

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

<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, 31,  
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 [41, 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 [38, 42, 62, 69].

50 The difference between “understanding” and “memorizing,” as framed by researchers in ed-  
51 ucation, cognitive psychology, and cognitive neuroscience [e.g., 24, 29, 34, 42, 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, 40, 51, 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 [43, 44]. To be clear, this is not  
61 to say that text embedding models themselves are capable of “understanding” deep conceptual  
62 meaning in any traditional sense. But rather, their ability to capture the underlying *structure* of  
63 text documents beyond their surface-level contents provides a computational framework through  
64 which those documents’ deeper conceptual meanings may be quantified, explored, and under-  
65 stood. According to these models, the deep conceptual meaning of a document may be captured  
66 by a feature vector in a high-dimensional representation space, wherein nearby vectors reflect con-  
67 ceptually related documents. A model that succeeds at capturing an analogue of “understanding”  
68 is able to assign nearby feature vectors to two conceptually related documents, *even when the specific*  
69 *words contained in those documents have limited overlap*. In this way, “concepts” are defined implicitly  
70 by the model’s geometry [e.g., how the embedding coordinate of a given word or document relates  
71 to the coordinates of other text embeddings; 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.

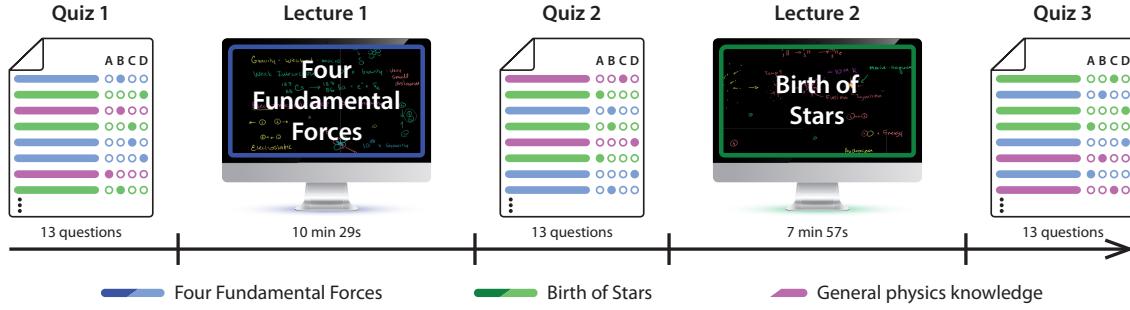
83 Here we develop a framework for modeling how conceptual knowledge is acquired during  
84 learning. The central idea behind our framework is to use text embedding models to define the  
85 coordinate systems of two maps: a *knowledge map* that describes the extent to which each concept is  
86 currently 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, 25, 32, 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 [29, 42]. 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 [37] 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



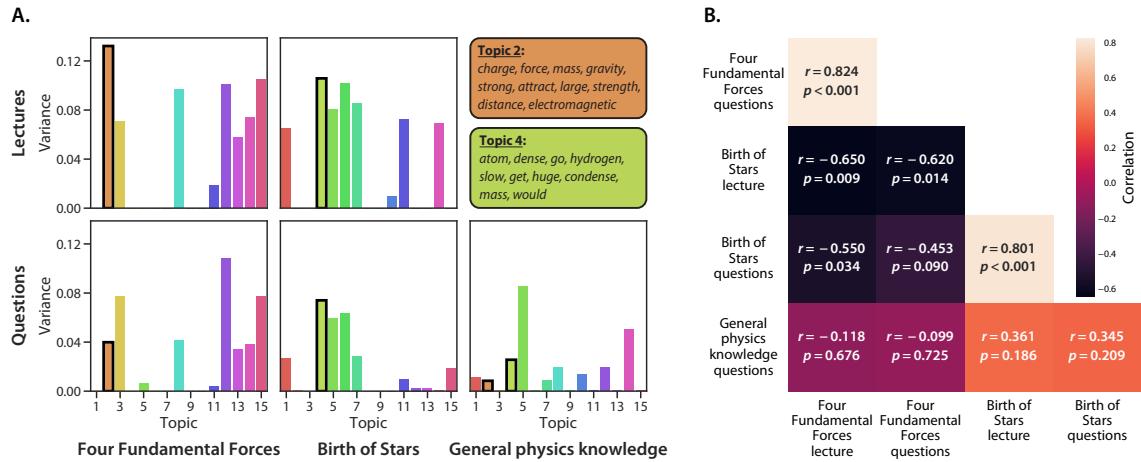
**Figure 2: Modeling course content.** **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The full set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text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

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

To study in detail how participants’ conceptual knowledge changed over the course of the experiment, we first sought to model the conceptual content presented to them at each moment throughout each of the two lectures. We adapted an approach we developed in prior work [30] to identify the latent themes in the lectures using a topic model [8]. Briefly, topic models take as input a collection of text documents, and learn a set of “topics” (i.e., latent themes) from their contents. