

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 estimate each participant’s knowledge at the given question’s embedding space coordinate, using all *other* questions that participant answered on the same quiz. For each quiz, we grouped these estimates into two distributions: one for the estimated knowledge at the coordinates of *correctly* answered questions, and another for the estimated knowledge at the coordinates of *incorrectly* answered questions (Fig. 6). We then used Mann-Whitney U-tests to compare the means of these distributions of estimated knowledge.

For the initial quizzes participants took (prior to watching either lecture), participants’ estimated knowledge tended to be low overall, and relatively unstructured (Fig. 6, left column). When we held out individual questions and estimated their knowledge at the held-out questions’ embedding coordinates, we found no reliable differences in the estimates when the held-out question had been correctly versus incorrectly answered. This “null” effect persisted when we used *all* of the Quiz 1 questions from a given participant to predict a held-out question (“All questions”; $U = 50587, p = 0.723$), when we used questions from one lecture to predict knowledge at the embedding coordinate of a held-out question about the *other* lecture (“Across-lecture”; predicting knowledge for held-out *Four Fundamental Forces Questions* using *Birth of Stars* questions: $U = 8244, p = 0.184$; predicting knowledge for held-out *Birth of Stars* questions: $U = 8202.5, p = 0.161$), and when we used questions from one lecture to predict knowledge at the embedding coordinate of a held-out question about the *same* lecture (“Within-lecture”; *Four Fundamental Forces*: $U = 7681.5, p = 0.746$; *Birth of Stars*: $U = 8125, p = 0.204$). We believe that this reflects a floor effect: when knowledge is low everywhere,

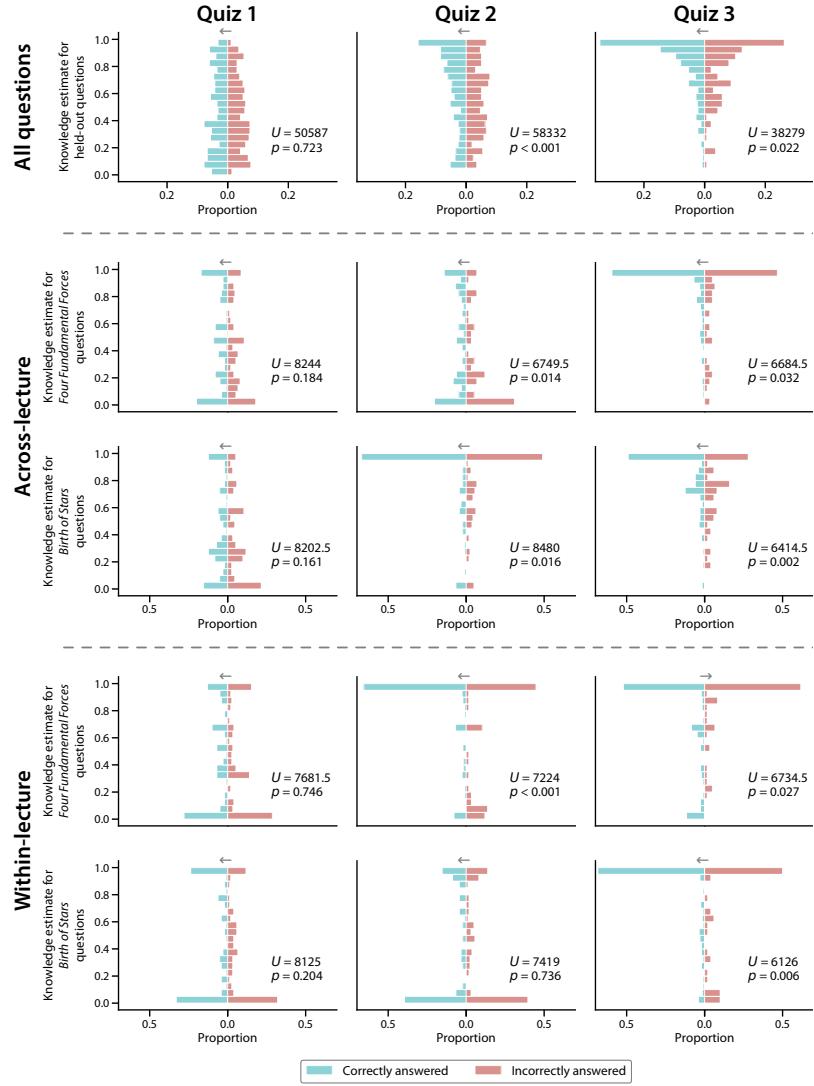


Figure 6: Estimating knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held-out correctly (blue) or incorrectly (red) answered question. The Mann-Whitney U-tests reported in each panel are between the distributions of estimated knowledge at the coordinates of correctly versus 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 estimate knowledge at the held-out question’s embedding coordinate. In the middle rows (“Across-lecture”), we used all questions about one lecture to estimate 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 estimate 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 there is little signal to differentiate between what is known versus unknown.

300 After watching *Four Fundamental Forces*, estimated knowledge for held-out correctly answered
301 questions (from the second quiz; Fig. 6, middle column) exhibited a positive shift relative to held-
302 out incorrectly answered questions. This held when we included all questions in the analysis
303 ($U = 58332, p < 0.001$), when we predicted knowledge across-lectures (*Four Fundamental Forces*:
304 $U = 6749.5, p = 0.014$; *Birth of Stars*: $U = 8480, p = 0.016$), and when we predicted knowledge at the
305 embedding coordinates of held-out *Four Fundamental Forces* questions using other *Four Fundamental*
306 *Forces* questions from the same quiz and participant ($U = 7224, p < 0.001$). This difference did *not*
307 hold for within-lecture predictions of *Birth of Stars* knowledge ($U = 7419, p = 0.739$). Again, we
308 suggest that this might reflect a floor effect whereby knowledge about the content of the *Birth of*
309 *Stars* material is relatively low everywhere in that region of text embedding space.

310 Finally, after watching *Birth of Stars*, estimated knowledge for held-out correctly answered
311 questions (from the third quiz; Fig. 6, right column) was higher for held-out correctly answered
312 questions than for held-out incorrectly answered questions. This held when we included all
313 questions in the analysis ($U = 38279, p = 0.022$), when we carried out across-lecture predictions
314 (*Four Fundamental Forces*: $U = 6684.5, p = 0.032$; *Birth of Stars*: $U = 6414.5, p = 0.002$), and when
315 we carried out within-lecture predictions of held-out *Birth of Stars* questions using other *Birth of*
316 *Stars* questions from the same quiz and participant ($U = 6126, p = 0.006$). However, we found
317 the *opposite* effect when we carried out within-lecture predictions of held-out *Four Fundamental*
318 *Forces* questions using other *Four Fundamental Forces* questions from the same quiz and participant
319 ($U = 6734, p = 0.027$). Specifically, held-out correctly answered Quiz 3 questions about *Four*
320 *Fundamental Forces* had reliably *lower* estimated knowledge than held-out incorrectly answered
321 questions. Speculatively, we suggest that this may reflect participants forgetting some of the *Four*
322 *Fundamental Forces* content. If this forgetting happens in a relatively “random” way (with respect
323 to spatial distance within the text embedding space), then it could explain why some held-out
324 questions about *Four Fundamental Forces* were answered incorrectly, even if questions at nearby
325 coordinates (i.e., about similar content) were answered correctly. This might lead our approach
326 to over-estimate knowledge for held-out questions about “forgotten” knowledge that participants

327 answered incorrectly.

328 That the knowledge estimates derived from the text embedding space reliably distinguish
329 between held-out correctly versus incorrectly answered questions (Fig. 6) suggests that the text
330 embedding space bears at least some relationship to participants’ knowledge. But what does that
331 relationship look like as we move through the embedding space? For example, suppose we know
332 that a participant answers a question (at embedding coordinate X) correctly. As we move away
333 from X in the embedding space, how does quiz performance “fall off” with distance? Or, suppose
334 the participant instead answered that same question *incorrectly*. Again, as we move away from X
335 in the embedding space, how do the chances that the participant does *not* know about the content
336 change with distance? We reasoned that, assuming our space is capturing something about how
337 participants actually organize their knowledge, conceptual knowledge right around X should be
338 similar to the participant’s knowledge of the content at X . And at another extreme, at some distance
339 (after moving sufficiently far away from X), our guesses about what participants know (based on
340 their response to the question at location X) should be no better than guessing based on their
341 overall proportion of correctly answered questions—i.e., if Y is very far away from X , all we can
342 do with the participant’s response to X is guess that “their performance on quiz questions about Y
343 is about equal to their average performance on quiz questions about any material.”

344 With these ideas in mind, we asked: conditioned on answering a question correctly, what
345 proportion of all questions (within some radius, r , of that question’s embedding coordinate)
346 were answered correctly? We plotted this proportion as a function of r . Similarly, we could
347 ask, conditioned on answering a question incorrectly, how the proportion of correct responses
348 changed with r . As shown in Figure 7, we found that quiz performance falls off smoothly with
349 distance, and the “rate” of the falloff does not appear to change across the different quizzes, as
350 measured by the distance at which performance becomes statistically indistinguishable from a
351 simple proportion correct score (see *Estimating the “smoothness” of knowledge*). This suggests that,
352 at least within the region of text embedding space covered by the questions our participants
353 answered (and as characterized using our topic model), the rate at which knowledge changes
354 with distance is relatively constant, even as participants’ overall level of knowledge varies across

