

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

<sup>4</sup> Paxton C. Fitzpatrick<sup>1</sup>, Andrew C. Heusser<sup>1, 2</sup>, and Jeremy R. Manning<sup>1,\*</sup>

<sup>1</sup>Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

<sup>2</sup>Akili Interactive Labs

Boston, MA 02110, USA

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

<sup>5</sup> **Abstract**

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

<sup>17</sup> **Keywords:** education, learning, knowledge, concepts, natural language processing

<sup>18</sup> **Introduction**

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

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

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

45 acquired information to the scaffolding of one’s existing knowledge or experience [6, 11, 13, 15, 30,  
46 65]? 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 [40, 70]? 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 [37, 41, 62, 69].

50 The difference between “understanding” and “memorizing,” as framed by researchers in ed-  
51 ucation, cognitive psychology, and cognitive neuroscience [e.g., 23, 28, 33, 41, 62], 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 passage  
55 might provide some ability to answer simple questions about it. However, text embedding mod-  
56 els [e.g., 7, 8, 10, 12, 16, 39, 50, 71] 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 [42?][42, 43]. To be clear, this is  
61 not 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~~ documents’ deeper conceptual meanings  
65 may be quantified, explored, and understood. According to these models, the deep conceptual  
66 meaning of a document may be captured by a feature vector in a high-dimensional representation  
67 space, wherein nearby vectors reflect conceptually related documents. A model that succeeds at  
68 capturing an analogue of “understanding” is able to assign nearby feature vectors to two conceptu-  
69 ally related documents, *even when the specific words contained in those documents have limited overlap*.  
70 In this way, “concepts” are defined implicitly by the model’s geometry [e.g., how the embedding  
71 coordinate of a given word or document relates to the coordinates of other text embeddings; 56].

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, 20, 21, 24, 31, 59]. 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 [28, 41]. To help bridge this gap, our study uses course materials from real online courses to inform, fit, and test models of real-world conceptual learning. We also provide a demonstration of how our models can be used to construct “maps” of what students know, and how their knowledge changes with training. In addition to helping to visually capture knowledge (and changes in knowledge), we hope that such maps might lead to real-world tools for improving how we educate. Taken together, our work shows that existing course materials and evaluative tools like short multiple-choice quizzes may be leveraged to gain highly detailed insights into what students know and how they learn.

## Results

At its core, our main modeling approach is based around a simple assumption that we sought to test empirically: all else being equal, knowledge about a given concept is predictive of knowledge about similar or related concepts. From a geometric perspective, this assumption implies that knowledge is fundamentally “smooth.” In other words, as one moves through a space representing an individual’s knowledge (where similar concepts occupy nearby coordinates), their “level of knowledge” should change relatively gradually. To begin to test this smoothness assumption, we sought to track participants’ knowledge and how it changed over time in response to training. Two overarching goals guide our approach. First, we want to gain detailed insights into what learners know at different points in their training. For example, rather than simply reporting on the proportions of questions participants answer correctly (i.e., their overall performance), we seek estimates of their knowledge about a variety of specific concepts. Second, we want our approach to be potentially scalable to large numbers of diverse concepts, courses, and students. This requires that the conceptual content of interest be discovered *automatically*, rather than relying on manually produced ratings or labels.



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

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 [36] 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



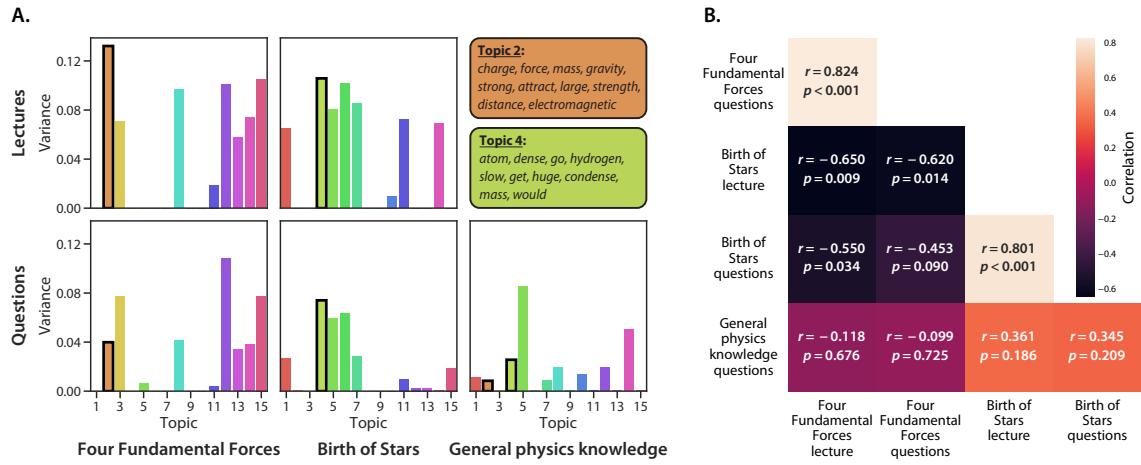
**Figure 2: Modeling course content.** **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The full set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text 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 [29]  
 154 to identify the latent themes in the lectures using a topic model [8]. 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 [the average topic weights from](#) their non-matching question sets (Lecture 1 video vs. Lecture 2 questions:  $r(13) = -0.547$ ,  $p = 0.035$ , 95% CI = [-0.812, -0.231]; Lecture 2 video vs. Lecture 1 questions:  $r(13) = -0.612$ ,  $p = 0.015$ , 95% CI = [-0.874, -0.281]), indicating that the topic model also exhibits some degree of specificity. The full set of pairwise comparisons between average topic weights for the lectures and question sets



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

188 is reported in 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” [22] 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  $95\% \text{ CI} = [0.696, 0.973]$ ) and the Lecture 2 video and questions ( $r(13) = 0.801, p < 0.001, 95\%$   
200  $\text{CI} = [0.539, 0.958]$ ). Simultaneously, as reported in Figure 3B, the variability-variabilities in topic  
201 expression across *different* videos and lecture-specific questions (i.e., Lecture 1 video vs. Lecture 2

202 questions; Lecture 2 video vs. Lecture 1 questions) were negatively correlated, and neither video's  
203 topic 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** **variabilities** in topic expression across mo-  
210 ments of each lecture and questions from its corresponding set (Fig. 3), we wondered whether the  
211 text embedding model might additionally capture these conceptual relationships at a finer scale.  
212 For example, if a particular question asks about the content from one small part of a lecture, we  
213 wondered whether the text embeddings could be used to automatically identify the "matching"  
214 moment(s) in the lecture. To explore this, we computed the correlation between each question's  
215 topic weights and the topic weights for each second of its corresponding lecture, and found that  
216 each question appeared to be temporally specific (Fig. 4). In particular, most questions' topic  
217 vectors were maximally correlated with a well-defined (and relatively narrow) range of time-  
218 points from their corresponding lectures, and the correlations fell off sharply outside of that range  
219 (Supp. Figs. 3, 4). We also qualitatively examined the best-matching intervals for each question by  
220 comparing the question's text to the **text of transcribed text from** the most-correlated parts of the  
221 lectures (Supp. Tab. 3). Despite that the questions were excluded from the text embedding model's  
222 training set, in general we found (through manual inspection) a close correspondence between  
223 the conceptual content that each question probed and the content covered by the best-matching  
224 moments of the lectures. Two 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 **administering some form of assessment (e.g., a quiz) and** computing the proportion of correctly  
229 answered questions. But if two students receive identical scores on **such** an exam, might our



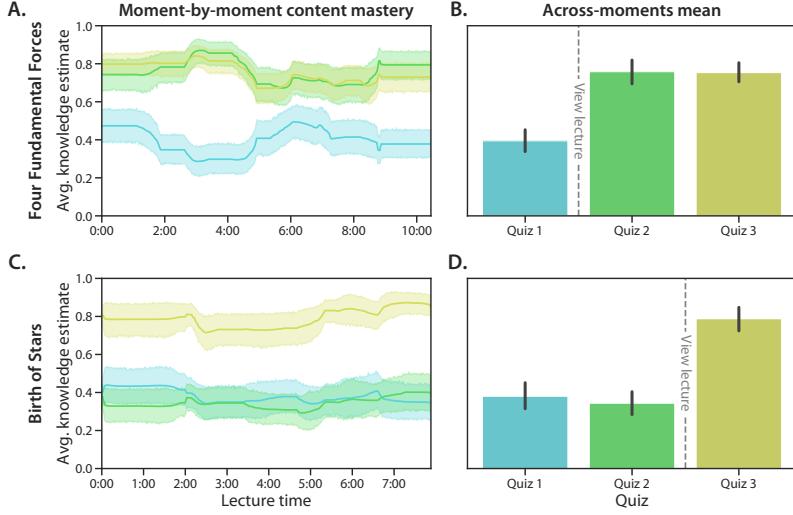
**Figure 4: Which parts of each lecture are captured by each question?** Each panel displays time series plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel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

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

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

243 moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially, the  
244 estimated knowledge at coordinate  $x$  is given by the weighted ~~average~~ proportion of quiz questions  
245 the participant answered correctly, where the weights reflect how much each question is “about”  
246 the content at  $x$ . When we apply this approach to estimate the participant’s knowledge about the  
247 content presented in each moment of each lecture, we can obtain a detailed time course describing  
248 how much “knowledge” ~~the~~that participant has about the content presented at any part of the  
249 lecture. As shown in Figure 5A and C, we can apply this approach separately for the questions  
250 from each quiz participants took throughout the experiment. From just a few questions per quiz  
251 (see *Estimating dynamic knowledge traces*), we obtain a high-resolution snapshot (at the time each  
252 quiz was taken) of what the participants knew about any moment’s content, from either of the two  
253 lectures they watched (comprising a total of 1,100 samples across the two lectures).

