

¹ Text embedding models yield high-resolution insights
² into conceptual knowledge from short multiple-choice
³ quizzes

⁴ Paxton C. Fitzpatrick¹, Andrew C. Heusser^{1, 2}, and Jeremy R. Manning^{1,*}

¹Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

²Akili Interactive Labs

Boston, MA 02110, USA

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

⁵ **Abstract**

⁶ We develop a mathematical framework, based on natural language processing models, for track-
⁷ ing and characterizing the acquisition of conceptual knowledge. Our approach embeds each
⁸ concept in a high-dimensional representation space, where nearby coordinates reflect similar or
⁹ related concepts. We test our approach using behavioral data from participants who answered
¹⁰ small sets of multiple-choice quiz questions interleaved between watching two course videos
¹¹ from the Khan Academy platform. We apply our framework to the videos' transcripts and
¹² the text of the quiz questions to quantify the content of each moment of video and each quiz
¹³ question. We use these embeddings, along with participants' quiz responses, to track how the
¹⁴ learners' knowledge changed after watching each video. Our findings show how a small set of
¹⁵ quiz questions may be used to obtain rich and meaningful high-resolution insights into what
¹⁶ each learner knows, and how their knowledge changes over time as they learn.

¹⁷ **Keywords:** education, learning, knowledge, concepts, natural language processing

¹⁸ **Introduction**

¹⁹ Suppose that a teacher had access to a complete, tangible “map” of everything a student knows.
²⁰ Defining what such a map might even look like, let alone how it might be constructed or filled in, is
²¹ itself a non-trivial problem. But if a teacher *were* to gain access to such a map, how might it change
²² their ability to teach that student? Perhaps they might start by checking how well the student
²³ knows the to-be-learned information already, or how much they know about related concepts.
²⁴ For some students, they could potentially optimize their teaching efforts to maximize efficiency
²⁵ by focusing primarily on not-yet-known content. For other students (or other content areas), it
²⁶ might be more effective to optimize for direct connections between already known content and
²⁷ new material. Observing how the student’s knowledge changed over time, in response to their
²⁸ teaching, could also help to guide the teacher towards the most effective strategy for that individual
²⁹ student.

³⁰ A common approach to assessing a student’s knowledge is to present them with a set of quiz
³¹ questions, calculate the proportion they answer correctly, and provide them with feedback in the
³² form of a simple numeric or letter grade. While such a grade can provide *some* indication of whether
³³ the student has mastered the to-be-learned material, any univariate measure of performance on a
³⁴ complex task sacrifices certain relevant information, risks conflating underlying factors, and so on.
³⁵ For example, consider the relative utility of the theoretical map described above that characterizes
³⁶ a student’s knowledge in detail, versus a single annotation saying that the student answered 85%
³⁷ of their quiz questions correctly, or that they received a ‘B’. Here, we show that the same quiz data
³⁸ required to compute proportion-correct scores or letter grades can instead be used to obtain far
³⁹ more detailed insights into what a student knew at the time they took the quiz.

⁴⁰ Designing and building procedures and tools for mapping out knowledge touches on deep
⁴¹ questions about what it means to learn. For example, how do we acquire conceptual knowledge?
⁴² Memorizing course lectures or textbook chapters by rote can lead to the superficial *appearance*
⁴³ of understanding the underlying content, but achieving true conceptual understanding seems to
⁴⁴ require something deeper and richer. Does conceptual understanding entail connecting newly

45 acquired information to the scaffolding of one’s existing knowledge or experience [4, 9, 11, 12, 25,
46 57]? Or weaving a lecture’s atomic elements (e.g., its component words) into a structured network
47 that describes how those individual elements are related [35, 61]? Conceptual understanding
48 could also involve building a mental model that transcends the meanings of those individual
49 atomic elements by reflecting the deeper meaning underlying the gestalt whole [32, 36, 54, 60].

50 The difference between “understanding” and “memorizing,” as framed by researchers in ed-
51 ucation, cognitive psychology, and cognitive neuroscience [e.g., 20, 23, 28, 36, 54], has profound
52 analogs in the fields of natural language processing and natural language understanding. For
53 example, considering the raw contents of a document (e.g., its constituent symbols, letters, and
54 words) might provide some clues as to what the document is about, just as memorizing a pas-
55 sage might provide some ability to answer simple questions about it. However, text embedding
56 models [e.g., 5, 6, 8, 10, 13, 34, 44, 62] also attempt to capture the deeper meaning *underlying* those
57 atomic elements. These models consider not only the co-occurrences of those elements within and
58 across documents, but (in many cases) also patterns in how those elements appear across different
59 scales (e.g., sentences, paragraphs, chapters, etc.), the temporal and grammatical properties of the
60 elements, and other high-level characteristics of how they are used [37, 38]. To be clear, this is not
61 to say that text embedding models themselves are capable of “understanding” deep conceptual
62 meaning in any traditional sense. But rather, their ability to capture the underlying *structure* of
63 text documents beyond their surface-level contents provides a computational framework through
64 which those document’s deeper conceptual meaning may be quantified, explored, and understood.
65 According to these models, the deep conceptual meaning of a document may be captured by a
66 feature vector in a high-dimensional representation space, wherein nearby vectors reflect concep-
67 tually related documents. A model that succeeds at capturing an analogue of “understanding” is
68 able to assign nearby feature vectors to two conceptually related documents, *even when the specific*
69 *words contained in those documents have limited overlap*. In this way, “concepts” are defined implicitly
70 by the model’s geometry [e.g., how the embedding coordinate of a given word or document relates
71 to the coordinates of other text embeddings; 49].

72 Given these insights, what form might a representation of the sum total of a person’s knowledge

73 take? First, we might require a means of systematically describing or representing (at least some
74 subset of) the nearly infinite set of possible things a person could know. Second, we might want to
75 account for potential associations between different concepts. For example, the concepts of “fish”
76 and “water” might be associated in the sense that fish live in water. Third, knowledge may have
77 a critical dependency structure, such that knowing about a particular concept might require first
78 knowing about a set of other concepts. For example, understanding the concept of a fish swimming
79 in water first requires understanding what fish and water *are*. Fourth, as we learn, our “current
80 state of knowledge” should change accordingly. Learning new concepts should both update our
81 characterizations of “what is known” and also unlock any now-satisfied dependencies of those
82 newly learned concepts so that they are “tagged” as available for future learning.

83 Here we develop a framework for modeling how conceptual knowledge is acquired during
84 learning. The central idea behind our framework is to use text embedding models to define the
85 coordinate systems of two maps: a *knowledge map* that describes the extent to which each concept is
86 currently known, and a *learning map* that describes changes in knowledge over time. Each location
87 on these maps represents a single concept, and the maps’ geometries are defined such that related
88 concepts are located nearby in space. We use this framework to analyze and interpret behavioral
89 data collected from an experiment that had participants answer sets of multiple-choice questions
90 about a series of recorded course lectures.

91 Our primary research goal is to advance our understanding of what it means to acquire deep,
92 real-world conceptual knowledge. Traditional laboratory approaches to studying learning and
93 memory (e.g., list-learning studies) often draw little distinction between memorization and under-
94 standing. Instead, these studies typically focus on whether information is effectively encoded or
95 retrieved, rather than whether the information is *understood*. Approaches to studying conceptual
96 learning, such as category learning experiments, can begin to investigate the distinction between
97 memorization and understanding, often by training participants to distinguish arbitrary or random
98 features in otherwise meaningless categorized stimuli [1, 17, 18, 21, 26, 52]. However the objective
99 of real-world training, or learning from life experiences more generally, is often to develop new
100 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 [23, 36]. 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.