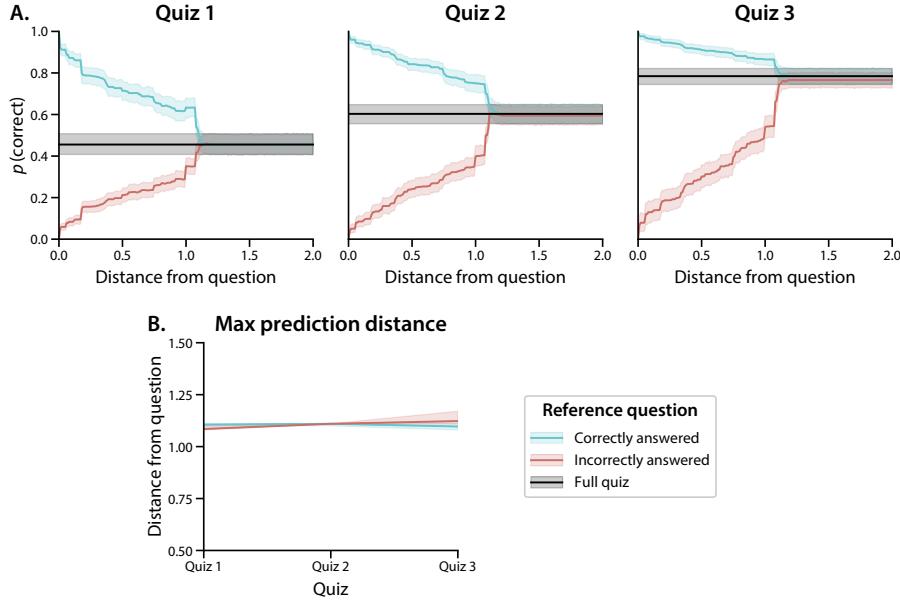


Figure 7: Quiz performance falls off gradually in text embedding space. A. Performance versus distance. For each participant, for each correctly answered question (blue) or incorrectly answered question (red), we computed the proportion of correctly answered questions within a given distance of that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

355 quizzes or regions of the embedding space.

356 Knowledge estimates need not be limited to the content of the lectures. As illustrated in
 357 Figure 8, our general approach to estimating knowledge from a small number of quiz questions
 358 may be extended to *any* content, given its text embedding coordinate. To visualize how knowledge
 359 “spreads” through text embedding space to content beyond the lectures participants watched, we
 360 first fit a new topic model to the lectures’ sliding windows with (up to) $k = 100$ topics. Conceptually,
 361 increasing the number of topics used by the model functions to increase the “resolution” of the
 362 embedding space, providing a greater ability to estimate knowledge for content that is highly
 363 similar to (but not precisely the same as) that contained in the two lectures. This change in the

364 number of topics overcame an undesirable behavior in the UMAP embedding procedure [40],
365 whereby embedding coordinates for the 15-topic model tended to be “clumped” into separated
366 clusters, rather than forming a smooth trajectory through the 2D space. When we increased the
367 number of topics to 100, the embedding coordinates in the 2D space formed a smooth trajectory
368 through the space, with substantially less clumping (Fig. 8). We note that we used these 2D maps
369 solely for visualization; all relevant comparisons, distance computations, and statistical tests we
370 report above were carried out in the original 15-dimensional space, using the 15-topic model. Aside
371 from increasing the number of topics from 15 to 100, all other procedures and model parameters
372 were carried over from the preceding analyses. As in our other analyses, we resampled each
373 lecture’s topic trajectory to 1 Hz and projected each question into a shared text embedding space.

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

383 Several features of the resulting knowledge maps are worth noting. The average knowledge
384 map estimated from Quiz 1 responses (Fig. 8A, leftmost map) shows that participants tended to
385 have relatively little knowledge about any parts of the text embedding space (i.e., the shading is
386 relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked
387 increase in knowledge on the left side of the map (around roughly the same range of coordinates
388 traversed by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words,
389 participants’ estimated increase in knowledge is localized to conceptual content that is nearby (i.e.,
390 related to) the content from the lecture they watched prior to taking Quiz 2. This localization is
391 non-trivial: these knowledge estimates are informed only by the embedded coordinates of the