254 While the time courses in Figure 5A and C provide detailed *estimates* about participants’  
255 knowledge, these estimates are of course only *useful* to the extent that they accurately reflect what  
256 participants actually know. As one sanity check, we anticipated that the knowledge estimates  
257 should reflect a content-specific “boost” in participants’ knowledge after watching each lecture. In  
258 other words, if participants learn about each lecture’s content ~~when they watch each lecture upon~~  
259 watching it, the knowledge estimates should capture that. After watching the *Four Fundamental*  
260 *Forces* lecture, participants should exhibit more knowledge for the content of that lecture than they  
261 had before, and that knowledge should persist for the remainder of the experiment. Specifically,  
262 knowledge about that lecture’s content should be relatively low when estimated using Quiz 1  
263 responses, but should increase when estimated using Quiz 2 or 3 responses (Fig. 5B). Indeed, we  
264 found that participants’ estimated knowledge about the content of *Four Fundamental Forces* was  
265 substantially higher on Quiz 2 versus Quiz 1 ( $t(49) = 8.764, p < 0.001$ ) and on Quiz 3 versus Quiz 1  
266 ( $t(49) = 10.519, p < 0.001$ ). We found no reliable differences in estimated knowledge about that  
267 lecture’s content on Quiz 2 versus 3 ( $t(49) = 0.160, p = 0.874$ ). Similarly, we hypothesized (and  
268 subsequently confirmed) that participants should show greater estimated knowledge about the  
269 content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 5D). Specifically, since  
270 participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their

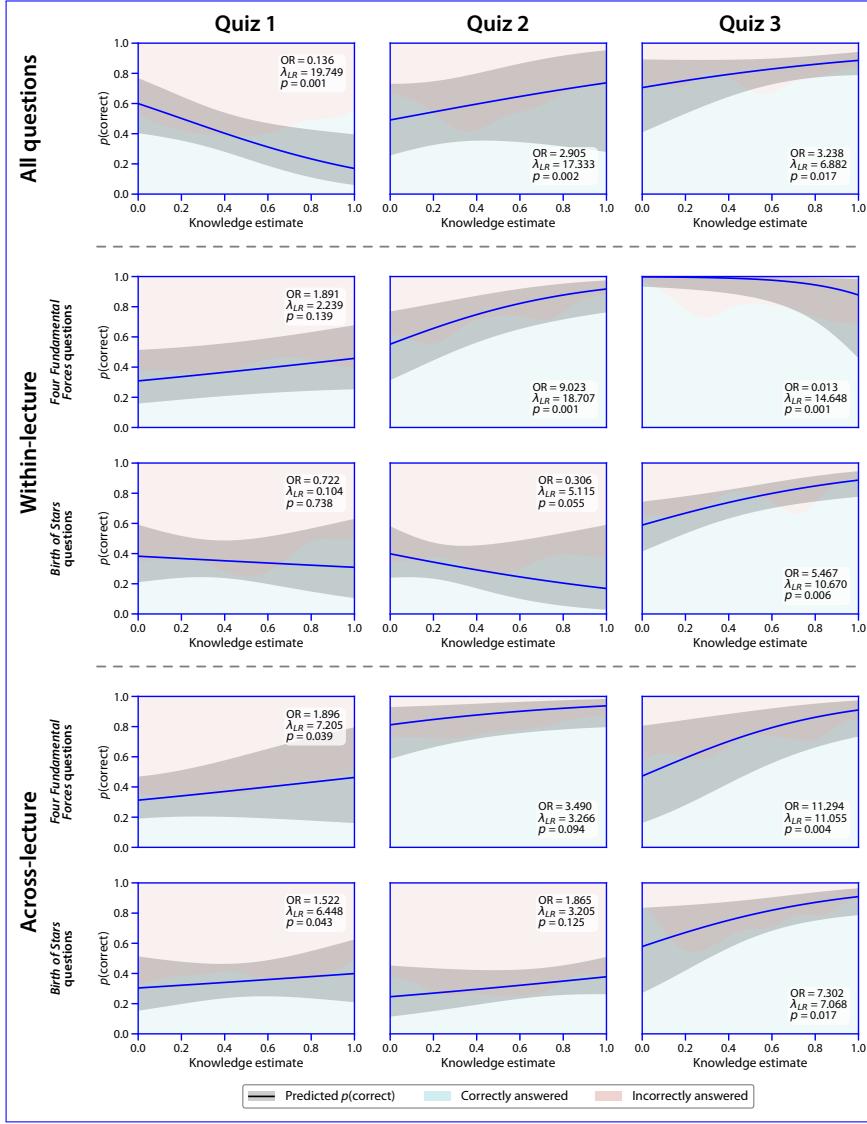


**Figure 5: Estimating knowledge about the content presented at each moment of each lecture.** **A. Knowledge about the time-varying content of *Four Fundamental Forces*.** Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (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*.** **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 they are~~ likely to answer that question correctly or incorrectly. We developed a statistical approach to test this claim. For each ~~question~~ ~~quiz~~ ~~question a participant answered~~, in turn, we used Equation 1 to ~~predict each participant’s estimate their~~ knowledge at the given question’s embedding space coordinate ~~, using all other based on other~~ questions that participant answered on the same quiz. ~~For each~~ ~~We repeated this for all participants, and for each of the three quizzes. Then, separately for each~~ quiz, we ~~grouped these predicted knowledge values into two distributions: one for the predicted knowledge at the coordinates of correctly answered questions, and another for the predicted knowledge at the coordinates of incorrectly answered questions (Fig. 6).~~ We then used Mann–Whitney U tests to compare the means of these distributions of predicted knowledge ~~fit a generalized linear mixed model (GLMM) with a logistic link function to explain the likelihood of correctly answering a question as a function of estimated knowledge for its embedding coordinate, while accounting for random variation among participants and questions (see *Generalized linear mixed models*).~~ To assess the predictive value of the knowledge estimates, we compared each GLMM to an analogous (i.e., nested) “null” model ~~that did not consider estimated knowledge using parametric bootstrap likelihood-ratio tests.~~

We carried out ~~these analyses in three different ways. First, we used all (but one) of the questions from a given quiz (and participant) to predict knowledge at the embedding coordinate of a held-out question~~ three different versions of the analyses described above, wherein we considered different sources of information in our estimates of participants’ knowledge for each quiz question. First, we estimated knowledge at each question’s embedding coordinate using ~~all other questions answered by the same participant on the same quiz~~ (“All questions”~~in~~ ~~; Fig. 6, top row~~). This test was



**Figure 6: Predicting knowledge at the embedding coordinates of held-out questions.** Predicting success on held-out questions using estimated knowledge. We used generalized linear mixed models (GLMMs) to model the likelihood of correctly answering a quiz question as a function of estimated knowledge for its embedding coordinate (see *Generalized linear mixed models*). Separately for each quiz (column), we plot the distributions examined this relationship based on three different sets of predicted knowledge at the embedding coordinates of estimates: knowledge for each held-out correctly (blue) or incorrectly (red) answered question. The Mann-Whitney U tests reported in each panel are between-based on all other questions the distributions of predicted knowledge at the coordinates of correctly and incorrectly same participant answered held-out questions. In on the top row same quiz ("All questions"; top row), we used all quiz questions (from each quiz, knowledge for each participant) except one to predict knowledge at the held-out question's embedding coordinate. In the middle rows ("Across-lecture"), we used all questions about one lecture to predict knowledge at based on all other questions (from the embedding coordinate of a held-out question same participant and quiz) about the other same lecture. In the bottom row ("Within-lecture"; middle rows), we used all but one and knowledge for each question about one lecture to predict knowledge at based on all questions (from the embedding coordinate of a held-out question same participant and quiz) about the same other lecture ("Across-lecture"; bottom rows). We repeated each of these analyses using all possible held-out questions for each quiz and participant. The arrows at the tops of backgrounds in each panel indicate whether display kernel density estimates of the average predicted knowledge was higher for held-out relative observed proportions of correctly answered (left blue) or versus incorrectly answered (right red) answered questions, for each level of estimated knowledge along the x-axis. The black curves display the (population-level) GLMM-predicted probabilities of correctly answering a question as a function of estimated knowledge. Error ribbons denote 95% confidence intervals.