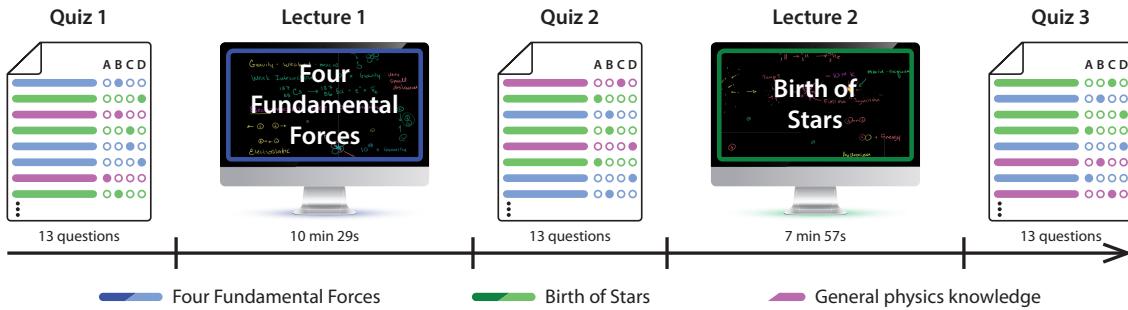


Figure 1: Experimental paradigm. Participants alternate between completing three 13-question multiple-choice quizzes and watching two Khan Academy lectures. Each quiz contains a mix of 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general physics knowledge. The specific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

127 We asked participants in our study to complete brief multiple-choice quizzes before, between,
 128 and after watching two lecture videos from the Khan Academy [31] platform (Fig. 1). The first
 129 lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics:
 130 gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*,
 131 provided an overview of our current understanding of how stars form. We selected these particular
 132 lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad
 133 audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training
 134 on participants' abilities to learn from the lectures. To this end, we selected two introductory
 135 videos that were intended to be viewed at the start of students' training in their respective content
 136 areas. Second, we wanted the two lectures to have some related content, so that we could test
 137 our approach's ability to distinguish similar conceptual content. To this end, we chose two videos
 138 from the same Khan Academy course domain, "Cosmology and Astronomy." Third, we sought to
 139 minimize dependencies and specific overlap between the videos. For example, we did not want
 140 participants' abilities to understand one video to (directly) influence their abilities to understand the
 141 other. To satisfy this last criterion, we chose videos from two different lecture series (Lectures 1 and
 142 2 were from the "Scale of the Universe" and "Stars, Black Holes, and Galaxies" series, respectively).

143 We also wrote a set of multiple-choice quiz questions that we hoped would enable us to
 144 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.

145 about physics concepts not specifically presented in either video (see Supp. Tab. 1 for the full list
 146 of questions in our stimulus pool). Participants answered questions randomly drawn from each
 147 content area (Lecture 1, Lecture 2, and general physics knowledge) on each of the three quizzes.
 148 Quiz 1 was intended to assess participants’ “baseline” knowledge before training, Quiz 2 assessed
 149 knowledge after watching the *Four Fundamental Forces* video (i.e., Lecture 1), and Quiz 3 assessed
 150 knowledge after watching the *Birth of Stars* video (i.e., Lecture 2).

151 To study in detail how participants’ conceptual knowledge changed over the course of the
 152 experiment, we first sought to model the conceptual content presented to them at each moment
 153 throughout each of the two lectures. We adapted an approach we developed in prior work [24]
 154 to identify the latent themes in the lectures using a topic model [6]. Briefly, topic models take
 155 as input a collection of text documents, and learn a set of “topics” (i.e., latent themes) from their
 156 contents. Once fit, a topic model can be used to transform arbitrary (potentially new) documents
 157 into sets of “topic proportions,” describing the weighted blend of learned topics reflected in their
 158 texts. We parsed automatically generated transcripts of the two lectures into overlapping sliding
 159 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 their non-matching question sets (Lecture 1 video vs. Lecture 2 questions: $r(13) = -0.547$, $p = 0.035$, 95% CI = [-0.812, -0.231]; Lecture 2 video vs. Lecture 1 questions: $r(13) = -0.612$, $p = 0.015$, 95% CI = [-0.874, -0.281]), indicating that the topic model also exhibits some degree of specificity. The full set of pairwise comparisons between average topic weights for the lectures and question sets is reported in



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.

188 Supplementary Figure 2.

189 Another, more sensitive, way of summarizing the conceptual content of the lectures and ques-
190 tions is to look at *variability* in how topics are weighted over time and across different questions
191 (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “infor-
192 mation” [19] the lecture (or question set) reflects about that topic. For example, suppose a given
193 topic is weighted on heavily throughout a lecture. That topic might be characteristic of some
194 aspect or property of the lecture *overall* (conceptual or otherwise), but unless the topic’s weights
195 changed in meaningful ways over time, the topic would be a poor indicator of any *specific* concep-
196 tual content in the lecture. We therefore also compared the variances in topic weights (across time
197 or questions) between the lectures and questions. The variability in topic expression (over time
198 and across questions) was similar for the Lecture 1 video and questions ($r(13) = 0.824, p < 0.001,$
199 95% CI = [0.696, 0.973]) and the Lecture 2 video and questions ($r(13) = 0.801, p < 0.001, 95%$
200 CI = [0.539, 0.958]). Simultaneously, as reported in Figure 3B, the variability in topic expression
201 across *different* videos and lecture-specific questions (i.e., Lecture 1 video vs. Lecture 2 questions;

202 Lecture 2 video vs. Lecture 1 questions) were negatively correlated, and neither video’s topic
203 variability was reliably correlated with the topic variability across general physics knowledge
204 questions. Taken together, the analyses reported in Figure 3 and Supplementary Figure 2 indicate
205 that a topic model fit to the videos’ transcripts can also reveal correspondences (at a coarse scale)
206 between the lectures and questions.

207 While an individual lecture may be organized around a single broad theme at a coarse scale,
208 at a finer scale, each moment of a lecture typically covers a narrower range of content. Given
209 the correspondence we found between the variability in topic expression across moments of each
210 lecture and questions from its corresponding set (Fig. 3), we wondered whether the text embedding
211 model might additionally capture these conceptual relationships at a finer scale. For example, if a
212 particular question asks about the content from one small part of a lecture, we wondered whether
213 the text embeddings could be used to automatically identify the “matching” moment(s) in the
214 lecture. To explore this, we computed the correlation between each question’s topic weights
215 and the topic weights for each second of its corresponding lecture, and found that each question
216 appeared to be temporally specific (Fig. 4). In particular, most questions’ topic vectors were
217 maximally correlated with a well-defined (and relatively narrow) range of timepoints from their
218 corresponding lectures, and the correlations fell off sharply outside of that range (Supp. Figs. 3, 4).
219 We also qualitatively examined the best-matching intervals for each question by comparing the
220 question’s text to the text of the most-correlated parts of the lectures (Supp. Tab. 3). Despite that
221 the questions were excluded from the text embedding model’s training set, in general we found
222 (through manual inspection) a close correspondence between the conceptual content that each
223 question probed and the content covered by the best-matching moments of the lectures. Two
224 representative examples are shown at the bottom of Figure 4.

225 The ability to quantify how much each question is “asking about” the content from each moment
226 of the lectures could enable high-resolution insights into participants’ knowledge. Traditional
227 approaches to estimating how much a student “knows” about the content of a given lecture entail
228 computing the proportion of correctly answered questions. But if two students receive identical
229 scores on an exam, might our modeling framework help us to gain more nuanced insights into the



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.

230 specific content that each student has mastered (or failed to master)? For example, a student who
 231 misses three questions that were all about the same concept (e.g., concept *A*) will have gotten the
 232 same proportion of questions correct as another student who missed three questions about three
 233 different concepts (e.g., *A*, *B*, and *C*). But if we wanted to help these two students fill in the “gaps” in
 234 their understandings, we might do well to focus specifically on concept *A* for the first student, but
 235 to also add in materials pertaining to concepts *B* and *C* for the second student. In other words, raw
 236 “proportion-correct” measures may capture *how much* a student knows, but not *what* they know.
 237 We wondered whether our modeling framework might enable us to (formally and automatically)
 238 infer participants’ knowledge at the scale of individual concepts (e.g., as captured by a single
 239 moment of a lecture).

240 We developed a simple formula (Eqn. 1) for using a participant’s responses to a small set
 241 of multiple-choice questions to estimate how much the participant “knows” about the concept
 242 reflected by any arbitrary coordinate, x , in text embedding space (e.g., the content reflected by

any moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially, the estimated knowledge at coordinate x is given by the weighted average proportion of quiz questions the participant answered correctly, where the weights reflect how much each question is “about” the content at x . When we apply this approach to estimate the participant’s knowledge about the content presented in each moment of each lecture, we can obtain a detailed time course describing how much “knowledge” the 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’ knowledge, these estimates are of course only *useful* to the extent that they accurately reflect what participants actually know. As one sanity check, we anticipated that the knowledge estimates should reflect a content-specific “boost” in participants’ knowledge after watching each lecture. In other words, if participants learn about each lecture’s content when they watch each lecture, the knowledge estimates should capture that. After watching the *Four Fundamental Forces* lecture, participants should exhibit more knowledge for the content of that lecture than they had before, and that knowledge should persist for the remainder of the experiment. Specifically, knowledge about that lecture’s content should be relatively low when estimated using Quiz 1 responses, but should increase when estimated using Quiz 2 or 3 responses (Fig. 5B). Indeed, we found that participants’ estimated knowledge about the content of *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



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.

knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a “boost” on Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ($t(49) = 1.013, p = 0.316$), but the estimated knowledge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and Quiz 3 versus 1 ($t(49) = 8.969, p < 0.001$).

