lecture's trajectory are also similar. For example, as shown in Fig. 2C, the LDA embeddings for *Four Fundamental Forces* questions (blue dots) appear closer to the *Four Fundamental Forces* lecture trajectory (blue line), whereas the LDA embeddings for *Birth of Stars* questions (green dots) appear closer to the *Birth of Stars* lecture trajectory (green line). The BERT embeddings of the lectures and questions do not show this property (Supp. Fig. 6). We also examined per-question "content matches" between individual questions and individual moments of each lecture (Figs. 4, 6). The time series plot of individual questions' correlations are different from each other when computed using LDA (e.g., the traces can be clearly visually separated), whereas the correlations computed from BERT embeddings of different questions all look very similar. This tells us that LDA is capturing some differences in content between the questions, whereas BERT is not. The time series plots of individual questions' correlations have clear "peaks" when computed using LDA, but not when computed using BERT. This tells us that LDA is capturing a "match" between the content of each question and a relatively well-defined time window of the corresponding lectures. The BERT embeddings appear to blur together the content of the questions versus specific moments of each lecture. Finally, we also compared the pairwise correlations between embeddings of questions within versus across content areas (i.e., content covered by the individual lectures, lecture-specific questions, and by the "general physics knowledge" questions). The LDA embeddings show a strong contrast between same-content embeddings versus across-content embeddings. In other words, the embeddings of questions about the *Four Fundamental Forces* material are highly correlated with the embeddings of the *Four Fundamental Forces* lecture, but not with the embeddings of *Birth of Stars*, questions about *Birth of Stars*, or general physics knowledge questions. We see a similar pattern with the LDA embeddings of the *Birth of Stars* questions (Fig. 3, Supp. Fig. 2). In contrast, the BERT embeddings are all highly correlated with each other (Supp. Fig. 6). Taken together, these comparisons illustrate how LDA (trained on the specific content in question) provides both coverage of the requisite material and specificity at the level of the content covered by individual questions. BERT, on the other hand, essentially assigns both lectures and all of the questions (which are all broadly about "physics") into a tiny region of its embedding space, thereby blurring out meaningful distinctions between different specific concepts covered by the lectures and questions. We note that these are not criticisms of BERT (or other large language models trained on large and diverse corpora). Rather, our point is that simple fine-tuned models trained on a relatively small but specialized corpus can outperform much more complicated models trained on much larger corpora, when we are specifically interested in capturing subtle conceptual differences at the level of a single course lecture or question. Of course if our goal had been to find a model that generalized to many different content areas, we would expect our approach to perform comparatively poorly relative to BERT or other much larger models. We suggest that bridging the tradeoff between high resolution within each content area versus the ability to generalize to many different content areas will be an important challenge for future work in this domain.

Another application for large language models that does *not* require explicitly modeling the content of individual lectures or questions is to leverage the models' abilities to generate text. For example, generative text models like ChatGPT [54] and LLaMa [67] are already being used to build a new generation of interactive tutoring systems [e.g., 44]. Unlike the approach we have taken here, these generative text model-based systems do not explicitly model what learners know, or how their knowledge changes over time with training. One could imagine building a hybrid system that combines the best of both worlds: a large language model that can *generate* text, combined with a smaller model that can *infer* what learners know and how their knowledge changes over time. Such a hybrid system could potentially be used to build the next generation of interactive tutoring systems that are able to adapt to learners' needs in real time, and that are able to provide more nuanced feedback about what learners know and what they do not know.

At the opposite end of the spectrum from large language models, one could also imagine *simplifying* some aspects of our LDA-based approach by computing simple word overlap metrics. For example, the Jaccard similarity between text A and B is computed as the number of unique words in the intersection of words from A and B divided by the number of unique words in the union of words from A and B. In a supplementary analysis (Supp. Fig. 5), we compared the LDA-based question-lecture matches we reported in Figure 4 with the Jaccard similarities between each question and each sliding window of text from the corresponding lecture. As shown in Supplementary Figure 5, this simple word-matching approach does not appear to capture the same level of specificity as the LDA-based approach. Whereas the LDA-based approach often yields a clear peak in the time series of correlations between each question and the corresponding lecture, the Jaccard similarity-based approach does not. Furthermore, these LDA-based matches appear to capture conceptual overlaps between the questions and lectures (Supp. Tab. 3), whereas simple word matching does not. For example, one of the example questions examined in Supplementary Figure 5 asks "Which of the following occurs as a cloud of atoms gets more dense?" The LDA-based matches identify lecture timepoints where the relevant *topics* are discussed (e.g., when words like "cloud," "atom," "dense," etc., are mentioned *together*). The Jaccard similarity-based matches, on the other hand, are strong when *any* of these words are mentioned, even if they do not occur together.