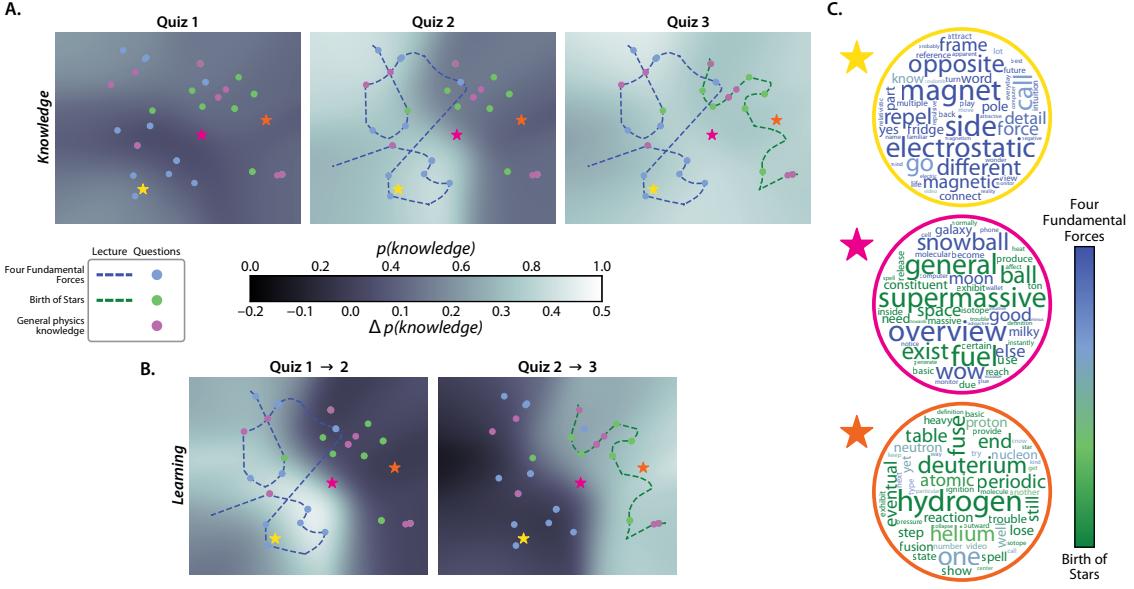


Figure 8: Mapping out the geometry of knowledge and learning. **A.** Average “knowledge maps” estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*).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

392 *quiz questions*, not by the embeddings of either lecture (see Eqn. 4). Finally, the knowledge map
393 estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region
394 surrounding the embedding of the *Birth of Stars* lecture participants watched immediately prior to
395 taking Quiz 3.

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

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

416 **Discussion**

417 We developed a computational framework that uses short multiple-choice quizzes to gain nuanced
418 insights into what learners know and how their knowledge changes with training. First, we show
419 that our approach can automatically match the conceptual knowledge probed by individual quiz
420 questions to the corresponding moments in lecture videos when those concepts were presented
421 (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment “knowledge traces”
422 that reflect the degree of knowledge participants have about each video’s time-varying content,
423 and capture temporally specific increases in knowledge after viewing each lecture (Fig. 5). We
424 also show that these knowledge estimates can generalize to held-out questions (Fig. 6). Finally,
425 we use our framework to construct visual maps that provide snapshot estimates of how much
426 participants know about any concept within the scope of our text embedding model, and how
427 much their knowledge of those concepts changes with training (Fig. 8).

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

442 In our study, we characterize the “coordinates” of participants’ knowledge using a relatively

443 simple “bag of words” text embedding model [LDA; 6]. More sophisticated text embedding mod-
444 els, such as transformer-based models [15, 48, 59, 62] can learn complex grammatical and semantic
445 relationships between words, higher-order syntactic structures, stylistic features, and more. We
446 considered using transformer-based models in our study, but we found that the text embeddings
447 derived from these models were surprisingly uninformative with respect to differentiating or oth-
448 erwise characterizing the conceptual content of the lectures and questions we used. We suspect
449 that this reflects a broader challenge in constructing models that are high-resolution within a given
450 domain (e.g., the domain of physics lectures and questions) *and* sufficiently broad so as to enable
451 them to cover a wide range of domains. For example, we found that the embeddings derived even
452 from much larger and more modern models like BERT [15], GPT [62], LLaMa [59], and others that
453 are trained on enormous text corpora, end up yielding poor resolution within the content space
454 spanned by individual course videos (Supp. Fig. 6). Whereas the LDA embeddings of the lectures
455 and questions are “near” each other (i.e., the convex hull enclosing the two lectures’ trajectories is
456 highly overlapping with the convex hull enclosing the questions’ embeddings), the BERT embed-
457 dings of the lectures and questions are instead largely distinct (top row of Supp. Fig. 6). The LDA
458 embeddings of the questions for each lecture and the corresponding lecture’s trajectory are also
459 similar. For example, as shown in Fig. 2C, the LDA embeddings for *Four Fundamental Forces* ques-
460 tions (blue dots) appear closer to the *Four Fundamental Forces* lecture trajectory (blue line), whereas
461 the LDA embeddings for *Birth of Stars* questions (green dots) appear closer to the *Birth of Stars*
462 lecture trajectory (green line). The BERT embeddings of the lectures and questions do not show
463 this property (Supp. Fig. 6). We also examined per-question “content matches” between individual
464 questions and individual moments of each lecture (Figs. 4, 6). The time series plot of individual
465 questions’ correlations are different from each other when computed using LDA (e.g., the traces
466 can be clearly visually separated), whereas the correlations computed from BERT embeddings of
467 different questions all look very similar. This tells us that LDA is capturing some differences in
468 content between the questions, whereas BERT is not. The time series plots of individual ques-
469 tions’ correlations have clear “peaks” when computed using LDA, but not when computed using
470 BERT. This tells us that LDA is capturing a “match” between the content of each question and a