If we are able to accurately estimate a participant’s knowledge about the content tested by a given question, our estimates of their knowledge should carry some predictive information about whether the participant is likely to answer that question correctly or incorrectly. We developed a statistical approach to test this claim. For each question, in turn, we used Equation 1 to predict each participant’s knowledge at the given question’s embedding space coordinate, using all *other* questions that participant answered on the same quiz. For each quiz, we grouped these predicted knowledge values into two distributions: one for the predicted knowledge at the coordinates of *correctly* answered questions, and another for the predicted knowledge at the coordinates of *incorrectly* answered questions (Fig. 6). We then used Mann-Whitney U-tests to compare the means of these distributions of predicted knowledge.

We carried out these analyses in three different ways. First, we used all (but one) of the questions from a given quiz (and participant) to predict knowledge at the embedding coordinate of a held-out question (“All questions” in Fig. 6). This test was intended to serve as an overall baseline for the predictive power of our approach. Second, we used questions about one lecture to predict knowledge at the embedding coordinate of a held-out question about the *other* lecture, from the same quiz and participant (“Across-lecture” in Fig. 6). This test was intended to test the *generalizability* of our approach by asking whether our knowledge predictions held across the content areas of the two lectures. Third, we used questions about one lecture to predict knowledge at the embedding coordinate of a held-out question about the *same* lecture, from the same quiz and participant (“Within-lecture” in Fig. 6). This test was intended to test the *specificity* of our approach by asking whether our knowledge predictions could distinguish between questions about different content covered by the same lecture. We repeated each of these analyses using all possible held-out questions for each quiz and participant.

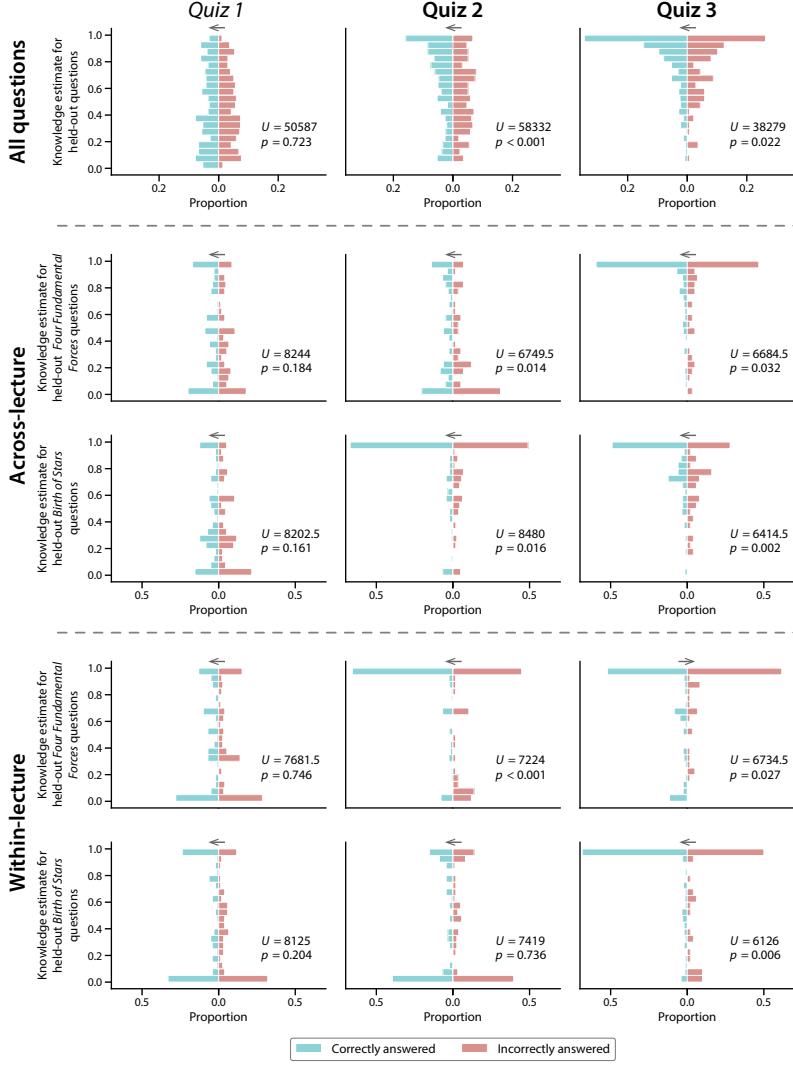


Figure 6: Predicting knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held-out correctly (blue) or incorrectly (red) answered question. The Mann-Whitney U-tests reported in each panel are between the distributions of predicted knowledge at the coordinates of correctly and incorrectly answered held-out questions. In the top row (“All questions”), we used all quiz questions (from each quiz, for each participant) except one to predict knowledge at the held-out question’s embedding coordinate. In the middle rows (“Across-lecture”), we used all questions about one lecture to predict knowledge at the embedding coordinate of a held-out question about the *other* lecture. In the bottom row (“Within-lecture”), we used all but one question about one lecture to predict knowledge at the embedding coordinate of a held-out question about the *same* lecture. We repeated each of these analyses using all possible held-out questions for each quiz and participant.

299 For the initial quizzes participants took (prior to watching either lecture), predicted knowledge
300 tended to be low overall, and relatively unstructured (Fig. 6, left column). When we held out indi-
301 vidual questions and predicted their knowledge at the held-out questions' embedding coordinates,
302 we found no reliable differences in the predictions when the held-out question had been correctly
303 versus incorrectly answered. This "null" effect persisted when we used *all* of the Quiz 1 questions
304 from a given participant to predict a held-out question ("All questions"; $U = 50587$, $p = 0.723$),
305 when we used questions from one lecture to predict knowledge at the embedding coordinate of
306 a held-out question about the *other* lecture ("Across-lecture"; predicting knowledge for held-out
307 *Four Fundamental Forces Questions* using *Birth of Stars* questions: $U = 8244$, $p = 0.184$; predicting
308 knowledge for held-out *Birth of Stars* questions: $U = 8202.5$, $p = 0.161$), and when we used ques-
309 tions from one lecture to predict knowledge at the embedding coordinate of a held-out question
310 about the *same* lecture ("Within-lecture"; *Four Fundamental Forces*: $U = 7681.5$, $p = 0.746$; *Birth of*
311 *Stars*: $U = 8125$, $p = 0.204$). We believe that this reflects a floor effect: when knowledge is low
312 everywhere, there is little signal to differentiate between what is known versus unknown.

313 After watching *Four Fundamental Forces*, predicted knowledge for held-out questions that were
314 answered correctly (from the second quiz; Fig. 6, middle column) exhibited a significant positive
315 shift relative to held-out questions that were answered incorrectly. This held when we included
316 all questions in the analysis ($U = 58332$, $p < 0.001$), when we predicted knowledge across-
317 lectures (*Four Fundamental Forces*: $U = 6749.5$, $p = 0.014$; *Birth of Stars*: $U = 8480$, $p = 0.016$),
318 and when we predicted knowledge at the embedding coordinates of held-out *Four Fundamental*
319 *Forces* questions using other *Four Fundamental Forces* questions from the same quiz and participant
320 ($U = 7224$, $p < 0.001$). This difference did *not* hold for within-lecture knowledge predictions at
321 knowledge at embedding space coordinates of *Birth of Stars* questions ($U = 7419$, $p = 0.739$). Again,
322 we suggest that this might reflect a floor effect whereby, at that point in the participants' training,
323 their knowledge about the content of the *Birth of Stars* material is relatively low everywhere in that
324 region of text embedding space.

325 Finally, after watching *Birth of Stars*, predicted knowledge for held-out correctly answered ques-
326 tions (from the third quiz; Fig. 6, right column) was higher than for held-out incorrectly answered

327 questions. This held when we included all questions in the analysis ($U = 38279$, $p = 0.022$), when
328 we carried out across-lecture predictions (*Four Fundamental Forces*: $U = 6684.5$, $p = 0.032$; *Birth*
329 *of Stars*: $U = 6414.5$, $p = 0.002$), and when we carried out within-lecture knowledge predic-
330 tions for held-out *Birth of Stars* questions using other *Birth of Stars* questions from the same quiz
331 and participant ($U = 6126$, $p = 0.006$). However, we found the *opposite* effect when we carried
332 out within-lecture knowledge predictions for held-out *Four Fundamental Forces* questions using
333 other *Four Fundamental Forces* questions from the same quiz and participant ($U = 6734$, $p = 0.027$).
334 Specifically, on Quiz 3, our knowledge predictions for held-out correctly answered questions about
335 *Four Fundamental Forces* were reliably *lower* than those for their incorrectly answered counterparts.
336 Speculatively, we suggest that this may reflect participants forgetting some of the *Four Fundamental*
337 *Forces* content. If this forgetting happens in a relatively “random” way (with respect to spatial dis-
338 tance within the text embedding space), then it could explain why some held-out questions about
339 *Four Fundamental Forces* were answered incorrectly, even if questions at nearby coordinates (i.e.,
340 about similar content) were answered correctly. This might lead our approach to over-estimate
341 knowledge for held-out questions about “forgotten” knowledge that participants answered in-
342 correctly. Taken together, the results in Figure 6 indicate that our approach can reliably predict
343 acquired knowledge (especially about recently learned content), and that the knowledge predic-
344 tions are generalizable across the content areas spanned by the two lectures, while also specific
345 enough to distinguish between questions about more subtly different content within the same
346 lecture.