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

For the initial quizzes participants took (prior to watching either lecture), predicted knowledge tended to be low overall, and relatively unstructured (Fig. 6, left column). When we held out individual questions and predicted their knowledge at the held-out questions' embedding coordinates, we found no reliable differences in the predictions when the In performing these analyses, our null hypothesis is that the knowledge estimates we compute based on the quiz questions' embedding coordinates do not provide useful information about participants' abilities to answer those questions. What result might we expect to see if this is the case? To gain an intuition for this scenario, consider the expected outcome if we carried out these same analyses using a simple proportion-correct measure in lieu of our knowledge estimates. Suppose a participant correctly answered  $n$  out of  $q$  questions on a given quiz. If we hold out a single correctly answered question, the proportion of remaining questions answered correctly would be  $\frac{n-1}{q-1}$ . Whereas if we hold out a single incorrectly answered question, the proportion of remaining questions answered correctly

327 would be  $\frac{n}{q-1}$ . In this way, the proportion of correctly answered remaining questions is always  
328 lower when the held-out question had been correctly versus incorrectly answered. This “null” effect  
329 persisted when we used all of the Quiz 1 questions from a given participant to predict a held-out  
330 question (“All questions”;  $U = 50587$ ,  $p = 0.723$ ), when we used questions from one lecture to  
331 predict knowledge at the embedding coordinate of a question was answered correctly than when  
332 it was answered incorrectly. Because our knowledge estimates are computed as a weighted  
333 version of this same proportion-correct score (where each held-in question’s weight reflects its  
334 embedding-space distance from the held-out question about the other lecture (“Across-lecture”;  
335 predicting knowledge for question; see Eqn. 1), if these weights are uninformative (e.g., randomly  
336 distributed), then we should expect to see this same inverse relationship between estimated  
337 knowledge and performance, on average. On the other hand, if the spatial relationships among  
338 the quiz questions’ embeddings are predictive of participants’ knowledge about the questions’  
339 content, then we would expect higher estimated knowledge for held-out *Four Fundamental Forces*  
340 Questions using *Birth of Stars* questions:  $U = 8244$ ,  $p = 0.184$ ; predicting knowledge for held-out  
341 *Birth of Stars* questions:  $U = 8202.5$ ,  $p = 0.161$ ), and when we used questions from one lecture to  
342 predict correctly (versus incorrectly) answered questions.

343 Before presenting our results, it is worth considering three possible explanations of why a  
344 participant might answer a given question correctly or incorrectly. One possibility is that the  
345 participant simply guessed the answer. A second is that they selected an answer by mistake, despite  
346 “knowing” the correct answer. In both of these scenarios, the participant’s knowledge about the  
347 question’s content should be uninformative about their observed response. A third possibility is  
348 that the participant’s response reflects their *actual* knowledge about the question’s content. In this  
349 case, we might expect to see a positive relationship between the participant’s knowledge and their  
350 likelihood of answering the question correctly. However, in order to see this positive relationship,  
351 the participant’s knowledge must be structured in a way that is reflected (at least partially) by the  
352 embedding space. In other words, if the participant’s performance reflects their true knowledge,  
353 but our text embedding space does not sufficiently capture the structure of that knowledge, then  
354 the knowledge estimates we generate will not be predictive of the participant’s performance. In the

355 extreme, if the embedding space is completely unstructured with respect to the content of the quiz  
356 questions, then we would expect to see the negative relationship between estimated knowledge  
357 and performance that we described above.

358 When we fit a GLMM to estimates of participants' knowledge for each Quiz 1 question based  
359 on all other Quiz 1 questions, we observed an outcome consistent with our null hypothesis:  
360 higher estimated knowledge at the embedding coordinate of a held-out question about the  
361 same lecture ("Within lecture"; *Four Fundamental Forces*:  $U = 7681.5, p = 0.746$ ; *Birth of Stars*:  
362  $U = 8125, p = 0.204$ ). We believe that this reflects a floor effect: when knowledge is low everywhere,  
363 there is little signal to differentiate between what is known versus unknown. was associated with  
364 a lower likelihood of answering the question correctly (odds ratio (*OR*) = 0.136, likelihood-ratio  
365 test statistic ( $\lambda_{LR}$ ) = 19.749, 95% CI = [14.352, 26.545],  $p = 0.001$ ). This outcome suggests that our  
366 knowledge estimates do not provide useful information about participants' Quiz 1 performance  
367 when we aggregated across all question content areas. We speculated that this might either  
368 indicate that the knowledge estimates are uninformative in general, or about Quiz 1 performance  
369 in particular. This would be expected, for example, if participants were guessing about the answers  
370 to the Quiz 1 questions (prior to having watched either lecture). When we repeated this analysis for  
371 Quizzes 2 and 3, we found that higher estimated knowledge for a given question predicted a greater  
372 likelihood of answering it correctly (Quiz 2: *OR* = 2.905,  $\lambda_{LR} = 17.333$ , 95% CI = [14.966, 29.309],  $p = 0.002$ ;  
373 Quiz 3: *OR* = 3.238,  $\lambda_{LR} = 6.882$ , 95% CI = [6.228, 8.184],  $p = 0.017$ ). Taken together, these results  
374 suggest that our knowledge estimates reliably predict participants' performance on individual  
375 held-out quiz questions, but only after participants have received at least some training.

376 After watching *Four Fundamental Forces*, predicted knowledge for We observed a similar pattern  
377 of results when used this approach to estimate participants' knowledge about held-out questions  
378 that were answered correctly (from the second quiz; Fig. 6, middle column) exhibited a significant  
379 positive shift relative to held-out questions that were answered incorrectly. This held when  
380 we included all questions in the analysis ( $U = 58332, p < 0.001$ ), when we predicted knowledge  
381 across lectures (*Four Fundamental Forces*:  $U = 6749.5, p = 0.014$ ; *Birth of Stars*:  $U = 8480, p = 0.016$ ),  
382 and when we predicted knowledge at the questions from one lecture using their performance on

383 other questions from the same lecture. Specifically, for Quiz 1 questions (i.e., prior to watching  
384 either), participants' estimated knowledge for the embedding coordinates of held-out *Four Fundamental Forces*  
385 questions-related questions estimated using other *Four Fundamental Forces* questions  
386 from the same quiz and participant ( $U = 7224, p < 0.001$ ). This difference did not hold for -related  
387 questions did not reliably predict whether those questions were answered correctly ( $OR = 1.891, \lambda_{LR} = 2.293, 95\% CI = [$   
388 The same was true of knowledge estimates for held-out *Birth of Stars*-related questions based on  
389 other *Birth of Stars*-related questions ( $OR = 0.722, \lambda_{LR} = 5.115, 95\% CI = [0.094, 0.146], p = 0.738$ ).  
390 As in our analysis that included all questions, we speculate that these "null" results might reflect  
391 some degree of random guessing on Quiz 1. When we repeated these within-lecture knowledge  
392 predictions at knowledge at embedding space coordinates of analyses using questions from Quiz 2  
393 (which participants took immediately after viewing *Four Fundamental Forces* but prior to viewing  
394 *Birth of Stars* questions), we found that they now reliably predicted success on *Four Fundamental*  
395 *Forces*-related questions ( $OR = 9.023, \lambda_{LR} = 18.707, 95\% CI = [10.877, 22.222], p = 0.001$ ) but not  
396 on *Birth of Stars*-related questions ( $U = 7419, p = 0.739$ ). Again, we suggest that this might reflect  
397 a floor effect whereby, at that point in the participants' training, their knowledge about the  
398 content of the  $OR = 0.306, \lambda_{LR} = 5.115, 95\% CI = [4.624, 5.655], p = 0.055$ ). Here, we speculate  
399 that participants might have been guessing about the *Birth of Stars* material is relatively low  
400 everywhere in that region of text embedding space.