We view our approach as occupying a sort of "sweet spot," between much larger language models and simple word matching-based approaches, that enables us to capture the relevant conceptual content of course materials at an appropriate semantic scale. Our approach enables us to accurately and consistently identify each question's content in a way that also matches up with what is presented in the lectures. In turn, this enables us to construct accurate predictions about participants' knowledge of the conceptual content tested by held-out questions (Fig. 6).

One limitation of our approach is that topic models contain no explicit internal representations of more complex aspects of "knowledge," like knowledge graphs, dependencies or associations between concepts, causality, and so on. These representations might (in principle) be added as extensions to our approach to more accurately and precisely capture, characterize, and track learners' knowledge. However, modeling these aspects of knowledge will likely require substantial additional research effort.

Within the past several years, the global pandemic forced many educators to suddenly adapt to teaching remotely [35, 51, 63, 71]. 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 [36], Coursera [72], EdX [38], and others [59]. 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 to receiving 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 [34]. 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 [52]. 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; 4, 19, 49, 53, 56] 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 [26, 32, 48]. Considering others' unique perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and communicate [57, 62, 66]. One could imagine future extensions of our work (e.g., analogous to the knowledge and learning maps shown in Fig. 8), 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). All participants viewed the videos in the same order (i.e., *Four Fundamental Forces* followed by *Birth of Stars*).

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). To help broaden the set of lecture-specific questions, our team worked through each lecture in small segments to identify what each segment was "about" conceptually, and then write a question about that concept. The general physics questions were drawn our team's prior coursework and areas of interest, along with internet searches and brainstorming with the project team and other members of J.R.M.'s lab. Although we attempted to design the questions to test "conceptual knowledge," we note that estimating the specific "amount" of conceptual understanding that each question "requires" to answer is somewhat subjective, and might even come down to the "strategy" a given participant uses to answer the question at that particular moment. The full set of questions and answer choices may be found in Supplementary Table 1. The final set of questions (and response options) was reviewed and approved by J.R.M. before we collected or analyzed the text or experimental data.

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 contained 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. We obtained informed consent from all participants, and 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

Statistics

All of the statistical tests performed in our study were two-sided. The 95% confidence intervals we reported for each correlation were estimated by generating 10,000 bootstrap distributions of correlation coefficients by sampling (with replacement) from the observed data.

Constructing text embeddings of multiple lectures and questions

We adapted an approach we developed in prior work [29] 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; 8] trained on a set of documents, to discover a set of k "topics" or "themes." Formally, each topic is defined as a distribution of weights over words 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 (as inferred from their tendency to co-occur in the same document). 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, as a starting point, 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. We found that with a limited number of additional adjustments following Boyd-Graber et al. [9], such as removing corpus-specific stop-words, the model yielded (subjectively) sensible and coherent 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; 73] 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 [17]. 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. This w parameter was chosen to match the same number of words per sliding window (rounded to the nearest whole word, and before preprocessing) as the sliding windows we defined in our prior work [29; i.e., 185 words per sliding window].

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. We next performed a series of standard text preprocessing steps: normalizing case, lemmatizing, removing punctuation and removing stop-words. We constructed our corpus of stop words by augmenting the Natural Language Toolkit [NLTK; 5] English stop word list with the following additional words, selected using one of the approaches suggested by Boyd-Graber et al. [9]: "actual," "actually," "also," "bit," "could," "e," "even," "first," "follow," "following," "four," "let," "like," "mc," "really,", "saw," "see," "seen," "thing," and "two." This yielded sliding windows with an average of 73.8 remaining words, and lasting for an average of 62.22 seconds. 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 time series 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 Q used to estimate the knowledge trace. 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) ensures that every question contributes some non-negative amount to the knowledge estimate.

Generalized linear mixed models

In the set of analyses reported in Figure 6, we assessed whether estimates of participants' knowledge at the embedding coordinates of individual quiz questions could be used to reliably predict their abilities to correctly answer those questions. In essence, we treated each question a given participant answered on a given quiz as a "lecture" consisting of a single timepoint, and used Equation 1 to estimate the participant's knowledge for its embedding coordinate based on their performance on all *other* questions they answered on that same quiz ("All questions"; Fig. 6, top row). Additionally, for each lecture-related question (i.e., excluding questions about general physics knowledge), we computed analogous knowledge estimates based on two different subsets of questions the participant answered on the same quiz: (1) all *other* questions about the same lecture as the target question ("Within-lecture"; Fig. 6, middle rows), and (2) all questions about the other of the two lectures ("Across-lecture"; Fig. 6, bottom rows).