347 That the knowledge predictions derived from the text embedding space reliably distinguish
348 between held-out correctly versus incorrectly answered questions (Fig. 6) suggests that spatial
349 relationships ~~in the text embedding within this~~ space can help explain what participants know.
350 But how far does this explanatory power extend? For example, suppose we know that a participant
351 ~~answers a question~~ (correctly answered a question) at embedding coordinate x ~~correctly~~. As we
352 move farther away from x in the embedding space, how does ~~their knowledge~~ the likelihood that
353 ~~the participant knows about the content at a given location~~ “fall off” with distance? ~~Or~~ Conversely,
354 suppose the participant instead answered that same question *incorrectly*. Again, as we move farther

away from x in the embedding space, how ~~do the chances does the likelihood~~ that the participant does *not* know about ~~the a coordinate's~~ content change with distance? We reasoned that, assuming our embedding space is capturing something about how ~~participants~~individuals actually organize their knowledge, ~~conceptual knowledge right around a participant's ability to answer questions~~ embedded very close to x should tend to be similar to ~~the participant's knowledge of the content at their ability to answer the question embedded at x~~ . ~~And Whereas~~ at another extreme, ~~at some distance (after moving sufficiently far away once we reach some sufficiently large distance from x), our guesses about what participants know (our ability to infer whether or not a participant will correctly answer a question) based on their response to ability to answer the question at location x)~~ should be no better than guessing based on their ~~overall overall~~ proportion of correctly answered questions—~~i.e., if y is very far away from x , all we can do with the~~. In other words, beyond the maximum distance at which the participant's response to ability to answer the question at x is ~~guess that "their performance on quiz questions about~~ informative of their ability to answer a second question at location y ~~will be about equal to their average performance on quiz questions about any material.~~”, 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 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.

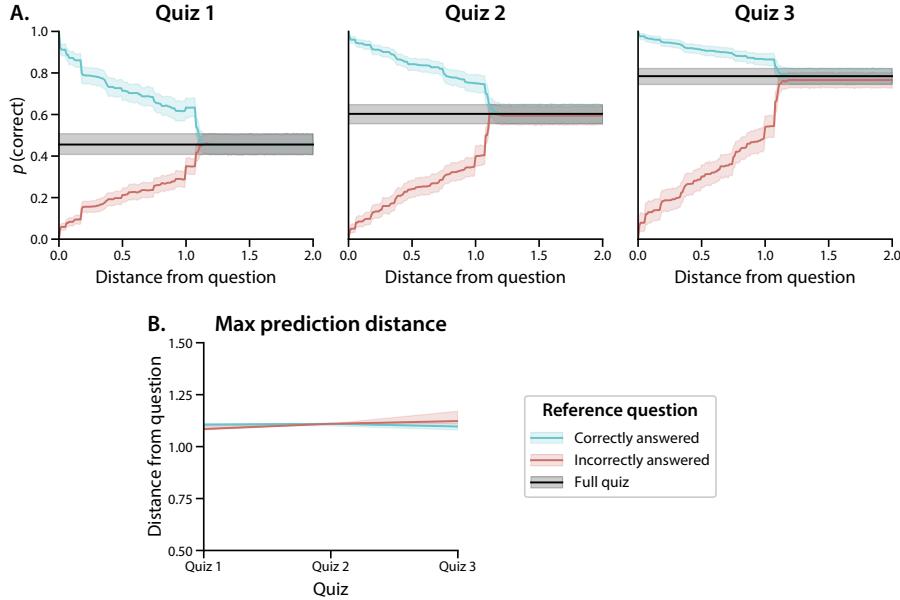


Figure 7: Quiz performance falls off gradually in text embedding space. A. Performance versus distance. For each participant, for each correctly answered question (blue) or incorrectly answered question (red), we computed the proportion of correctly answered questions within a given distance of that question’s embedding coordinate. We 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.

383 Knowledge estimates need not be limited to the content of the lectures. As illustrated in
 384 Figure 8, our general approach to estimating knowledge from a small number of quiz questions
 385 may be extended to *any* content, given its text embedding coordinate. To visualize how knowledge
 386 “spreads” through text embedding space to content beyond the lectures participants watched, we
 387 first fit a new topic model to the lectures’ sliding windows with (up to) $k = 100$ topics. Conceptually,
 388 increasing the number of topics used by the model functions to increase the “resolution” of the
 389 embedding space, providing a greater ability to estimate knowledge for content that is highly
 390 similar to (but not precisely the same as) that contained in the two lectures. ~~This change in the
 391 number of topics overcame an undesirable behavior in the UMAP embedding procedure [40],~~

392 whereby embedding coordinates for the 15 topic model tended to be “clumped” into separated
393 clusters, rather than forming a smooth trajectory through the 2D space. When we increased the
394 number of topics to 100, the embedding coordinates in the 2D space formed a smooth trajectory
395 through the space, with substantially less clumping (Fig. 8). We note that we used these 2D maps
396 solely for visualization; all relevant comparisons, distance computations, and statistical tests we
397 report above were carried out in the original 15-dimensional space, using the 15-topic model. Aside
398 from increasing the number of topics from 15 to 100, all other procedures and model parameters
399 were carried over from the preceding analyses. As in our other analyses, we resampled each
400 lecture’s topic trajectory to 1 Hz and projected each question into a shared text embedding space.

401 We projected the resulting 100-dimensional topic vectors (for each second of video and each quiz
402 question) onto a shared 2-dimensional plane (see *Creating knowledge and learning map visualizations*).
403 Next, we sampled points from a 100×100 grid of coordinates that evenly tiled a rectangle enclosing
404 the 2D projections of the videos and questions. We used Equation 4 to estimate participants’
405 knowledge at each of these 10,000 sampled locations, and averaged these estimates across par-
406 ticipants to obtain an estimated average *knowledge map* (Fig. 8A). Intuitively, the knowledge map
407 constructed from a given quiz’s responses provides a visualization of how “much” participants
408 knew about any content expressible by the fitted text embedding model at the point in time when
409 they completed that quiz.

410 Several features of the resulting knowledge maps are worth noting. The average knowledge
411 map estimated from Quiz 1 responses (Fig. 8A, leftmost map) shows that participants tended to
412 have relatively little knowledge about any parts of the text embedding space (i.e., the shading is
413 relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked
414 increase in knowledge on the left side of the map (around roughly the same range of coordinates
415 traversed by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words,
416 participants’ estimated increase in knowledge is localized to conceptual content that is nearby (i.e.,
417 related to) the content from the lecture they watched prior to taking Quiz 2. This localization is
418 non-trivial: these knowledge estimates are informed only by the embedded coordinates of the
419 *quiz questions*, not by the embeddings of either lecture (see Eqn. 4). Finally, the knowledge map

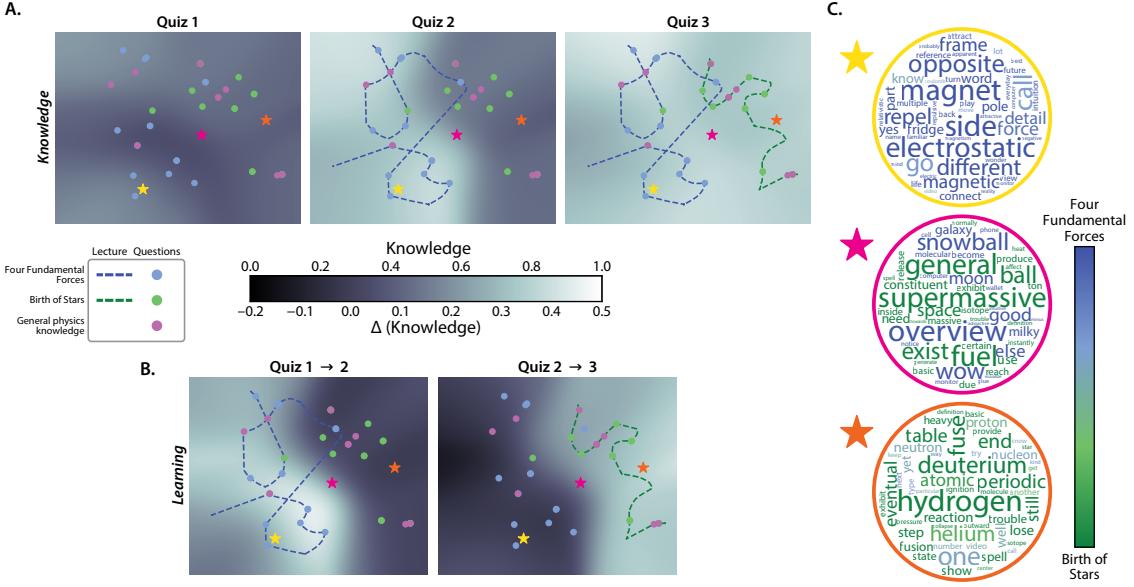


Figure 8: Mapping out the geometry of knowledge and learning. **A.** Average “knowledge maps” estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*). 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.

420 estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region
421 surrounding the embedding of the *Birth of Stars* lecture participants watched immediately prior to
422 taking Quiz 3.

423 Another way of visualizing these content-specific increases in knowledge after participants
424 viewed each lecture is displayed in Figure 8B. Taking the point-by-point difference between the
425 knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map*
426 that describes the *change* in knowledge estimates from one quiz to the next. These learning maps
427 highlight that the estimated knowledge increases we observed across maps were specific to the
428 regions around the embeddings of each lecture, in turn.

429 Because the 2D projection we used to construct the knowledge and learning maps is invertible,
430 we may gain additional insights into these maps' meanings by reconstructing the original high-
431 dimensional topic vector for any location on the map we are interested in. For example, this could
432 serve as a useful tool for an instructor looking to better understand which content areas a student
433 (or a group of students) knows well (or poorly). As a demonstration, we show the top-weighted
434 words from the blends of topics reconstructed from three example locations on the maps (Fig. 8C):
435 one point near the *Four Fundamental Forces* embedding (yellow), a second point near the *Birth of*
436 *Stars* embedding (orange), and a third point between the two lectures' embeddings (pink). As
437 shown in the word clouds in the panel, the top-weighted words at the example coordinate near the
438 *Four Fundamental Forces* embedding tended to be weighted more heavily by the topics expressed
439 in that lecture. Similarly, the top-weighted words at the example coordinate near the *Birth of Stars*
440 embedding tended to be weighted more heavily by the topics expressed in *that* lecture. And the
441 top-weighted words at the example coordinate between the two lectures' embeddings show a
442 roughly even mix of words most strongly associated with each lecture.

443 Discussion

444 We developed a computational framework that uses short multiple-choice quizzes to gain nuanced
445 insights into what learners know and how their knowledge changes with training. First, we show

446 that our approach can automatically match the conceptual knowledge probed by individual quiz
447 questions to the corresponding moments in lecture videos when those concepts were presented
448 (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment “knowledge traces”
449 that reflect the degree of knowledge participants have about each video’s time-varying content,
450 and capture temporally specific increases in knowledge after viewing each lecture (Fig. 5). We
451 also show that these knowledge estimates can generalize to held-out questions (Fig. 6). Finally,
452 we use our framework to construct visual maps that provide snapshot estimates of how much
453 participants know about any concept within the scope of our text embedding model, and how
454 much their knowledge of those concepts changes with training (Fig. 8).

455 We view our work as making several contributions to the study of how people acquire con-
456 ceptual knowledge. First, from a methodological standpoint, our modeling framework provides
457 a systematic means of mapping out and characterizing knowledge in maps that have infinite (ar-
458 bitrarily many) numbers of coordinates, and of “filling out” those maps using relatively small
459 numbers of multiple choice quiz questions. Our experimental finding that we can use these maps
460 to predict responses to held-out questions has several psychological implications as well. For ex-
461 ample, concepts that are assigned to nearby coordinates by the text embedding model also appear
462 to be “known to a similar extent” (as reflected by participants’ responses to held-out questions;
463 Fig. 6). This suggests that participants also *conceptualize* similarly the content reflected by nearby
464 embedding coordinates. The “spatial smoothness” of participants’ knowledge (as estimated using
465 quiz performance) is being captured by the knowledge maps we are inferring from their quiz
466 responses (e.g., Figs. 7, 8). In other words, our study shows that knowledge about a given concept
467 implies knowledge about related concepts, and we also show how estimated knowledge falls off
468 with distance in text embedding space.

469 In our study, we characterize the “coordinates” of participants’ knowledge using a relatively
470 simple “bag of words” text embedding model [LDA; 6]. More sophisticated text embedding mod-
471 els, such as transformer-based models [15, 48, 59, 62] can learn complex grammatical and semantic
472 relationships between words, higher-order syntactic structures, stylistic features, and more. We
473 considered using transformer-based models in our study, but we found that the text embeddings

474 derived from these models were surprisingly uninformative with respect to differentiating or oth-
475 erwise characterizing the conceptual content of the lectures and questions we used. We suspect
476 that this reflects a broader challenge in constructing models that are high-resolution within a given
477 domain (e.g., the domain of physics lectures and questions) *and* sufficiently broad so as to enable
478 them to cover a wide range of domains. For example, we found that the embeddings derived even
479 from much larger and more modern models like BERT [15], GPT [62], LLaMa [59], and others that
480 are trained on enormous text corpora, end up yielding poor resolution within the content space
481 spanned by individual course videos (Supp. Fig. 6). Whereas the LDA embeddings of the lectures
482 and questions are “near” each other (i.e., the convex hull enclosing the two lectures’ trajectories is
483 highly overlapping with the convex hull enclosing the questions’ embeddings), the BERT embed-
484 dings of the lectures and questions are instead largely distinct (top row of Supp. Fig. 6). The LDA
485 embeddings of the questions for each lecture and the corresponding lecture’s trajectory are also
486 similar. For example, as shown in Fig. 2C, the LDA embeddings for *Four Fundamental Forces* ques-
487 tions (blue dots) appear closer to the *Four Fundamental Forces* lecture trajectory (blue line), whereas
488 the LDA embeddings for *Birth of Stars* questions (green dots) appear closer to the *Birth of Stars*
489 lecture trajectory (green line). The BERT embeddings of the lectures and questions do not show
490 this property (Supp. Fig. 6). We also examined per-question “content matches” between individual
491 questions and individual moments of each lecture (Figs. 4, 6). The time series plot of individual
492 questions’ correlations are different from each other when computed using LDA (e.g., the traces
493 can be clearly visually separated), whereas the correlations computed from BERT embeddings of
494 different questions all look very similar. This tells us that LDA is capturing some differences in
495 content between the questions, whereas BERT is not. The time series plots of individual ques-
496 tions’ correlations have clear “peaks” when computed using LDA, but not when computed using
497 BERT. This tells us that LDA is capturing a “match” between the content of each question and a
498 relatively well-defined time window of the corresponding lectures. The BERT embeddings appear
499 to blur together the content of the questions versus specific moments of each lecture. Finally, we
500 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

502 “general physics knowledge” questions). The LDA embeddings show a strong contrast between
503 same-content embeddings versus across-content embeddings. In other words, the embeddings of
504 questions about the *Four Fundamental Forces* material are highly correlated with the embeddings of
505 the *Four Fundamental Forces* lecture, but not with the embeddings of *Birth of Stars*, questions about
506 *Birth of Stars*, or general physics knowledge questions. We see a similar pattern with the LDA
507 embeddings of the *Birth of Stars* questions (Fig. 3, Supp. Fig. 2). In contrast, the BERT embeddings
508 are all highly correlated with each other (Supp. Fig. 6). Taken together, these comparisons illus-
509 trate how LDA (trained on the specific content in question) provides both coverage of the requisite
510 material and specificity at the level of the content covered by individual questions. BERT, on the
511 other hand, essentially assigns both lectures and all of the questions (which are all broadly about
512 “physics”) into a tiny region of its embedding space, thereby blurring out meaningful distinctions
513 between different specific concepts covered by the lectures and questions. We note that these are
514 not criticisms of BERT (or other large language models trained on large and diverse corpora).
515 Rather, our point is that simple fine-tuned models trained on a relatively small but specialized
516 corpus can outperform much more complicated models trained on much larger corpora, when we
517 are specifically interested in capturing subtle conceptual differences at the level of a single course
518 lecture or question. Of course if our goal had been to find a model that generalized to many
519 different content areas, we would expect our approach to perform comparatively poorly relative to
520 BERT or other much larger models. We suggest that bridging the tradeoff between high resolution
521 within each content area versus the ability to generalize to many different content areas will be an
522 important challenge for future work in this domain.