401 Finally, after watching *Birth of Stars*, predicted knowledge for held-out correctly answered  
402 questions (from the third quiz; Fig. 6, right column) was higher than for held-out incorrectly  
403 answered questions. This held when we included all questions in the analysis ( $U = 38279, p = 0.022$ ),  
404 when we carried out across-lecture predictions (content (e.g., prior to having watched it), whereas  
405 they might have been drawing on some structured knowledge about the *Four Fundamental Forces*:  
406  $U = 6684.5, p = 0.032$ ; content (e.g., from having just watched it). When we applied this approach  
407 to Quiz 3 responses (given immediately after viewing *Birth of Stars*:  $U = 6414.5, p = 0.002$ ), and  
408 and when we carried out ), we found that within-lecture knowledge predictions for held-out  
409 estimates for *Birth of Stars* questions using other *Birth of Stars* questions from the same quiz and  
410 participant ( $U = 6126, p = 0.006$ -related questions could now reliably predict success on those

411 questions ( $OR = 5.467$ ,  $\lambda_{LR} = 10.670$ , 95% CI = [7.998, 12.532],  $p = 0.006$ ). However, we found the  
412 opposite effect when we carried out within-lecture knowledge predictions for held-out estimates  
413 for Four Fundamental Forces questions using other Four Fundamental Forces questions from the  
414 same quiz and participant ( $U = 6734$ ,  $p = 0.027$ ). Specifically, on Quiz answered on Quiz 3,  
415 our knowledge predictions for held-out correctly answered questions about Four Fundamental  
416 Forces were reliably lower than those for their incorrectly answered counterparts. were no longer  
417 directly related to the likelihood of successfully answering them and instead exhibited the inverse  
418 relationship we would expect to arise from unstructured knowledge (with respect to the embedding  
419 space;  $OR = 0.013$ ,  $\lambda_{LR} = 14.648$ , 95% CI = [10.695, 23.096],  $p = 0.001$ ). Speculatively, we suggest  
420 that this may reflect participants forgetting some of the Four Fundamental Forces content (e.g.,  
421 perhaps in favor of prioritizing encoding the just-watched Birth of Stars content in preparation for  
422 the third quiz). If this forgetting happens in a relatively “random” way (with respect to spatial  
423 distance within the text-embedding space), then it could explain why some held-out questions  
424 about Four Fundamental Forces were answered incorrectly, even if questions at nearby coordinates  
425 (i.e., about similar content) were answered correctly. This might lead our approach to over-estimate  
426 knowledge for held-out questions about “forgotten” knowledge that participants answered incor-  
427 rectly. Taken together, the results in Figure 6 indicate these within-lecture results suggest that  
428 our approach can reliably predict acquired knowledge (especially about recently learned content)  
429 , and distinguish between questions about different content covered by a single lecture when  
430 participants have sufficiently structured knowledge about its contents, though this specificity may  
431 decrease with time since the relevant material was learned.

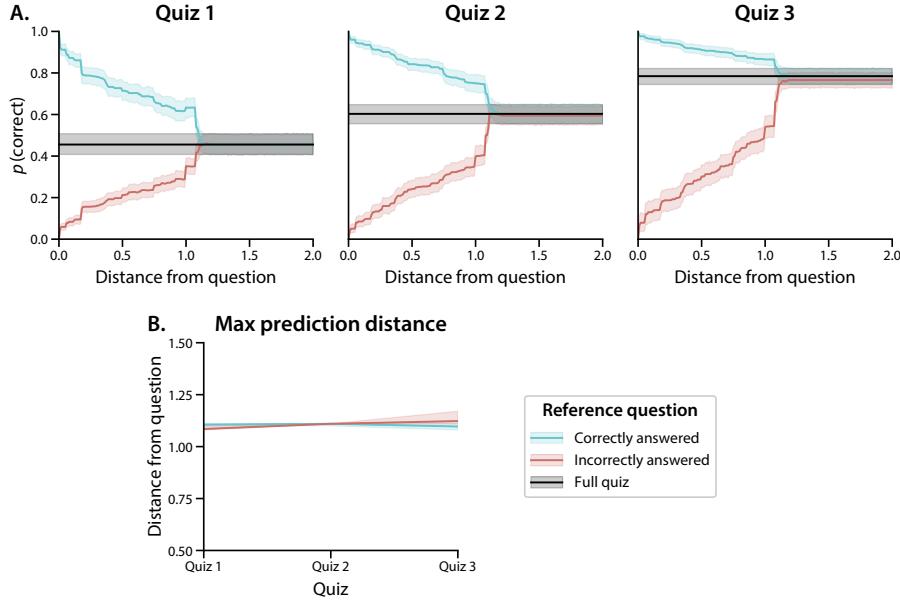
432 Finally, we used this approach to estimate participants’ knowledge about held-out questions  
433 from one lecture using their performance on questions from the other lecture. Here we again  
434 observed a similar pattern of results, though with some notable differences. On Quiz 1, we found  
435 that participants’ abilities to correctly answer questions about Four Fundamental Forces could be  
436 predicted from their responses to questions about Birth of Stars ( $OR = 1.896$ ,  $\lambda_{LR} = 7.205$ , 95% CI = [6.224, 7.524],  $p = 0.000$ )  
437 and similarly, that their ability to correctly answer Birth of Stars-related questions could be predicted  
438 from their responses to Four Fundamental Forces-related questions ( $OR = 1.522$ ,  $\lambda_{LR} = 6.448$ , 95% CI = [5.656, 6.843],  $p = 0.000$ ).

Given the results from our analyses that included all questions and within-lecture predictions, we were surprised to find that the knowledge predictions are generalizable across the content areas spanned by the two lectures, while also specific enough to estimates could reliably (if weakly) predict participants' performance across content from different lectures. It is possible that this result reflects a combination of random guessing prior to training (leading to a weak effect size), alongside some coarse-scale structured knowledge that participants had about the content prior to watching either lecture. When we repeated this analysis using questions from Quiz 2, we found participants' responses to *Four Fundamental Forces*-related questions did not reliably predict their success on *Birth of Stars*-related questions ( $OR = 1.865, \lambda_{LR} = 3.205, 95\% CI = [3.027, 3.600], p = 0.125$ ), nor did their responses to *Birth of Stars*-related questions reliably predict their success on *Four Fundamental Forces*-related questions ( $OR = 3.490, \lambda_{LR} = 3.266, 95\% CI = [3.033, 3.866], p = 0.094$ ). These "prediction failures" appear to come from the fact that any signal derived from participants' knowledge about the content of the *Birth of Stars* lecture (prior to watching it) is swamped by the much more dramatic increase in their knowledge about the content of the *Four Fundamental Forces* (which they watched just prior to taking Quiz 2). This is reflected in their Quiz 2 performance for questions about each lecture (mean proportion correct for *Four Fundamental Forces*-related questions on Quiz 2: 0.77; mean proportion correct for *Birth of Stars*-related questions on Quiz 2: 0.36). When we carried out these across-lecture knowledge predictions using questions from Quiz 3 (when participants had now viewed both lectures), we could again reliably predict success on questions about both *Four Fundamental Forces* ( $OR = 11.294, \lambda_{LR} = 11.055, 95\% CI = [9.126, 18.476], p = 0.004$ ) and *Birth of Stars* ( $OR = 7.302, \lambda_{LR} = 7.068, 95\% CI = [6.490, 8.584], p = 0.017$ ) using responses to questions about the other lecture's content. Across all three versions of these analyses, our results suggest that (by and large) our knowledge estimates can reliably predict participants' abilities to answer individual quiz questions, distinguish between questions about more subtly different content within the same lecture, similar content, and generalize across content areas, provided that participants' quiz responses reflect a minimum level of "real" knowledge about both content on which these predictions are based and that for which they are made. Our results also indicate some important limitations of our approach: if participants' quiz performance does not reflect what they know

467 (e.g., when they “guess”), or if their knowledge is not structured in a way that is reflected by the  
468 embedding space, then our knowledge estimates will not be predictive of their performance.

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

488 With these ideas in mind, we asked: conditioned on answering a question correctly, what  
489 proportion of all questions (within some radius,  $r$ , of that question’s embedding coordinate)  
490 were answered correctly? We plotted this proportion as a function of  $r$ . Similarly, we could  
491 ask, conditioned on answering a question incorrectly, how the proportion of correct responses  
492 changed with  $r$ . As shown in Figure 7, we found that quiz performance falls off smoothly with  
493 distance, and the “rate” of the falloff does not appear to change across the different quizzes, as  
494 measured by the distance at which performance becomes statistically indistinguishable from a



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

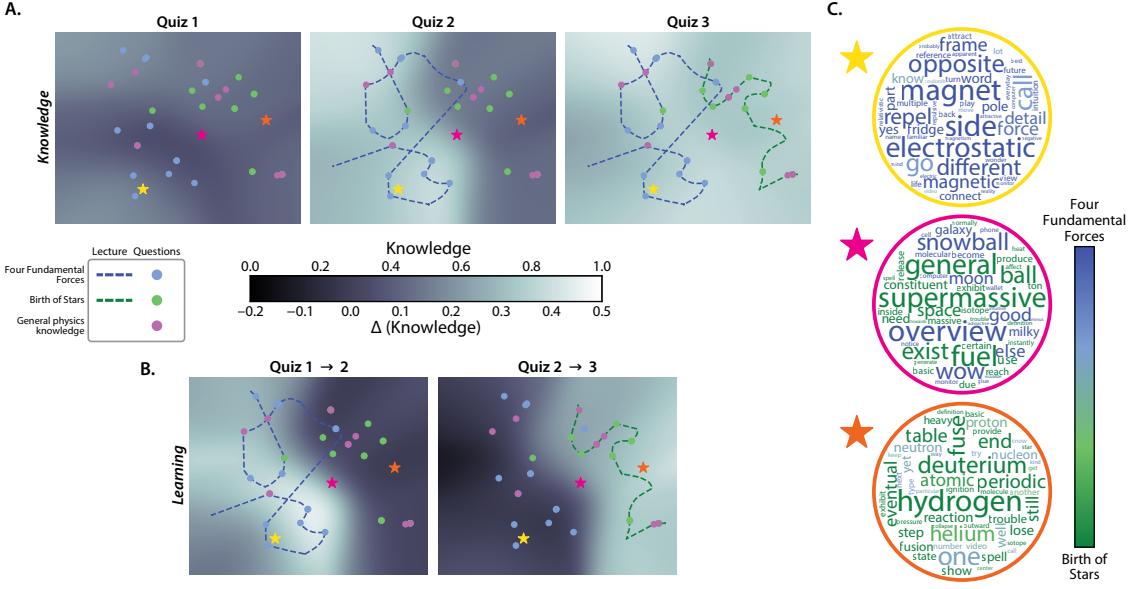
495 simple proportion correct score (see *Estimating the “smoothness” of knowledge*). This suggests that,  
 496 at least within the region of text embedding space covered by the questions our participants  
 497 answered (and as characterized using our topic model), the rate at which knowledge changes  
 498 with distance is relatively constant, even as participants’ overall level of knowledge varies across  
 499 quizzes or regions of the embedding space.

500 Knowledge estimates need not be limited to the content of the lectures. As illustrated in  
 501 Figure 8, our general approach to estimating knowledge from a small number of quiz questions  
 502 may be extended to *any* content, given its text embedding coordinate. To visualize how knowledge  
 503 “spreads” through text embedding space to content beyond the lectures participants watched, we

504 first fit a new topic model to the lectures' sliding windows with  $k = 100$  topics. Conceptually,  
505 increasing the number of topics used by the model functions to increase the "resolution" of the  
506 embedding space, providing a greater ability to estimate knowledge for content that is highly  
507 similar to (but not precisely the same as) that contained in the two lectures. We note that we  
508 used these 2D maps solely for visualization; all relevant comparisons, distance computations, and  
509 statistical tests we report above were carried out in the original 15-dimensional space, using the  
510 15-topic model. Aside from increasing the number of topics from 15 to 100, all other procedures  
511 and model parameters were carried over from the preceding analyses. As in our other analyses,  
512 we resampled each lecture's topic trajectory to 1 Hz and projected each question into a shared text  
513 embedding space.

514 We projected the resulting 100-dimensional topic vectors (for each second of video and each quiz  
515 question) onto a shared 2-dimensional plane (see *Creating knowledge and learning map visualizations*).  
516 Next, we sampled points from a  $100 \times 100$  grid of coordinates that evenly tiled a rectangle enclosing  
517 the 2D projections of the videos and questions. We used Equation 4 to estimate participants'  
518 knowledge at each of these 10,000 sampled locations, and averaged these estimates across par-  
519 ticipants to obtain an estimated average *knowledge map* (Fig. 8A). Intuitively, the knowledge map  
520 constructed from a given quiz's responses provides a visualization of how "much" participants  
521 knew about any content expressible by the fitted text embedding model at the point in time when  
522 they completed that quiz.

523 Several features of the resulting knowledge maps are worth noting. The average knowledge  
524 map estimated from Quiz 1 responses (Fig. 8A, leftmost map) shows that participants tended to  
525 have relatively little knowledge about any parts of the text embedding space (i.e., the shading is  
526 relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked  
527 increase in knowledge on the left side of the map (around roughly the same range of coordinates  
528 traversed by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words,  
529 participants' estimated increase in knowledge is localized to conceptual content that is nearby (i.e.,  
530 related to) the content from the lecture they watched prior to taking Quiz 2. This localization is  
531 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.

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

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

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

556 **Discussion**

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

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

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

583 simple “bag of words” text embedding model [LDA; 8]. More sophisticated text embedding mod-  
584 els, such as transformer-based models [18, 55, 68, 71] can learn complex grammatical and semantic  
585 relationships between words, higher-order syntactic structures, stylistic features, and more. We  
586 considered using transformer-based models in our study, but we found that the text embeddings  
587 derived from these models were surprisingly uninformative with respect to differentiating or oth-  
588 erwise characterizing the conceptual content of the lectures and questions we used. We suspect  
589 that this reflects a broader challenge in constructing models that are high-resolution within a given  
590 domain (e.g., the domain of physics lectures and questions) *and* sufficiently broad so as to enable  
591 them to cover a wide range of domains. For example, we found that the embeddings derived even  
592 from much larger and more modern models like BERT [18], GPT [71], LLaMa [68], and others that  
593 are trained on enormous text corpora, end up yielding poor resolution within the content space  
594 spanned by individual course videos (Supp. Fig. 6). Whereas the LDA embeddings of the lectures  
595 and questions are “near” each other (i.e., the convex hull enclosing the two lectures’ trajectories is  
596 highly overlapping with the convex hull enclosing the questions’ embeddings), the BERT embed-  
597 dings of the lectures and questions are instead largely distinct (top row of Supp. Fig. 6). The LDA  
598 embeddings of the questions for each lecture and the corresponding lecture’s trajectory are also  
599 similar. For example, as shown in Fig. 2C, the LDA embeddings for *Four Fundamental Forces* ques-  
600 tions (blue dots) appear closer to the *Four Fundamental Forces* lecture trajectory (blue line), whereas  
601 the LDA embeddings for *Birth of Stars* questions (green dots) appear closer to the *Birth of Stars*  
602 lecture trajectory (green line). The BERT embeddings of the lectures and questions do not show  
603 this property (Supp. Fig. 6). We also examined per-question “content matches” between individual  
604 questions and individual moments of each lecture (Figs. 4, 6). The time series plot of individual  
605 questions’ correlations are different from each other when computed using LDA (e.g., the traces  
606 can be clearly visually separated), whereas the correlations computed from BERT embeddings of  
607 different questions all look very similar. This tells us that LDA is capturing some differences in  
608 content between the questions, whereas BERT is not. The time series plots of individual ques-  
609 tions’ correlations have clear “peaks” when computed using LDA, but not when computed using  
610 BERT. This tells us that LDA is capturing a “match” between the content of each question and a

611 relatively well-defined time window of the corresponding lectures. The BERT embeddings appear  
612 to blur together the content of the questions versus specific moments of each lecture. Finally, we  
613 also compared the pairwise correlations between embeddings of questions within versus across  
614 content areas (i.e., content covered by the individual lectures, lecture-specific questions, and by the  
615 “general physics knowledge” questions). The LDA embeddings show a strong contrast between  
616 same-content embeddings versus across-content embeddings. In other words, the embeddings of  
617 questions about the *Four Fundamental Forces* material are highly correlated with the embeddings of  
618 the *Four Fundamental Forces* lecture, but not with the embeddings of *Birth of Stars*, questions about  
619 *Birth of Stars*, or general physics knowledge questions. We see a similar pattern with the LDA  
620 embeddings of the *Birth of Stars* questions (Fig. 3, Supp. Fig. 2). In contrast, the BERT embeddings  
621 are all highly correlated with each other (Supp. Fig. 6). Taken together, these comparisons illus-  
622 trate how LDA (trained on the specific content in question) provides both coverage of the requisite  
623 material and specificity at the level of the content covered by individual questions. BERT, on the  
624 other hand, essentially assigns both lectures and all of the questions (which are all broadly about  
625 “physics”) into a tiny region of its embedding space, thereby blurring out meaningful distinctions  
626 between different specific concepts covered by the lectures and questions. We note that these are  
627 not criticisms of BERT (or other large language models trained on large and diverse corpora).  
628 Rather, our point is that simple fine-tuned models trained on a relatively small but specialized  
629 corpus can outperform much more complicated models trained on much larger corpora, when we  
630 are specifically interested in capturing subtle conceptual differences at the level of a single course  
631 lecture or question. Of course if our goal had been to find a model that generalized to many  
632 different content areas, we would expect our approach to perform comparatively poorly relative to  
633 BERT or other much larger models. We suggest that bridging the tradeoff between high resolution  
634 within each content area versus the ability to generalize to many different content areas will be an  
635 important challenge for future work in this domain.

636 Another application for large language models that does *not* require explicitly modeling the  
637 content of individual lectures or questions is to leverage the models’ abilities to generate text. For  
638 example, generative text models like ChatGPT [55] and LLaMa [68] are already being used to build

639 a new generation of interactive tutoring systems [e.g., 44]. Unlike the approach we have taken here,  
640 these generative text model-based systems do not explicitly model what learners know, or how  
641 their knowledge changes over time with training. One could imagine building a hybrid system  
642 that combines the best of both worlds: a large language model that can *generate* text, combined  
643 with a smaller model that can *infer* what learners know and how their knowledge changes over  
644 time. Such a hybrid system could potentially be used to build the next generation of interactive  
645 tutoring systems that are able to adapt to learners' needs in real time, and that are able to provide  
646 more nuanced feedback about what learners know and what they do not know.

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

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

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

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

677 Within the past several years, the global pandemic forced many educators to suddenly adapt to  
678 teaching remotely [35, 51, 64, 72]. This change in world circumstances is happening alongside (and  
679 perhaps accelerating) geometric growth in the availability of high-quality online courses from plat-  
680 forms such as Khan Academy [36], Coursera [73], EdX [38], and others [60]. Continued expansion  
681 of the global internet backbone and improvements in computing hardware have also facilitated  
682 improvements in video streaming, enabling videos to be easily shared and viewed by increasingly  
683 large segments of the world's population. This exciting time for online course instruction provides  
684 an opportunity to re-evaluate how we, as a global community, educate ourselves and each other.  
685 For example, we can ask: what defines an effective course or training program? Which aspects of  
686 teaching might be optimized and/or augmented by automated tools? How and why do learning  
687 needs and goals vary across people? How might we lower barriers to receiving a high-quality  
688 education?

689 Alongside these questions, there is a growing desire to extend existing theories beyond the  
690 domain of lab testing rooms and into real classrooms [34]. In part, this has led to a recent  
691 resurgence of "naturalistic" or "observational" experimental paradigms that attempt to better  
692 reflect more ethologically valid phenomena that are more directly relevant to real-world situations  
693 and behaviors [52]. In turn, this has brought new challenges in data analysis and interpretation. A  
694 key step towards solving these challenges will be to build explicit models of real-world scenarios

695 and how people behave in them (e.g., models of how people learn conceptual content from real-  
696 world courses, as in our current study). A second key step will be to understand which sorts  
697 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 4,  
698 19, 49, 53, 57] might help to inform these models. A third major step will be to develop and  
699 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic  
700 paradigms.

701 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also  
702 relate to the notion of “theory of mind” of other individuals [26, 32, 48]. Considering others’ unique  
703 perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and  
704 communicate [58, 63, 67]. One could imagine future extensions of our work (e.g., analogous to  
705 the knowledge and learning maps shown in Fig. 8), that attempt to characterize how well-aligned  
706 different people’s knowledge bases or backgrounds are. In turn, this might be used to model how  
707 knowledge (or other forms of communicable information) flows not just between teachers and  
708 students, but between friends having a conversation, individuals on a first date, participants at  
709 a business meeting, doctors and patients, experts and non-experts, political allies or adversaries,  
710 and more. For example, the extent to which two people’s knowledge maps “match” or “align” in  
711 a given region of text embedding space might serve as a predictor of how effectively they will be  
712 able to communicate about the corresponding conceptual content.

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

719 **Materials and methods**

720 **Participants**

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

726 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09  
727 years). A total of 15 participants reported their gender as male and 35 participants reported their  
728 gender as female. A total of 49 participants reported their native language as "English" and 1  
729 reported having another native language. A total of 47 participants reported their ethnicity as  
730 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants  
731 reported their races as White (32 participants), Asian (14 participants), Black or African American  
732 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other  
733 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

746 Participants reported their undergraduate major(s) as “social sciences” (28 participants), “nat-  
747 ural sciences” (16 participants), “professional” (e.g., pre-med or pre-law; 8 participants), “mathe-  
748 matics and engineering” (7 participants), “humanities” (4 participants), or “undecided” (3 partici-  
749 pants). Note that some participants selected multiple categories for their undergraduate major(s).  
750 We also asked participants about the courses they had taken. In total, 45 participants reported hav-  
751 ing taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan  
752 Academy courses. Of those who reported having watched at least one Khan Academy course,  
753 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8  
754 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We  
755 also asked participants about the specific courses they had watched, categorized under different  
756 subject areas. In the “Mathematics” area, participants reported having watched videos on AP  
757 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-  
758 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry  
759 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential  
760 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),  
761 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other  
762 videos not listed in our survey (5 participants). In the “Science and engineering” area, participants  
763 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-  
764 ipants); Physics, AP Physics I, or AP Physics II (18 participants); Biology, AP Biology; or High  
765 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed  
766 in our survey (5 participants). We also asked participants whether they had specifically seen the  
767 videos used in our experiment. Of the 45 participants who reported having having taken at least  
768 one Khan Academy course in the past, 44 participants reported that they had not watched the *Four*  
769 *Fundamental Forces* video, and 1 participant reported that they were not sure whether they had  
770 watched it. All participants reported that they had not watched the *Birth of Stars* video. When  
771 we asked participants about non-Khan Academy online courses, they reported having watched  
772 or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test  
773 preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 partic-

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

780 **Experiment**

781 We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces*  
782 (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force;  
783 duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed;  
784 duration: 7 minutes and 57 seconds). All participants viewed the videos in the same order (i.e.,  
785 *Four Fundamental Forces* followed by *Birth of Stars*).

786 We then hand-created 39 multiple-choice questions: 15 about the conceptual content of *Four*  
787 *Fundamental Forces* (i.e., Lecture 1), 15 about the conceptual content of *Birth of Stars* (i.e., Lecture 2),  
788 and 9 questions that tested for general conceptual knowledge about basic physics (covering material  
789 that was not presented in either video). To help broaden the set of lecture-specific questions,  
790 our team worked through each lecture in small segments to identify what each segment was  
791 “about” conceptually, and then write a question about that concept. The general physics questions  
792 were drawn our team’s prior coursework and areas of interest, along with internet searches and  
793 brainstorming with the project team and other members of J.R.M.’s lab. Although we attempted to  
794 design the questions to test “conceptual knowledge,” we note that estimating the specific “amount”  
795 of conceptual understanding that each question “requires” to answer is somewhat subjective, and  
796 might even come down to the “strategy” a given participant uses to answer the question at that  
797 particular moment. The full set of questions and answer choices may be found in Supplementary  
798 Table 1. The final set of questions (and response options) was reviewed and approved by J.R.M.  
799 before we collected or analyzed the text or experimental data.

800 Over the course of the experiment, participants completed three 13-question multiple-choice

801 quizzes: the first before viewing Lecture 1, the second between Lectures 1 and 2, and the third  
802 after viewing Lecture 2 (see Fig. 1). The questions appearing on each quiz, for each participant,  
803 were randomly chosen from the full set of 39, with the constraints that (a) each quiz contained  
804 exactly 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general  
805 physics knowledge, and (b) each question appear exactly once for each participant. The orders of  
806 questions on each quiz, and the orders of answer options for each question, were also randomized.  
807 We obtained informed consent from all participants, and our experimental protocol was approved  
808 by the Committee for the Protection of Human Subjects at Dartmouth College. We used this  
809 experiment to develop and test our computational framework for estimating knowledge and  
810 learning.

811 **Analysis**

812 **Statistics**

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

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

817 We adapted an approach we developed in prior work [29] to embed each moment of the two  
818 lectures and each question in our pool in a common representational space. Briefly, our approach  
819 uses a topic model [Latent Dirichlet Allocation; 8] trained on a set of documents, to discover a set  
820 of  $k$  “topics” or “themes.” Formally, each topic is defined as a distribution of weights over words  
821 in the model’s vocabulary (i.e., the union of all unique words, across all documents, excluding  
822 “stop words.”). Conceptually, each topic is intended to give larger weights to words that are  
823 semantically related (as inferred from their tendency to co-occur in the same document). After  
824 fitting a topic model, each document in the training set, or any *new* document that contains at  
825 least some of the words in the model’s vocabulary, may be represented as a  $k$ -dimensional vector

describing how much the document (most probably) reflects each topic. To select an appropriate  $k$  for our model, as a starting point, we identified the minimum number of topics that yielded at least one “unused” topic (i.e., in which all words in the vocabulary were assigned uniform weights) after training. This indicated that the number of topics was sufficient to capture the set of latent themes present in the two lectures (from which we constructed our document corpus, as described below). We found this value to be  $k = 15$  topics. We found that with a limited number of additional adjustments following [9]Boyd-Graber et al. [9], such as removing corpus-specific stop-words, the model yielded (subjectively) sensible and coherent topics. The distribution of weights over words in the vocabulary for each discovered topic is shown in Supplementary Figure 1, and each topic’s top-weighted words may be found in Supplementary Table 2.

As illustrated in Figure 2A, we start by building up a corpus of documents using overlapping sliding windows that span each video’s transcript. Khan Academy provides professionally created, manual transcriptions of all videos for closed captioning. However, such transcripts would not be readily available in all contexts to which our framework could potentially be applied. Khan Academy videos are hosted on the YouTube platform, which additionally provides automated captions. We opted to use these automated transcripts [which, in prior work, we have found to be of sufficiently near-human quality to yield reliable data in behavioral studies; 74] when developing our framework in order to make it more directly extensible and adaptable by others in the future.

We fetched these automated transcripts using the `youtube-transcript-api` Python package [17]. The transcripts consisted of one timestamped line of text for every few seconds (mean: 2.34 s; standard deviation: 0.83 s) of spoken content in the video (i.e., corresponding to each individual caption that would appear on-screen if viewing the lecture via YouTube, and when those lines would appear). We defined a sliding window length of (up to)  $w = 30$  transcript lines, and assigned each window a timestamp corresponding to the midpoint between the timestamps for its first and last lines. This  $w$  parameter was chosen to match the same number of words per sliding window (rounded to the nearest whole word, and before preprocessing) as the sliding windows we defined in our prior work [29] (i.e., 185 words per sliding window)[29; i.e., 185 words per sliding window]

.