471 relatively well-defined time window of the corresponding lectures. The BERT embeddings appear
472 to blur together the content of the questions versus specific moments of each lecture. Finally, we
473 also compared the pairwise correlations between embeddings of questions within versus across
474 content areas (i.e., content covered by the individual lectures, lecture-specific questions, and by the
475 “general physics knowledge” questions). The LDA embeddings show a strong contrast between
476 same-content embeddings versus across-content embeddings. In other words, the embeddings of
477 questions about the *Four Fundamental Forces* material are highly correlated with the embeddings of
478 the *Four Fundamental Forces* lecture, but not with the embeddings of *Birth of Stars*, questions about
479 *Birth of Stars*, or general physics knowledge questions. We see a similar pattern with the LDA
480 embeddings of the *Birth of Stars* questions (Fig. 3, Supp. Fig. 2). In contrast, the BERT embeddings
481 are all highly correlated with each other (Supp. Fig. 6). Taken together, these comparisons illus-
482 trate how LDA (trained on the specific content in question) provides both coverage of the requisite
483 material and specificity at the level of the content covered by individual questions. BERT, on the
484 other hand, essentially assigns both lectures and all of the questions (which are all broadly about
485 “physics”) into a tiny region of its embedding space, thereby blurring out meaningful distinctions
486 between different specific concepts covered by the lectures and questions. We note that these are
487 not criticisms of BERT (or other large language models trained on large and diverse corpora).
488 Rather, our point is that simple fine-tuned models trained on a relatively small but specialized
489 corpus can outperform much more complicated models trained on much larger corpora, when we
490 are specifically interested in capturing subtle conceptual differences at the level of a single course
491 lecture or question. Of course if our goal had been to find a model that generalized to many
492 different content areas, we would expect our approach to perform comparatively poorly relative to
493 BERT or other much larger models. We suggest that bridging the tradeoff between high resolution
494 within each content area versus the ability to generalize to many different content areas will be an
495 important challenge for future work in this domain.

496 Another application for large language models that does *not* require explicitly modeling the
497 content of individual lectures or questions is to leverage the models’ ability to generate text. For
498 example, generative text models like ChatGPT [48] and LLaMa [59] are already being used to build

499 a new generation of interactive tutoring systems [e.g., 39]. Unlike the approach we have taken here,
500 these generative text model-based systems do not explicitly model what learners know, or how
501 their knowledge changes over time with training. One could imagine building a hybrid system
502 that combines the best of both worlds: a large language model that can *generate* text, combined
503 with a smaller model that can *infer* what learners know and how their knowledge changes over
504 time. Such a hybrid system could potentially be used to build the next generation of interactive
505 tutoring systems that are able to adapt to learners' needs in real time, and that are able to provide
506 more nuanced feedback about what learners know and what they do not know.

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

525 We view our approach as occupying a sort of “sweet spot,” between much larger language
526 models and simple word matching-based approaches, that enables us to capture the relevant

527 conceptual content of course materials at an appropriate semantic scale. Our approach enables us
528 to accurately and consistently identify each question’s content in a way that also matches up with
529 what is presented in the lectures. In turn, this enables us to construct accurate predictions about
530 participants’ knowledge of the conceptual content tested by held-out questions (Fig. 6).

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

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

549 Alongside these questions, there is a growing desire to extend existing theories beyond the
550 domain of lab testing rooms and into real classrooms [29]. In part, this has led to a recent
551 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better
552 reflect more ethologically valid phenomena that are more directly relevant to real-world situations
553 and behaviors [46]. In turn, this has brought new challenges in data analysis and interpretation. A
554 key step towards solving these challenges will be to build explicit models of real-world scenarios

555 and how people behave in them (e.g., models of how people learn conceptual content from real-
556 world courses, as in our current study). A second key step will be to understand which sorts
557 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 2,
558 16, 43, 47, 50] might help to inform these models. A third major step will be to develop and
559 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic
560 paradigms.