523 Another application for large language models that does *not* require explicitly modeling the
524 content of individual lectures or questions is to leverage the models’ ability to generate text. For
525 example, generative text models like ChatGPT [48] and LLaMa [59] are already being used to build
526 a new generation of interactive tutoring systems [e.g., 39]. Unlike the approach we have taken here,
527 these generative text model-based systems do not explicitly model what learners know, or how
528 their knowledge changes over time with training. One could imagine building a hybrid system
529 that combines the best of both worlds: a large language model that can *generate* text, combined

530 with a smaller model that can *infer* what learners know and how their knowledge changes over
531 time. Such a hybrid system could potentially be used to build the next generation of interactive
532 tutoring systems that are able to adapt to learners' needs in real time, and that are able to provide
533 more nuanced feedback about what learners know and what they do not know.

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

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

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

564 Within the past several years, the global pandemic forced many educators to suddenly adapt to
565 teaching remotely [30, 45, 56, 63]. This change in world circumstances is happening alongside (and
566 perhaps accelerating) geometric growth in the availability of high-quality online courses from plat-
567 forms such as Khan Academy [31], Coursera [64], EdX [33], and others [53]. Continued expansion
568 of the global internet backbone and improvements in computing hardware have also facilitated
569 improvements in video streaming, enabling videos to be easily shared and viewed by increasingly
570 large segments of the world’s population. This exciting time for online course instruction provides
571 an opportunity to re-evaluate how we, as a global community, educate ourselves and each other.
572 For example, we can ask: what defines an effective course or training program? Which aspects of
573 teaching might be optimized and/or augmented by automated tools? How and why do learning
574 needs and goals vary across people? How might we lower barriers of access to a high-quality
575 education?

576 Alongside these questions, there is a growing desire to extend existing theories beyond the
577 domain of lab testing rooms and into real classrooms [29]. In part, this has led to a recent
578 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better
579 reflect more ethologically valid phenomena that are more directly relevant to real-world situations
580 and behaviors [46]. In turn, this has brought new challenges in data analysis and interpretation. A
581 key step towards solving these challenges will be to build explicit models of real-world scenarios
582 and how people behave in them (e.g., models of how people learn conceptual content from real-
583 world courses, as in our current study). A second key step will be to understand which sorts
584 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 2,
585 16, 43, 47, 50] might help to inform these models. A third major step will be to develop and

586 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic
587 paradigms.

588 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also
589 relate to the notion of “theory of mind” of other individuals [22, 27, 42]. Considering others’ unique
590 perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and
591 communicate [51, 55, 58]. One could imagine future extensions of our work (e.g., analogous to
592 the knowledge and learning maps shown in Fig. 8), that attempt to characterize how well-aligned
593 different people’s knowledge bases or backgrounds are. In turn, this might be used to model how
594 knowledge (or other forms of communicable information) flows not just between teachers and
595 students, but between friends having a conversation, individuals on a first date, participants at
596 a business meeting, doctors and patients, experts and non-experts, political allies or adversaries,
597 and more. For example, the extent to which two people’s knowledge maps “match” or “align” in
598 a given region of text embedding space might serve as a predictor of how effectively they will be
599 able to communicate about the corresponding conceptual content.

600 Ultimately, our work suggests a rich new line of questions about the geometric “form” of
601 knowledge, how knowledge changes over time, and how we might map out the full space of
602 what an individual knows. Our finding that detailed estimates about knowledge may be obtained
603 from short quizzes shows one way that traditional approaches to evaluation in education may be
604 extended. We hope that these advances might help pave the way for new approaches to teaching
605 or delivering educational content that are tailored to individual students’ learning needs and goals.

606 **Materials and methods**

607 **Participants**

608 We enrolled a total of 50 Dartmouth undergraduate students in our study. Participants received
609 optional course credit for enrolling. We asked each participant to complete a demographic survey
610 that included questions about their age, gender, native spoken language, ethnicity, race, hearing,

611 color vision, sleep, coffee consumption, level of alertness, and several aspects of their educational
612 background and prior coursework.

613 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
614 years). A total of 15 participants reported their gender as male and 35 participants reported their
615 gender as female. A total of 49 participants reported their native language as "English" and 1
616 reported having another native language. A total of 47 participants reported their ethnicity as
617 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants
618 reported their races as White (32 participants), Asian (14 participants), Black or African American
619 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
620 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

633 Participants reported their undergraduate major(s) as "social sciences" (28 participants), "nat-
634 ural sciences" (16 participants), "professional" (e.g., pre-med or pre-law; 8 participants), "mathe-
635 matics and engineering" (7 participants), "humanities" (4 participants), or "undecided" (3 partici-
636 pants). Note that some participants selected multiple categories for their undergraduate major(s).
637 We also asked participants about the courses they had taken. In total, 45 participants reported hav-
638 ing taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan

639 Academy courses. Of those who reported having watched at least one Khan Academy course,
640 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8
641 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We
642 also asked participants about the specific courses they had watched, categorized under different
643 subject areas. In the “Mathematics” area, participants reported having watched videos on AP
644 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-
645 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry
646 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential
647 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),
648 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other
649 videos not listed in our survey (5 participants). In the “Science and engineering” area, participants
650 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-
651 ipants); Physics, AP Physics I, or AP Physics II (18 participants); Biology, AP Biology; or High
652 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed
653 in our survey (5 participants). We also asked participants whether they had specifically seen the
654 videos used in our experiment. Of the 45 participants who reported having having taken at least
655 one Khan Academy course in the past, 44 participants reported that they had not watched the *Four*
656 *Fundamental Forces* video, and 1 participant reported that they were not sure whether they had
657 watched it. All participants reported that they had not watched the *Birth of Stars* video. When
658 we asked participants about non-Khan Academy online courses, they reported having watched
659 or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test
660 preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 partic-
661 ipants), Computing (2 participants), and other categories not listed in our survey (17 participants).
662 Finally, we asked participants about in-person courses they had taken in different subject areas.
663 They reported taking courses in Mathematics (38 participants), Science and engineering (37 par-
664 ticipants), Arts and humanities (34 participants), Test preparation (27 participants), Economics
665 and finance (26 participants), Computing (14 participants), College and careers (7 participants), or
666 other courses not listed in our survey (6 participants).

667 **Experiment**

668 We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces*
669 (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force;
670 duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed;
671 duration: 7 minutes and 57 seconds). All participants viewed the videos in the same order (i.e., *Four
672 Fundamental Forces* followed by *Birth of Stars*). ~~While we are not aware of any specific confounds
673 of viewing order, nor have we are aware of how or why viewing order might influence our main
674 findings, we acknowledge that we did not control for potential order effects in our study.~~

675 We then hand-created 39 multiple-choice questions: 15 about the conceptual content of *Four
676 Fundamental Forces* (i.e., Lecture 1), 15 about the conceptual content of *Birth of Stars* (i.e., Lecture 2),
677 and 9 questions that tested for general conceptual knowledge about basic physics (covering material
678 that was not presented in either video). ~~One of our group's undergraduate research assistants
679 worked alongside a rotating Masters student to develop this set of questions (these researchers
680 are acknowledged in our paper for their contribution, although they did not meet the criteria for
681 authorship discussed with all team members at the start of the project, as determined by J.R.M.) The
682 senior author (J.R.M.) tasked the pair of researchers with coming up with "15 conceptual questions
683 about each lecture, along with 9 additional questions about general physics knowledge."~~ To help
684 broaden the set of lecture-specific questions, ~~the researchers were further instructed to work our
685 team worked~~ through each lecture in small segments ~~to~~ identify what each segment was "about"
686 conceptually, and then write a question about that concept. The general physics questions were
687 drawn ~~from the researchers' coursework our team's prior coursework and areas of interest~~, along
688 with internet searches and brainstorming with the project team and other members of J.R.M.'s lab.
689 The final set of questions (and response options) was reviewed and approved by J.R.M. before we
690 collected or analyzed the text or experimental data.

691 We note that estimating the specific "amount" of conceptual understanding that each question
692 "requires" to answer is somewhat subjective, and might even come down to the "strategy" a given
693 participant uses to answer the question at that particular moment. The full set of questions and

694 answer choices may be found in Supplementary Table 1.

695 Over the course of the experiment, participants completed three 13-question multiple-choice
696 quizzes: the first before viewing Lecture 1, the second between Lectures 1 and 2, and the third
697 after viewing Lecture 2 (see Fig. 1). The questions appearing on each quiz, for each participant,
698 were randomly chosen from the full set of 39, with the constraints that (a) each quiz contained
699 exactly 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general
700 physics knowledge, and (b) each question appear exactly once for each participant. The orders of
701 questions on each quiz, and the orders of answer options for each question, were also randomized.
702 We obtained informed consent from all participants, and our experimental protocol was approved
703 by the Committee for the Protection of Human Subjects at Dartmouth College. We used this
704 experiment to develop and test our computational framework for estimating knowledge and
705 learning.