In performing these analyses, our null hypothesis is that the knowledge estimates we compute based on the quiz questions' embedding coordinates do *not* provide useful information about participants' abilities to answer those questions—in other words, that there is no meaningful difference (on average) between the knowledge estimates we compute for questions participants answered correctly and those they answered incorrectly. Specifically, since we estimate knowledge for a given embedding coordinate as a weighted proportion-correct score (where each question's weight reflects its embedding-space distance from the target coordinate; see Eqn. 1), if these weights are uninformative (e.g., randomly distributed), then our estimates of participants' knowledge should be equivalent (on average) to the *unweighted* proportion of correctly answered questions used to compute them. In general, for a given participant and quiz, this expected value (i.e., that participant's proportion-correct score on that quiz) is the same for any coordinate in the embedding space (e.g., any lecture timepoint, quiz question, etc.). However, in the "All questions" and "Within-lecture" versions of the analyses shown in Figure 6, we estimate each participant's knowledge for each target question using all *other* questions (or all *other* questions about the same lecture) they answered on the same quiz. This introduces a systematic dependency between a participant's success on a target question and their proportion-correct score on the remaining questions available to estimate their knowledge for it. For example, suppose a participant correctly answered n out of q questions on a given quiz. If we hold out a single *correctly* answered question as the target, the proportion of remaining questions answered correctly would be $\frac{n-1}{q-1}$. Whereas if we hold out a single *incorrectly* answered question, the proportion of remaining questions answered correctly would be $\frac{n}{q-1}$. Thus, the proportion of correctly answered remaining questions (and therefore the null-hypothesized value of a knowledge estimate computed from them) is always *lower* for target questions a participant answered correctly than for those they answered incorrectly.

To correct for this baseline inverse relationship between a participant's success on a target question and their estimated knowledge for it, we used a rebalancing procedure that ensured our knowledge estimates for questions each participant answered correctly and incorrectly were computed from the *same* proportion of correctly answered questions. For each target question on a given participant's quiz, we identified all remaining questions with the opposite "correctness" label (i.e., if the target question was answered correctly, we identified all remaining incorrectly answered questions, and vice versa). We then held out each of these opposite-label questions, in turn, along with the target question, and estimated the participant's knowledge for the target question using all *other* remaining questions. Since each of these subsets of remaining questions was constructed by holding out one correctly answered question and one incorrectly answered question from the participant's quiz, if the participant correctly answered n out of q questions total, then their proportion-correct score on each subset of questions used to estimate their knowledge for the target question is $\frac{n-1}{q-2}$, regardless of whether they answered the target question correctly or incorrectly. Finally, averaging over these per-subset knowledge estimates yielded a rebalanced estimate of the participant's knowledge for the target question that leveraged information from all remaining questions' embedding coordinates, but whose expected value under our null hypothesis was the same as that of each individual subset ($\frac{n-1}{q-2}$). By equalizing the null-hypothesized values of knowledge estimates for correctly and incorrectly answered questions, this procedure ensures that any meaningful relationships we observe between participants' estimated knowledge for individual quiz questions and their abilities to correctly answer them are attributable to the predictive power of the embedding-space distances used to weight questions' contributions to the knowledge estimates, rather than an artifact of our estimation procedure. Note that if a participant answered all or no questions on a given quiz correctly, their responses contained no opposite-label questions with which to perform this rebalancing, and we therefore excluded their data from our analyses for that quiz. We used this rebalancing procedure when constructing knowledge estimates for the

"All questions" and "Within-lecture" versions of the analyses shown in Figure 6, but not for the "Across-lecture" analyses as, in this case, the target questions and the questions used to estimate participants' knowledge for them were drawn from different subsets of quiz questions (those about one lecture, and those about the other), and were therefore independent.