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

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

579 **Materials and methods**

580 **Participants**

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

586 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
587 years). A total of 15 participants reported their gender as male and 35 participants reported their
588 gender as female. A total of 49 participants reported their native language as "English" and 1
589 reported having another native language. A total of 47 participants reported their ethnicity as
590 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants
591 reported their races as White (32 participants), Asian (14 participants), Black or African American
592 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
593 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

Participants reported their undergraduate major(s) as “social sciences” (28 participants), “natural sciences” (16 participants), “professional” (e.g., pre-med or pre-law; 8 participants), “mathematics and engineering” (7 participants), “humanities” (4 participants), or “undecided” (3 participants). Note that some participants selected multiple categories for their undergraduate major(s).

We also asked participants about the courses they had taken. In total, 45 participants reported having taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan Academy courses. Of those who reported having watched at least one Khan Academy course, 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We also asked participants about the specific courses they had watched, categorized under different subject areas. In the “Mathematics” area, participants reported having watched videos on AP Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Calculus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants), Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other videos not listed in our survey (5 participants). In the “Science and engineering” area, participants reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 participants); Physics, AP Physics I, or AP Physics II (18 participants); Biology, AP Biology; or High school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed in our survey (5 participants). We also asked participants whether they had specifically seen the videos used in our experiment. Of the 45 participants who reported having having taken at least one Khan Academy course in the past, 44 participants reported that they had not watched the *Four Fundamental Forces* video, and 1 participant reported that they were not sure whether they had watched it. All participants reported that they had not watched the *Birth of Stars* video. When we asked participants about non-Khan Academy online courses, they reported having watched or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 participants).

634 ipants), Computing (2 participants), and other categories not listed in our survey (17 participants).
635 Finally, we asked participants about in-person courses they had taken in different subject areas.
636 They reported taking courses in Mathematics (38 participants), Science and engineering (37 par-
637 ticipants), Arts and humanities (34 participants), Test preparation (27 participants), Economics
638 and finance (26 participants), Computing (14 participants), College and careers (7 participants), or
639 other courses not listed in our survey (6 participants).

640 Experiment

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

648 We then hand-created 39 multiple-choice questions: 15 about the conceptual content of *Four*
649 *Fundamental Forces* (i.e., Lecture 1), 15 about the conceptual content of *Birth of Stars* (i.e., Lecture 2),
650 and 9 questions that tested for general conceptual knowledge about basic physics (covering material
651 that was not presented in either video). One of our group's undergraduate research assistants
652 worked alongside a rotating Masters student to develop this set of questions (these researchers
653 are acknowledged in our paper for their contribution, although they did not meet the criteria for
654 authorship discussed with all team members at the start of the project, as determined by J.R.M.) The
655 senior author (J.R.M.) tasked the pair of researchers with coming up with "15 conceptual questions
656 about each lecture, along with 9 additional questions about general physics knowledge." To
657 help broaden the set of lecture-specific questions, the researchers were further instructed to work
658 through each lecture in small segments, identify what each segment was "about" conceptually,
659 and then write a question about that concept. The general physics questions were drawn from the
660 researchers' coursework along with internet searches and brainstorming with the project team and

661 other members of J.R.M.’s lab. The final set of questions (and response options) was reviewed and
662 approved by J.R.M. before we collected or analyzed the text or experimental data.

663 We note that estimating the specific “amount” of conceptual understanding that each question
664 “requires” to answer is somewhat subjective, and might even come down to the “strategy” a given
665 participant uses to answer the question at that particular moment. The full set of questions and
666 answer choices may be found in Supplementary Table 1.

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

678 **Analysis**

679 **Statistics**

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

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

684 We adapted an approach we developed in prior work [24] to embed each moment of the two
685 lectures and each question in our pool in a common representational space. Briefly, our approach

686 uses a topic model [Latent Dirichlet Allocation; 6] trained on a set of documents, to discover a
687 set of (up to) k “topics” or “themes.” Formally, each topic is defined as a distribution of weights
688 over words in the model’s vocabulary (i.e., the union of all unique words, across all documents,
689 excluding “stop words.”). Conceptually, each topic is intended to give larger weights to words
690 that are semantically related (as inferred from their tendency to co-occur in the same document).
691 After fitting a topic model, each document in the training set, or any *new* document that contains at
692 least some of the words in the model’s vocabulary, may be represented as a k -dimensional vector
693 describing how much the document (most probably) reflects each topic. To select an appropriate
694 k for our model, as a starting point, we identified the minimum number of topics that yielded
695 at least one “unused” topic (i.e., in which all words in the vocabulary were assigned uniform
696 weights) after training. This indicated that the number of topics was sufficient to capture the set
697 of latent themes present in the two lectures (from which we constructed our document corpus, as
698 described below). We found this value to be $k = 15$ topics. We found that with a limited number
699 of additional adjustments following [7], such as removing corpus-specific stop-words, the model
700 yielded (subjectively) sensible and coherent topics. The distribution of weights over words in
701 the vocabulary for each discovered topic is shown in Supplementary Figure 1, and each topic’s
702 top-weighted words may be found in Supplementary Table 2.