706 **Analysis**

707 **Statistics**

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

711 **Constructing text embeddings of multiple lectures and questions**

712 We adapted an approach we developed in prior work [24] to embed each moment of the two
713 lectures and each question in our pool in a common representational space. Briefly, our approach
714 uses a topic model [Latent Dirichlet Allocation; 6] trained on a set of documents, to discover a
715 set of (up to) k “topics” or “themes.” Formally, each topic is defined as a distribution of weights
716 over words in the model’s vocabulary (i.e., the union of all unique words, across all documents,
717 excluding “stop words.”). Conceptually, each topic is intended to give larger weights to words
718 that are semantically related (as inferred from their tendency to co-occur in the same document).

719 After fitting a topic model, each document in the training set, or any *new* document that contains at
720 least some of the words in the model’s vocabulary, may be represented as a k -dimensional vector
721 describing how much the document (most probably) reflects each topic. To select an appropriate
722 k for our model, as a starting point, we identified the minimum number of topics that yielded
723 at least one “unused” topic (i.e., in which all words in the vocabulary were assigned uniform
724 weights) after training. This indicated that the number of topics was sufficient to capture the set
725 of latent themes present in the two lectures (from which we constructed our document corpus, as
726 described below). We found this value to be $k = 15$ topics. We found that with a limited number
727 of additional adjustments following [7], such as removing corpus-specific stop-words, the model
728 yielded (subjectively) sensible and coherent topics. The distribution of weights over words in
729 the vocabulary for each discovered topic is shown in Supplementary Figure 1, and each topic’s
730 top-weighted words may be found in Supplementary Table 2.

731 As illustrated in Figure 2A, we start by building up a corpus of documents using overlapping
732 sliding windows that span each video’s transcript. Khan Academy provides professionally created,
733 manual transcriptions of all videos for closed captioning. However, such transcripts would not
734 be readily available in all contexts to which our framework could potentially be applied. Khan
735 Academy videos are hosted on the YouTube platform, which additionally provides automated
736 captions. We opted to use these automated transcripts [which, in prior work, we have found to be
737 of sufficiently near-human quality to yield reliable data in behavioral studies; 65] when developing
738 our framework in order to make it more directly extensible and adaptable by others in the future.

739 We fetched these automated transcripts using the `youtube-transcript-api` Python pack-
740 age [14]. The transcripts consisted of one timestamped line of text for every few seconds (mean:
741 2.34 s; standard deviation: 0.83 s) of spoken content in the video (i.e., corresponding to each indi-
742 vidual caption that would appear on-screen if viewing the lecture via YouTube, and when those
743 lines would appear). We defined a sliding window length of (up to) $w = 30$ transcript lines, and
744 assigned each window a timestamp corresponding to the midpoint between the timestamps for its
745 first and last lines. This w parameter was chosen to match the same number of words per sliding
746 window (rounded to the nearest whole word, and before preprocessing) as the sliding windows

747 we defined in our prior work [24] (i.e., 185 words per sliding window).

748 These sliding windows ramped up and down in length at the beginning and end of each
749 transcript, respectively. In other words, each transcript’s first sliding window covered only its first
750 line, the second sliding window covered the first two lines, and so on. This ensured that each line
751 from the transcripts appeared in the same number (w) of sliding windows. We next performed a
752 series of standard text preprocessing steps: normalizing case, lemmatizing, removing punctuation
753 and removing stop-words. We constructed our corpus of stop words by augmenting the Natural
754 Language Toolkit [NLTK; 3] English stop word list with the following additional words, selected
755 using the approach suggested by [7]: “actual,” “actually,” “also,” “bit,” “could,” “e,” “even,”
756 “first,” “follow,” “following,” “four,” “let,” “like,” “mc,” “really”, “saw,” “see,” “seen,” “thing,”
757 and “two.” This yielded sliding windows with an average of 73.8 remaining words, and lasting for
758 an average of 62.22 seconds. We treated the text from each sliding window as a single “document,”
759 and combined these documents across the two videos’ windows to create a single training corpus
760 for the topic model.

761 After fitting a topic model to the two videos’ transcripts, we could use the trained model to
762 transform arbitrary (potentially new) documents into k -dimensional topic vectors. A convenient
763 property of these topic vectors is that documents that reflect similar blends of topics (i.e., documents
764 that reflect similar themes, according to the model) will yield similar coordinates (in terms of
765 correlation, cosine similarity, Kullback-Leibler divergence, Euclidean distance, or other geometric
766 measures). In general, the similarity between different documents’ topic vectors may be used to
767 characterize the similarity in conceptual content between the documents.

768 We transformed each sliding window’s text into a topic vector, and then used linear interpolation
769 (independently for each topic dimension) to resample the resulting time series to one vector
770 per second. We also used the fitted model to obtain topic vectors for each question in our pool (see
771 Supp. Tab. 1). Taken together, we obtained a *trajectory* for each video, describing its path through
772 topic space, and a single coordinate for each question (Fig. 2C). Embedding both videos and all of
773 the questions using a common model enables us to compare the content from different moments
774 of videos, compare the content across videos, and estimate potential associations between specific

775 questions and specific moments of video.

776 **Estimating dynamic knowledge traces**

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

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

779 where

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

780 and where mincorr and maxcorr are the minimum and maximum correlations between any lecture
781 timepoint and question, taken over all timepoints in the given lecture, and all five questions *about*
782 that lecture appearing on the given quiz. We also define $f(s, \Omega)$ as the s^{th} topic vector from the set
783 of topic vectors Ω . Here t indexes the set of lecture topic vectors, L , and i and j index the topic
784 vectors of questions used to estimate the knowledge trace, Q . Note that “correct” denotes the set
785 of indices of the questions the participant answered correctly on the given quiz.

786 Intuitively, $\text{ncorr}(x, y)$ is the correlation between two topic vectors (e.g., the topic vector from one
787 timepoint in a lecture, x , and the topic vector for one question, y), normalized by the minimum and
788 maximum correlations (across all timepoints t and questions Q) to range between 0 and 1, inclusive.
789 Equation 1 then computes the weighted average proportion of correctly answered questions about
790 the content presented at timepoint t , where the weights are given by the normalized correlations
791 between timepoint t ’s topic vector and the topic vectors for each question. The normalization step
792 (i.e., using ncorr instead of the raw correlations) ensures that every question contributes some
793 non-negative amount to the knowledge estimate.

794 **Estimating the “smoothness” of knowledge**

795 In the analysis reported in Figure 7A, we show how participants’ quiz performance changes as
796 a function of distance to a given correctly or incorrectly answered reference question. We used

797 a bootstrap-based approach to estimate the maximum distances over which these proportions of
798 correctly answered questions could be reliably distinguished from participants' overall average
799 proportion of correctly answered questions.

800 In our bootstrap procedure, we ran 10,000 iterations to estimate the relationship between partic-
801 ipants' performance and the distance to a given reference question. For each of these iterations, for
802 every individual quiz (q), we first determined the across-participants average "simple" proportion
803 correct and its 95% confidence interval. This interval was established by repeatedly (1,000 times)
804 subsampling participants with replacement, computing the mean "simple" proportion correct for
805 each subsample, and then deriving the 2.5th and 97.5th percentiles from the distribution of these
806 subsample means. We used this interval as our benchmark for determining whether the propor-
807 tion of correctly answered questions for a given subset of questions was reliably different (at the
808 $p < 0.05$ significance level) from the average proportion correct across all questions.

809 Next, for each participant, we examined all 15 questions they answered on quiz q . We treated
810 each question as the "reference question" in turn. Around this reference, we constructed a series of
811 15-dimensional spheres (starting with a radius of 0), where each successive sphere had a radius of
812 0.01 (correlation distance) greater than its predecessor. Within each of these spheres, we calculated
813 the proportion of questions answered correctly by the participant. This yielded two distinct sets
814 of proportion-correct values for each binned distance (radius) for a specific participant and quiz:
815 one set of values where the reference questions had been answered correctly, and another set
816 where the reference questions had been answered incorrectly. From these, we established the
817 average proportion correct within each radius for both categories of reference questions. Finally,
818 we identified the minimum binned distance from the correctly answered reference questions for
819 which the average proportion correct intersected the 95% confidence interval of the simple average
820 proportion correct computed earlier. We display the resulting distance estimates, for each quiz
821 and reference question status, in Figure 7B.