854 These sliding windows ramped up and down in length at the beginning and end of each  
855 transcript, respectively. In other words, each transcript's first sliding window covered only its first  
856 line, the second sliding window covered the first two lines, and so on. This ensured that each line  
857 from the transcripts appeared in the same number ( $w$ ) of sliding windows. We next performed a  
858 series of standard text preprocessing steps: normalizing case, lemmatizing, removing punctuation  
859 and removing stop-words. We constructed our corpus of stop words by augmenting the Natural  
860 Language Toolkit [NLTK; 5] English stop word list with the following additional words, selected  
861 using one of the approaches suggested by [9]Boyd-Graber et al. [9]: "actual," "actually," "also,"  
862 "bit," "could," "e," "even," "first," "follow," "following," "four," "let," "like," "mc," "really,"  
863 "saw," "see," "seen," "thing," and "two." This yielded sliding windows with an average of 73.8  
864 remaining words, and lasting for an average of 62.22 seconds. We treated the text from each sliding  
865 window as a single "document," and combined these documents across the two videos' windows  
866 to create a single training corpus for the topic model.

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

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

882 **Estimating dynamic knowledge traces**

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

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

885 where

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

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

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

900 **Generalized linear mixed models**

901 In the set of analyses reported in Figure 6, we assessed whether estimates of participants' knowledge  
902 at the embedding coordinates of individual quiz questions could be used to reliably predict their  
903 ability to correctly answer those questions. In essence, we treated each question a given participant

904 answered on a given quiz as a “lecture” consisting of a single timepoint, and used Equation 1 to  
905 estimate the participant’s knowledge for its embedding coordinate based on their performance on  
906 all other questions they answered on that same quiz (“All questions”; Fig. 6, top row). Additionally,  
907 for each lecture-related question (i.e., excluding questions about general physics knowledge), we  
908 computed analogous knowledge estimates based on all other questions the participant answered  
909 on the same quiz about (1) the same lecture as the target question (“Within-lecture”; Fig. 6, middle  
910 rows), and (2) the other of the two lectures (“Across-lecture”; Fig. 6, bottom rows).

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

$$\text{accuracy} \sim \text{knowledge} + (\text{knowledge} | \text{participant}) + (\text{knowledge} | \text{question})$$

917 where “accuracy” is a binary value indicating whether each target question was answered  
918 correctly or incorrectly, “knowledge” is estimated knowledge at each target question’s embedding  
919 coordinate, “participant” is a unique identifier assigned to each participant, and “question” is a  
920 unique identifier assigned to each quiz question. For models we fit using knowledge estimates for  
921 target questions about multiple content areas (i.e., in the “All questions” version of the analysis),  
922 we also included an additional random effect term,  $(\text{knowledge} | \text{lecture})$ , where “lecture” is a  
923 categorical value denoting whether the target question was about *Four Fundamental Forces*, *Birth*  
924 *of Stars*, or general physics knowledge. Note that with our coding scheme, identifiers for each  
925 question are implicitly nested within levels of lecture and do not require explicit nesting in  
926 our model formula. We then iteratively removed random effects from the maximal model until  
927 it successfully converged with a full rank (i.e., non-singular) random effects variance-covariance  
928 matrix.

929 To assess the predictive value of our knowledge estimates, we compared each GLMM's ability  
930 to discriminate between correctly and incorrectly answered questions to that of an analogous model  
931 that did *not* consider estimated knowledge. Specifically, we used the same sets of observations  
932 with which we fit each "full" model to fit a second "null" model, with the formula:

$$\text{accuracy} \sim (1 | \text{participant}) + (1 | \text{question})$$

933 where "accuracy", "participant", and "question" are as defined above. As with our full models,  
934 the null models we fit for the "All questions" version of the analysis for each quiz contained an  
935 additional term,  $(1 | \text{lecture})$ , where "lecture" is as defined above. We then compared each  
936 full model to its reduced (null) equivalent using a likelihood-ratio test (LRT). Because the typical  
937 asymptotic  $\chi^2$  approximation of the null distribution for the LRT statistic ( $\lambda_{LR}$ ) is anti-conservative  
938 for models that differ in their random slope terms [25, 61, 66], we computed *p*-values for these  
939 tests using a parametric bootstrapping procedure [27]. For each of 1,000 bootstraps, we used the  
940 fitted null model to simulate a sample of observations of equal size to our original sample. We  
941 then re-fit both the null and full models to this simulated sample and compared them via an LRT.  
942 This yielded a distribution of  $\lambda_{LR}$  statistics we may expect to observe under our null hypothesis.  
943 Following [14, 54], we computed a corrected *p*-value for our observed  $\lambda_{LR}$  as  $\frac{r+1}{n+1}$ , where  $r$  is the  
944 number of simulated model comparisons that yielded a  $\lambda_{LR}$  greater than or equal to our observed  
945 value and  $n$  is the number of simulations we ran (1,000).

946 **Estimating the "smoothness" of knowledge**

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

952 For each of 10,000 iterations, we drew a random subsample (with replacement) of 50 participants

953 from our dataset **full dataset**. Within each iteration, we first computed the 95% confidence interval  
954 (CI) of the across-subsample-participants mean proportion correct on each of the three quizzes,  
955 separately. To compute this interval for each quiz, we repeatedly (1,000 times) subsampled par-  
956 ticipants (with replacement, from the outer subsample for the current iteration) and computed  
957 the mean proportion correct of each of these inner subsamples. We then identified the 2.5<sup>th</sup> and  
958 97.5<sup>th</sup> percentiles of the resulting distributions of 1,000 means. These three intervals (one for each  
959 quiz) served as our thresholds for confidence that the proportion correct within a given distance  
960 from a reference question was reliably different (at the  $p < 0.05$  significance level) from the average  
961 proportion correct across all questions on the given quiz.

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

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

980 This resulted in two “intersection” distances for each quiz (for correctly answered and incor-

981 rectly answered reference questions). Repeating this full process for each of the 10,000 bootstrap  
982 iterations output two distributions of intersection distances for each of the three quizzes. The  
983 means and 95% confidence intervals for these distributions are plotted in Figure 7B.

984 **Creating knowledge and learning map visualizations**

985 An important feature of our approach is that, given a trained text embedding model and partic-  
986 ipants' quiz performance on each question, we can estimate their knowledge about *any* content  
987 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-  
988 tions, or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 7, 8, 9, 10,  
989 and 11), we used Uniform Manifold Approximation and Projection [UMAP; 46, 47] to construct a  
990 2D projection of the text embedding space. Whereas our main analyses used a 15-topic embedding  
991 space, we used a 100-topic embedding space for these visualizations. This change in the number  
992 of topics overcame an undesirable behavior in the UMAP embedding procedure, whereby embed-  
993 ding coordinates for the 15-topic model tended to be “clumped” into separated clusters, rather  
994 than forming a smooth trajectory through the 2D space. When we increased the number of topics  
995 to 100, the embedding coordinates in the 2D space formed a smooth trajectory through the space,  
996 with substantially less clumping (Fig. 8). Creating a “map” by sampling this 100-dimensional  
997 space at high resolution to obtain an adequate set of topic vectors spanning the embedding space  
998 would be computationally intractable. However, sampling a 2D grid is trivial.

999 At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing  
1000 the cross-entropy between the pairwise (clustered) distances between the observations in their  
1001 original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional  
1002 embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise  
1003 distances in the original high-dimensional space were defined as 1 minus the correlation between  
1004 each pair of coordinates, and pairwise distances in the low-dimensional embedding space were  
1005 defined as the Euclidean distance between each pair of coordinates.

1006 In our application, all of the coordinates we embedded were topic vectors, whose elements  
1007 are always non-negative and sum to one. Although UMAP is an invertible transformation at

1008 the embedding locations of the original data, other locations in the embedding space will not  
1009 necessarily follow the same implicit “rules” as the original high-dimensional data. For example,  
1010 inverting an arbitrary coordinate in the embedding space might result in negative-valued vectors,  
1011 which are incompatible with the topic modeling framework. To protect against this issue, we  
1012 log-transformed the topic vectors prior to embedding them in the 2D space. When we inverted  
1013 the embedded vectors (e.g., to estimate topic vectors or word clouds, as in Fig. 8C), we passed  
1014 the inverted (log-transformed) values through the exponential function to obtain a vector of non-  
1015 negative values, and normalized them to sum to one.

1016 After embedding both lectures’ topic trajectories and the topic vectors of every question, we  
1017 defined a rectangle enclosing the 2D projections of the lectures’ and quizzes’ embeddings. We then  
1018 sampled points from a regular  $100 \times 100$  grid of coordinates that evenly tiled this enclosing rectangle.  
1019 We sought to estimate participants’ knowledge (and learning, i.e., changes in knowledge) at each  
1020 of the resulting 10,000 coordinates.