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

711 We fetched these automated transcripts using the `youtube-transcript-api` Python pack-
712 age [14]. The transcripts consisted of one timestamped line of text for every few seconds (mean:
713 2.34 s; standard deviation: 0.83 s) of spoken content in the video (i.e., corresponding to each indi-

714 individual caption that would appear on-screen if viewing the lecture via YouTube, and when those
715 lines would appear). We defined a sliding window length of (up to) $w = 30$ transcript lines, and
716 assigned each window a timestamp corresponding to the midpoint between the timestamps for its
717 first and last lines. This w parameter was chosen to match the same number of words per sliding
718 window (rounded to the nearest whole word, and before preprocessing) as the sliding windows
719 we defined in our prior work [24] (i.e., 185 words per sliding window).

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

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

740 We transformed each sliding window's text into a topic vector, and then used linear interpolation
741 (independently for each topic dimension) to resample the resulting time series to one vector

per second. We also used the fitted model to obtain topic vectors for each question in our pool (see Supp. Tab. 1). Taken together, we obtained a *trajectory* for each video, describing its path through topic space, and a single coordinate for each question (Fig. 2C). Embedding both videos and all of the questions using a common model enables us to compare the content from different moments of videos, compare the content across videos, and estimate potential associations between specific questions and specific moments of video.

Estimating dynamic knowledge traces

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

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

where

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

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

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

765 non-negative amount to the knowledge estimate.

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

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

772 In our bootstrap procedure, we ran 10,000 iterations to estimate the relationship between partic-
773 ipants’ performance and the distance to a given reference question. For each of these iterations, for
774 every individual quiz (q), we first determined the across-participants average “simple” proportion
775 correct and its 95% confidence interval. This interval was established by repeatedly (1,000 times)
776 subsampling participants with replacement, computing the mean “simple” proportion correct for
777 each subsample, and then deriving the 2.5th and 97.5th percentiles from the distribution of these
778 subsample means. We used this interval as our benchmark for determining whether the propor-
779 tion of correctly answered questions for a given subset of questions was reliably different (at the
780 $p < 0.05$ significance level) from the average proportion correct across all questions.

781 Next, for each participant, we examined all 15 questions they answered on quiz q . We treated
782 each question as the “reference question” in turn. Around this reference, we constructed a series of
783 15-dimensional spheres (starting with a radius of 0), where each successive sphere had a radius of
784 0.01 (correlation distance) greater than its predecessor. Within each of these spheres, we calculated
785 the proportion of questions answered correctly by the participant. This yielded two distinct sets
786 of proportion-correct values for each binned distance (radius) for a specific participant and quiz:
787 one set of values where the reference questions had been answered correctly, and another set
788 where the reference questions had been answered incorrectly. From these, we established the
789 average proportion correct within each radius for both categories of reference questions. Finally,
790 we identified the minimum binned distance from the correctly answered reference questions for
791 which the average proportion correct intersected the 95% confidence interval of the simple average

792 proportion correct computed earlier. We display the resulting distance estimates, for each quiz
793 and reference question status, in Figure 7B.

794 **Creating knowledge and learning map visualizations**

795 An important feature of our approach is that, given a trained text embedding model and partic-
796 ipants’ quiz performance on each question, we can estimate their knowledge about *any* content
797 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-
798 tions, or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 7, 8, 9, 10,
799 and 11), we used Uniform Manifold Approximation and Projection [UMAP; 40, 41] to construct a
800 2D projection of the text embedding space. Sampling the original 100-dimensional space at high
801 resolution to obtain an adequate set of topic vectors spanning the embedding space would be
802 computationally intractable. However, sampling a 2D grid is trivial.

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

810 In our application, all of the coordinates we embedded were topic vectors, whose elements
811 are always non-negative and sum to one. Although UMAP is an invertible transformation at
812 the embedding locations of the original data, other locations in the embedding space will not
813 necessarily follow the same implicit “rules” as the original high-dimensional data. For example,
814 inverting an arbitrary coordinate in the embedding space might result in negative-valued vectors,
815 which are incompatible with the topic modeling framework. To protect against this issue, we
816 log-transformed the topic vectors prior to embedding them in the 2D space. When we inverted
817 the embedded vectors (e.g., to estimate topic vectors or word clouds, as in Fig. 8C), we passed
818 the inverted (log-transformed) values through the exponential function to obtain a vector of non-

819 negative values, and normalized them to sum to one.

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

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

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

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

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

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

836 Author contributions

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

841 **Data availability**

842 All of the data analyzed in this manuscript may be found at <https://github.com/ContextLab/efficient-learning-khan>.
843

844 **Code availability**

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

847 **Acknowledgements**

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

855 **References**

- 856 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,
857 56:149–178.
- 858 [2] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and
859 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
860 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.