In each version of this analysis (i.e., row in Fig. 6), and separately for each of the three quizzes (i.e., column in Fig. 6), we then fit a generalized linear mixed model (GLMM) with a logistic link function to the set of knowledge estimates for all questions (or all questions about a particular lecture) that participants answered on the given quiz. We implemented these models in `R` using the `lme4` package [3] and fit them following guidance from Bates et al. [2] and Matuschek et al. [45]. Specifically, we initially fit each model with the maximal random effects structure afforded by our design, which we identified as:

$$\texttt{accuracy} \sim \texttt{knowledge} + (\texttt{knowledge} \mid \texttt{participant}) + (\texttt{knowledge} \mid \texttt{question})$$

where "`accuracy`" is a binary value indicating whether each target question was answered correctly or incorrectly, "`knowledge`" is estimated knowledge at each target question's embedding coordinate, "`participant`" is a unique identifier assigned to each participant, and "`question`" is a unique identifier assigned to each quiz question. For models we fit using knowledge estimates for target questions about multiple content areas (i.e., in the "All questions" version of the analysis), we also included an additional random effect term, (`knowledge` | `lecture`), where "`lecture`" is a categorical value denoting whether the target question was about *Four Fundamental Forces*, *Birth of Stars*, or general physics knowledge. Note that with our coding scheme, identifiers for each `question` are implicitly nested within levels of `lecture` and so do not require explicit nesting in our model formula. We then iteratively removed random effects from the maximal model until it successfully converged with a full-rank random effects variance-covariance matrix. We obtained the odds ratios reported in Figure 6 by exponentiating the estimated coefficient for "`knowledge`" from each fitted model. Conceptually, these odds ratios represent how many times greater the odds are that a given participant will answer a given question correctly if their estimated knowledge for its embedding coordinate is 1, compared to if it is 0. We estimated 95% confidence intervals for each odds ratio by generating 10,000 random subsamples (of full size, with replacement) from the data used to fit each model, and refitting the models to each subsample to obtain bootstrap distributions of 10,000 odds ratios.

To assess the predictive value of our knowledge estimates, we compared each GLMM's ability to explain participants' success on individual quiz questions to that of an analogous model which assumed (as we assume under our null hypothesis) that knowledge estimates for correctly and incorrectly answered questions did *not* systematically differ, on average. Specifically, we used the same sets of observations with which we fit each "full" model to fit a second "null" model that had the same random effects structure, but in which the coefficient for the fixed effect of "`knowledge`" was fixed at 0 (i.e., we removed this term from the null model). We then compared each full model to its reduced (null) equivalent using a likelihood-ratio test (LRT). Because the standard asymptotic χ^2_d approximation of the null distribution for the LRT statistic (λ_{LR}) can be anti-conservative for finite sample sizes [25, 60, 65], we computed *p*-values for these tests using a parametric bootstrap procedure [14, 27]. For each of 10,000 bootstraps, we used the fitted null model to simulate a sample of observations of equal size to our original sample. We then re-fit both the null and full models to this simulated sample and compared them via an LRT. This yielded a distribution of λ_{LR} statistics we may expect to observe given data that conforms to our null hypothesis. We computed a corrected *p*-value for our observed λ_{LR} as $\frac{r+1}{n+1}$, where r is the number of simulated model comparisons that yielded a λ_{LR} greater than our observed value and n is the number of simulations we ran (10,000).

Estimating the "smoothness" of knowledge

In the analysis reported in Figure 7A, we show how participants' ability to correctly answer quiz questions changes as a function of distance from a given correctly or incorrectly answered reference question. We used a bootstrap-based approach to estimate the maximum distances over which these proportions of correctly answered questions could be reliably distinguished from participants' overall average proportion of correctly answered questions.

For each of 10,000 iterations, we drew a random subsample (with replacement) of 50 participants from our dataset. Within each iteration, we first computed the 95% confidence interval (CI) of the across-subsample-participants mean proportion correct on each of the three quizzes, separately. To compute this interval for each quiz, we repeatedly (1,000 times) subsampled participants (with replacement, from the outer subsample for the current iteration) and computed the mean proportion correct of each of these inner subsamples. We then identified the 2.5^{th} and 97.5^{th} percentiles of the resulting distributions of 1,000 means. These three intervals (one for each quiz) served as our thresholds for confidence that the proportion correct within a given distance from a reference question was reliably different (at the $p < 0.05$ significance level) from the average proportion correct across all questions on the given quiz.

Next, for each participant in the current subsample, and for each of the three quizzes they completed (separately), we iteratively treated each of the 15 questions appearing on the given quiz as the "reference" question. We constructed a series of concentric 15-dimensional "spheres" centered on the reference question's embedding space coordinate, where each successive sphere's radius increased by 0.01 (correlation distance) between 0 and 2, inclusive (i.e., tiling the range of possible correlation distances with 201 spheres in total). We then computed the proportion of questions enclosed within each sphere that the participant answered correctly, and averaged these per-radius proportion-correct scores across reference questions that were answered correctly, and those that were answered incorrectly. This resulted in two number-of-spheres sequences of proportion-correct scores for each subsample participant and quiz: one derived from correctly answered reference questions, and one derived from incorrectly answered reference questions.