1021 To generate our estimates, we placed a set of 39 radial basis functions (RBFs) throughout the  
1022 embedding space, centered on the 2D projections for each question (i.e., we included one RBF for  
1023 each question). At coordinate  $x$ , the value of an RBF centered on a question’s coordinate  $\mu$ , is given  
1024 by:

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

1025 The  $\lambda$  term in the RBF equation controls the “smoothness” of the function, where larger values  
1026 of  $\lambda$  result in smoother maps. In our implementation we used  $\lambda = 50$ . Next, we estimated the  
1027 “knowledge” at each coordinate,  $x$ , using:

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

1028 Intuitively, Equation 4 computes the weighted proportion of correctly answered questions, where  
1029 the weights are given by how nearby (in the 2D space) each question is to the  $x$ . We also defined  
1030 *learning maps* as the coordinate-by-coordinate differences between any pair of knowledge maps.  
1031 Intuitively, learning maps reflect the *change* in knowledge across two maps.

1032 **Author contributions**

1033 Conceptualization: P.C.F., A.C.H., and J.R.M. Methodology: P.C.F., A.C.H., and J.R.M. Software:  
1034 P.C.F. Validation: P.C.F. Formal analysis: P.C.F. Resources: P.C.F., A.C.H., and J.R.M. Data curation:  
1035 P.C.F. Writing (original draft): J.R.M. Writing (review and editing): P.C.F., A.C.H., and J.R.M. Visu-  
1036 alization: P.C.F. and J.R.M. Supervision: J.R.M. Project administration: P.C.F. Funding acquisition:  
1037 J.R.M.

1038 **Data availability**

1039 All of the data analyzed in this manuscript may be found at <https://github.com/ContextLab/effic->  
1040 [ient-learning-khan](https://github.com/ContextLab/efficient-learning-khan).

1041 **Code availability**

1042 All of the code for running our experiment and carrying out the analyses may be found at  
1043 <https://github.com/ContextLab/efficient-learning-khan>.

1044 **Acknowledgements**

1045 We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of  
1046 this study, and assistance with data collection efforts from Will Baxley, Max Bluestone, Daniel  
1047 Carstensen, Kunal Jha, Caroline Lee, Lucy Owen, Xinming Xu, and Kirsten Ziman. Our work was  
1048 supported in part by NSF CAREER Award Number 2145172 to J.R.M. The content is solely the  
1049 responsibility of the authors and does not necessarily represent the official views of our supporting  
1050 organizations. The funders had no role in study design, data collection and analysis, decision to  
1051 publish, or preparation of the manuscript.

## 1052 References

- 1053 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,  
1054 56:149–178.
- 1055 [2] Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015a). Parsimonious mixed models. *arXiv*,  
1056 1506.04967.
- 1057 [3] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015b). Fitting linear mixed-effects models  
1058 using lme4. *Journal of Statistical Software*, 67(1):1–48.
- 1059 [4] Bevilacqua, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and  
1060 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom  
1061 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 1062 [5] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text  
1063 with the natural language toolkit*. Reilly Media, Inc.
- 1064 [6] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-  
1065 dren: distinguishing response flexibility from conceptual flexibility; the protracted development  
1066 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 1067 [7] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International  
1068 Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing  
1069 Machinery.
- 1070 [8] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine  
1071 Learning Research*, 3:993–1022.
- 1072 [9] Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models:  
1073 problems, diagnostics, and improvements. In Airolidi, E. M., Blei, D. M., Erosheva, E. A., and  
1074 Fienberg, S. E., editors, *Handbook of Mixed Membership Models and Their Applications*. CRC Press.

- 1075 [10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,  
1076 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,  
1077 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,  
1078 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,  
1079 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 1080 [11] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the  
1081 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 1082 [12] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-  
1083 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal  
1084 sentence encoder. *arXiv*, 1803.11175.
- 1085 [13] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual  
1086 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 1087 [14] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge  
1088 Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- 1089 [15] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).  
1090 Evidence for a new conceptualization of semantic representation in the left and right cerebral  
1091 hemispheres. *Cortex*, 40(3):467–478.
- 1092 [16] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).  
1093 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,  
1094 41(6):391–407.
- 1095 [17] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 1096 [18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep  
1098 bidirectional transformers for language understanding. *arXiv*, 1810.04805.

- 1099 [19] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,  
1100 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony  
1101 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 1102 [20] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 1103 [21] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of*  
1104 *Experimental Psychology: General*, 115:155–174.
- 1105 [22] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*  
1106 *Transactions of the Royal Society A*, 222(602):309–368.
- 1107 [23] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.  
1108 *School Science and Mathematics*, 100(6):310–318.
- 1109 [24] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather  
1110 prediction” task? individual variability in strategies for probabilistic category learning. *Learning*  
1111 *and Memory*, 9:408–418.
- 1112 [25] Goldman, N. and Whelan, S. (2000). Statistical Tests of Gamma-Distributed Rate Heterogeneity  
1113 in Models of Sequence Evolution in Phylogenetics. *Molecular Biology and Evolution*, 17(6):975–978.
- 1114 [26] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*  
1115 *Cognition and Development*, 13(1):19–37.
- 1116 [27] Halekoh, U. and Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric  
1117 Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *Journal of*  
1118 *Statistical Software*, 59(9):1–32.
- 1119 [28] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual  
1120 learning, pages 212–221. Sage Publications.
- 1121 [29] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-  
1122 ioral and neural signatures of transforming experiences into memories. *Nature Human Behaviour*,  
1123 5:905–919.

- 1124 [30] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge au-  
1125 tomatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*,  
1126 9:doi.org/10.3389/fpsyg.2018.00133.
- 1127 [31] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-  
1128 peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–  
1129 4008.
- 1130 [32] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating  
1131 empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 1132 [33] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.  
1133 Columbia University Press.
- 1134 [34] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,  
1135 326(7382):213–216.
- 1136 [35] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).  
1137 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International  
1138 Journal of Environmental Research and Public Health*, 18(5):2672.
- 1139 [36] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 1140 [37] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 1141 [38] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.  
1142 *The Chronicle of Higher Education*, 21:1–5.
- 1143 [39] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic  
1144 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,  
1145 104:211–240.
- 1146 [40] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic  
1147 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.

- 1148 [41] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of*  
1149 *Educational Studies*, 53(2):129–147.
- 1150 [42] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”  
1151 function? *Psychological Review*, 128(4):711–725.
- 1152 [43] Manning, J. R. (2023). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,  
1153 *Handbook of Human Memory*. Oxford University Press.
- 1154 [44] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension  
1155 for adding LLM-driven chatbots to interactive notebooks. [https://github.com/ContextLab/  
1156 chatify](https://github.com/ContextLab/chatify).
- 1157 [45] Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type i error  
1158 and power in linear mixed models. *Journal of Memory and Language*, 94:305–315.
- 1159 [46] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and  
1160 projection for dimension reduction. *arXiv*, 1802(03426).
- 1161 [47] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold  
1162 Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 1163 [48] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of  
1164 mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 1165 [49] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,  
1166 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to  
1167 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 1168 [50] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-  
1169 tations in vector space. *arXiv*, 1301.3781.
- 1170 [51] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications  
1171 from a national survey of language educators. *System*, 97:102431.

- 1172 [52] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of  
1173 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1174 [53] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).  
1175 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*  
1176 *Neuroscience*, 17(4):367–376.
- 1177 [54] North, B. V., Curtis, D., and Sham, P. C. (2002). A note on the calculation of empirical p values  
1178 from monte carlo procedures. *American Journal of Human Genetics*, 71(2):439–441.
- 1179 [55] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 1180 [56] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models.  
1181 *arXiv*, 2208.02957.
- 1182 [57] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG  
1183 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,  
1184 7:43916.
- 1185 [58] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*  
1186 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 1187 [59] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.  
1188 *Biological Cybernetics*, 45(1):35–41.
- 1189 [60] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in  
1190 higher education: unmasking power and raising questions about the movement’s democratic  
1191 potential. *Educational Theory*, 63(1):87–110.
- 1192 [61] Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for a zero random  
1193 effect variance or polynomial regression in additive and linear mixed models. *Computational*  
1194 *Statistics & Data Analysis*, 52(7):3283–3299.
- 1195 [62] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter  
1196 Student conceptions and conceptual learning in science. Routledge.

- 1197 [63] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-  
1198 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*  
1199 *tion in Nursing*, 22:32–42.
- 1200 [64] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching  
1201 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 1202 [65] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for  
1203 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*  
1204 *Mathematics Education*, 35(5):305–329.
- 1205 [66] Snijders, T. A. B. and Bosker, R. (2011). More powerful tests for variance parameters. In  
1206 *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, chapter 6, pages  
1207 94–108. Sage Publications, 2nd edition.
- 1208 [67] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal*  
1209 *Medicine*, 21:524–530.
- 1210 [68] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,  
1211 Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023).  
1212 LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 1213 [69] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Pio-  
1214 ntkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust  
1215 detection of AI-generated texts. *arXiv*, 2306.04723.
- 1216 [70] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold?  
1217 distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of*  
1218 *the Cognitive Science Society*, 43(43).
- 1219 [71] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and  
1220 Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing*  
1221 *Systems*.

- 1222 [72] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned  
1223 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 1224 [73] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from  
1225 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 1226 [74] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is  
1227 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*  
1228 *Research Methods*, 50:2597–2605.