- 861 [3] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text*
862 *with the natural language toolkit*. Reilly Media, Inc.
- 863 [4] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
864 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
865 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 866 [5] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*
867 *Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
868 Machinery.
- 869 [6] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
870 *Learning Research*, 3:993–1022.
- 871 [7] Boyd-Graber, J. and Mimno, D. (2014). Care and feeding of topic models: problems, diagnostics,
872 and improvements. In Airolidi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E., editors,
873 *Handbook of Mixed Membership Models and Their Applications*. CRC Press.
- 874 [8] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
875 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
876 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
877 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
878 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 879 [9] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
880 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 881 [10] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
882 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
883 sentence encoder. *arXiv*, 1803.11175.
- 884 [11] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
885 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.

- 886 [12] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
887 Evidence for a new conceptualization of semantic representation in the left and right cerebral
888 hemispheres. *Cortex*, 40(3):467–478.
- 889 [13] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
890 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,
891 41(6):391–407.
- 892 [14] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 894 [15] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep
895 bidirectional transformers for language understanding. *arXiv*, 1810.04805.
- 896 [16] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
897 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
898 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 899 [17] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 900 [18] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of*
901 *Experimental Psychology: General*, 115:155–174.
- 902 [19] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*
903 *Transactions of the Royal Society A*, 222(602):309–368.
- 904 [20] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
905 *School Science and Mathematics*, 100(6):310–318.
- 906 [21] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather
907 prediction” task? individual variability in strategies for probabilistic category learning. *Learning*
908 *and Memory*, 9:408–418.
- 909 [22] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*
910 *Cognition and Development*, 13(1):19–37.

- 911 [23] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
912 learning, pages 212–221. Sage Publications.
- 913 [24] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-
914 ioral and neural signatures of transforming experiences into memories. *Nature Human Behavior*,
915 5:905–919.
- 916 [25] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge au-
917 tomatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*,
918 9:doi.org/10.3389/fpsyg.2018.00133.
- 919 [26] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-
920 peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–
921 4008.
- 922 [27] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
923 empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 924 [28] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
925 Columbia University Press.
- 926 [29] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
927 326(7382):213–216.
- 928 [30] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
929 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International
930 Journal of Environmental Research and Public Health*, 18(5):2672.
- 931 [31] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 932 [32] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 933 [33] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
934 *The Chronicle of Higher Education*, 21:1–5.

- 935 [34] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic
936 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
937 104:211–240.
- 938 [35] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic
939 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.
- 940 [36] Macellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of
941 Educational Studies*, 53(2):129–147.
- 942 [37] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
943 *Handbook of Human Memory*. Oxford University Press.
- 944 [38] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum "memory wave"
945 function? *Psychological Review*, 128(4):711–725.
- 946 [39] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension
947 for adding LLM-driven chatbots to interactive notebooks. [https://github.com/ContextLab/
948 chatify](https://github.com/ContextLab/chatify).
- 949 [40] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and
950 projection for dimension reduction. *arXiv*, 1802(03426).
- 951 [41] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold
952 Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 953 [42] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
954 mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 955 [43] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
956 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
957 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 958 [44] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
959 tations in vector space. *arXiv*, 1301.3781.

- 960 [45] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
961 from a national survey of language educators. *System*, 97:102431.
- 962 [46] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
963 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 964 [47] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
965 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*
966 *Neuroscience*, 17(4):367–376.
- 967 [48] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 968 [49] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models.
969 *arXiv*, 2208.02957.
- 970 [50] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
971 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
972 7:43916.
- 973 [51] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*
974 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 975 [52] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.
976 *Biological Cybernetics*, 45(1):35–41.
- 977 [53] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
978 higher education: unmasking power and raising questions about the movement’s democratic
979 potential. *Educational Theory*, 63(1):87–110.
- 980 [54] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
981 Student conceptions and conceptual learning in science. Routledge.
- 982 [55] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
983 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*
984 *tion in Nursing*, 22:32–42.

- 985 [56] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching
986 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 987 [57] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
988 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in
989 Mathematics Education*, 35(5):305–329.
- 990 [58] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal
991 Medicine*, 21:524–530.
- 992 [59] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,
993 Goyal, N., Hambro, E., Azhar, F., Rodriguz, A., Joulin, A., Grave, E., and Lample, G. (2023).
994 LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 995 [60] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Pio-
996 ntkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust
997 detection of AI-generated texts. *arXiv*, 2306.04723.
- 998 [61] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold?
999 distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of
1000 the Cognitive Science Society*, 43(43).
- 1001 [62] Viswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and
1002 Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing
1003 Systems*.
- 1004 [63] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
1005 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 1006 [64] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
1007 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 1008 [65] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is

1009 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*
1010 *Research Methods*, 50:2597–2605.