We computed the across-subsample-participants mean proportion correct for each radius value (i.e., sphere) and "correctness" of reference question. This yielded two sequences of proportion-correct scores for each quiz, analogous to the blue and red lines displayed in Figure 7A, but for the present subsample. For each quiz, we then found the minimum distance from the reference question (i.e., sphere radius) at which each of these two sequences of per-radius proportion-correct scores intersected the 95% confidence interval for the overall proportion correct (i.e., analogous to the black error bands in Fig. 7A).

This resulted in two "intersection" distances for each quiz (for correctly answered and incorrectly answered reference questions). Repeating this full process for each of the 10,000 bootstrap iterations output two distributions of intersection distances for each of the three quizzes. The means and 95% confidence intervals for these distributions are plotted in Figure 7B.

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. 8, Supp. Figs. 7, 8, 9, 10, and 11), we used Uniform Manifold Approximation and Projection [UMAP; 46, 47] to construct a 2D projection of the text embedding space. Whereas our main analyses used a 15-topic embedding space, we used a 100-topic embedding space for these visualizations. This change in the number of topics overcame an undesirable behavior in the UMAP embedding procedure, whereby embedding coordinates for the 15-topic model tended to be "clumped" into separated clusters, rather than forming a smooth trajectory through the 2D space. When we increased the number of topics to 100, the embedding coordinates in the 2D space formed a smooth trajectory through the space, with substantially less clumping (Fig. 8). Creating a "map" by sampling this 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 for word clouds, as in Fig. 8C), 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: P.C.F., A.C.H., and J.R.M. Methodology: P.C.F., A.C.H., and J.R.M. Software: P.C.F. Validation: P.C.F. Formal analysis: P.C.F. Resources: P.C.F., A.C.H., and J.R.M. Data curation: P.C.F. Writing (original draft): J.R.M. Writing (review and editing): P.C.F., A.C.H., and J.R.M. Visualization: P.C.F. and J.R.M. Supervision: J.R.M. Project administration: P.C.F. Funding acquisition: J.R.M.

Data availability

All of the data analyzed in this manuscript may be found at https://github.com/ContextLab/efficient-learning-khan.

Code availability

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 J.R.M. 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] Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015a). Parsimonious mixed models. *arXiv*, 1506.04967.

[3] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015b). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

[4] Bevilacqua, 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.

[5] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text with the natural language toolkit*. Reilly Media, Inc.

[6] 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.

[7] 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.

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

[9] Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models: problems, diagnostics, and improvements. In Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E., editors, *Handbook of Mixed Membership Models and Their Applications*. CRC Press.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[15] 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.

[16] 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.

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

[18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*, 1810.04805.

[19] 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.

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

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

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

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

[24] 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.

[25] Goldman, N. and Whelan, S. (2000). Statistical Tests of Gamma-Distributed Rate Heterogeneity in Models of Sequence Evolution in Phylogenetics. *Molecular Biology and Evolution*, 17(6):975–978.

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

[27] Halekoh, U. and Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *Journal of Statistical Software*, 59(9):1–32.

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

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

[30] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, 9:doi.org/10.3389/fpsyg.2018.00133.

[31] 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.

[32] 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.

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

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

[35] 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.

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

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

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

[39] 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.

[40] 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.

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

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

[43] Manning, J. R. (2023). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook of Human Memory*. Oxford University Press.

[44] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension for adding LLM-driven chatbots to interactive notebooks. `https://github.com/ContextLab/chatify`.

[45] Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type i error and power in linear mixed models. *Journal of Memory and Language*, 94:305–315.

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

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

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.

[54] OpenAI (2023). ChatGPT. https://chat.openai.com.

[55] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models. *arXiv*, 2208.02957.

[56] 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.

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

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

[59] 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.

[60] Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, 52(7):3283–3299.

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

[62] 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.

[63] 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.

[64] 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.

[65] Snijders, T. A. B. and Bosker, R. (2011). More powerful tests for variance parameters. In *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, chapter 6, pages 94–108. Sage Publications, 2nd edition.

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

[67] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguz, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.

[68] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Piontkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust detection of AI-generated texts. *arXiv*, 2306.04723.

[69] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold? distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).

[70] Viswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

[71] 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.

[72] 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.

[73] 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.