822 **Creating knowledge and learning map visualizations**

823 An important feature of our approach is that, given a trained text embedding model and partic-
824 ipants’ quiz performance on each question, we can estimate their knowledge about *any* content
825 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-
826 tions, or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 7, 8, 9, 10,
827 and 11), we used Uniform Manifold Approximation and Projection [UMAP; 40, 41] to construct
828 a 2D projection of the text embedding space. ~~Sampling the original Whereas our main analyses~~
829 ~~used a 15-topic embedding space, we used a 100-topic embedding space for these visualizations.~~
830 ~~This change in the number of topics overcame an undesirable behavior in the UMAP embedding~~
831 ~~procedure [40], whereby embedding coordinates for the 15-topic model tended to be “clumped”~~
832 ~~into separated clusters, rather than forming a smooth trajectory through the 2D space. When~~
833 ~~we increased the number of topics to 100, the embedding coordinates in the 2D space formed a~~
834 ~~smooth trajectory through the space, with substantially less clumping (Fig. 8). Creating a “map” by~~
835 ~~sampling this~~ 100-dimensional space at high resolution to obtain an adequate set of topic vectors
836 spanning the embedding space would be computationally intractable. However, sampling a 2D
837 grid is trivial.

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

845 In our application, all of the coordinates we embedded were topic vectors, whose elements
846 are always non-negative and sum to one. Although UMAP is an invertible transformation at
847 the embedding locations of the original data, other locations in the embedding space will not
848 necessarily follow the same implicit “rules” as the original high-dimensional data. For example,

849 inverting an arbitrary coordinate in the embedding space might result in negative-valued vectors,
850 which are incompatible with the topic modeling framework. To protect against this issue, we
851 log-transformed the topic vectors prior to embedding them in the 2D space. When we inverted
852 the embedded vectors (e.g., to estimate topic vectors or word clouds, as in Fig. 8C), we passed
853 the inverted (log-transformed) values through the exponential function to obtain a vector of non-
854 negative values, and normalized them to sum to one.

855 After embedding both lectures' topic trajectories and the topic vectors of every question, we
856 defined a rectangle enclosing the 2D projections of the lectures' and quizzes' embeddings. We then
857 sampled points from a regular 100×100 grid of coordinates that evenly tiled this enclosing rectangle.
858 We sought to estimate participants' knowledge (and learning, i.e., changes in knowledge) at each
859 of the resulting 10,000 coordinates.

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

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

864 The λ term in the RBF equation controls the "smoothness" of the function, where larger values
865 of λ result in smoother maps. In our implementation we used $\lambda = 50$. Next, we estimated the
866 "knowledge" at each coordinate, x , using:

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

867 Intuitively, Equation 4 computes the weighted proportion of correctly answered questions, where
868 the weights are given by how nearby (in the 2D space) each question is to the x . We also defined
869 *learning maps* as the coordinate-by-coordinate differences between any pair of knowledge maps.
870 Intuitively, learning maps reflect the *change* in knowledge across two maps.

871 **Author contributions**

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

876 **Data availability**

877 All of the data analyzed in this manuscript may be found at <https://github.com/ContextLab/efficient-learning-khan>.
878

879 **Code availability**

880 All of the code for running our experiment and carrying out the analyses may be found at
881 <https://github.com/ContextLab/efficient-learning-khan>.

882 **Acknowledgements**

883 We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of
884 this study, and assistance with data collection efforts from Will Baxley, Max Bluestone, Daniel
885 Carstensen, Kunal Jha, Caroline Lee, Lucy Owen, Xinming Xu, and Kirsten Ziman. Our work was
886 supported in part by NSF CAREER Award Number 2145172 to J.R.M. The content is solely the
887 responsibility of the authors and does not necessarily represent the official views of our supporting
888 organizations. The funders had no role in study design, data collection and analysis, decision to
889 publish, or preparation of the manuscript.

890 **References**

- 891 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,
892 56:149–178.
- 893 [2] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and
894 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
895 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 896 [3] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text
897 with the natural language toolkit*. Reilly Media, Inc.
- 898 [4] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
899 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
900 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 901 [5] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International
902 Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
903 Machinery.
- 904 [6] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine
905 Learning Research*, 3:993–1022.
- 906 [7] Boyd-Graber, J. and Mimno, D. (2014). Care and feeding of topic models: problems, diagnostics,
907 and improvements. In Airolidi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E., editors,
908 *Handbook of Mixed Membership Models and Their Applications*. CRC Press.
- 909 [8] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
910 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
911 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
912 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
913 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.

- 914 [9] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
915 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 916 [10] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
917 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
918 sentence encoder. *arXiv*, 1803.11175.
- 919 [11] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
920 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 921 [12] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
922 Evidence for a new conceptualization of semantic representation in the left and right cerebral
923 hemispheres. *Cortex*, 40(3):467–478.
- 924 [13] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
925 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,
926 41(6):391–407.
- 927 [14] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 928
- 929 [15] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep
930 bidirectional transformers for language understanding. *arXiv*, 1810.04805.
- 931 [16] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
932 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
933 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 934 [17] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 935 [18] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of
936 Experimental Psychology: General*, 115:155–174.
- 937 [19] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical
938 Transactions of the Royal Society A*, 222(602):309–368.

- 939 [20] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
- 940 *School Science and Mathematics*, 100(6):310–318.
- 941 [21] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather
- 942 prediction” task? individual variability in strategies for probabilistic category learning. *Learning*
- 943 and *Memory*, 9:408–418.
- 944 [22] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*
- 945 *Cognition and Development*, 13(1):19–37.
- 946 [23] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
- 947 learning, pages 212–221. Sage Publications.
- 948 [24] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-
- 949 ioral and neural signatures of transforming experiences into memories. *Nature Human Behavior*,
- 950 5:905–919.
- 951 [25] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge au-
- 952 tomatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*,
- 953 9:doi.org/10.3389/fpsyg.2018.00133.
- 954 [26] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-
- 955 peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–
- 956 4008.
- 957 [27] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
- 958 empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 959 [28] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
- 960 Columbia University Press.
- 961 [29] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
- 962 326(7382):213–216.

- 963 [30] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
964 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International*
965 *Journal of Environmental Research and Public Health*, 18(5):2672.
- 966 [31] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 967 [32] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 968 [33] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
969 *The Chronicle of Higher Education*, 21:1–5.
- 970 [34] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic
971 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
972 104:211–240.
- 973 [35] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic
974 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.
- 975 [36] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of*
976 *Educational Studies*, 53(2):129–147.
- 977 [37] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
978 *Handbook of Human Memory*. Oxford University Press.
- 979 [38] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum "memory wave"
980 function? *Psychological Review*, 128(4):711–725.
- 981 [39] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension
982 for adding LLM-driven chatbots to interactive notebooks. <https://github.com/ContextLab/chatify>.
- 984 [40] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and
985 projection for dimension reduction. *arXiv*, 1802(03426).

- 986 [41] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold
987 Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 988 [42] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
989 mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 990 [43] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
991 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
992 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 993 [44] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
994 tations in vector space. *arXiv*, 1301.3781.
- 995 [45] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
996 from a national survey of language educators. *System*, 97:102431.
- 997 [46] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
998 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 999 [47] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
1000 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*
1001 *Neuroscience*, 17(4):367–376.
- 1002 [48] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 1003 [49] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models.
1004 *arXiv*, 2208.02957.
- 1005 [50] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
1006 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
1007 7:43916.
- 1008 [51] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*
1009 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.

- 1010 [52] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.
1011 *Biological Cybernetics*, 45(1):35–41.
- 1012 [53] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
1013 higher education: unmasking power and raising questions about the movement’s democratic
1014 potential. *Educational Theory*, 63(1):87–110.
- 1015 [54] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
1016 Student conceptions and conceptual learning in science. Routledge.
- 1017 [55] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
1018 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*
1019 *tion in Nursing*, 22:32–42.
- 1020 [56] Shim, T. E. and Lee, S. Y. (2020). College students’ experience of emergency remote teaching
1021 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 1022 [57] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
1023 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*
1024 *Mathematics Education*, 35(5):305–329.
- 1025 [58] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal*
1026 *Medicine*, 21:524–530.
- 1027 [59] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,
1028 Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023).
1029 LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 1030 [60] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Pio-
1031 ntkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust
1032 detection of AI-generated texts. *arXiv*, 2306.04723.
- 1033 [61] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold?

- 1034 distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of*
1035 *the Cognitive Science Society*, 43(43).
- 1036 [62] Viswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and
1037 Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing*
1038 *Systems*.
- 1039 [63] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
1040 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 1041 [64] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
1042 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 1043 [65] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is
1044 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*
1045 *Research Methods*, 50:2597–2605.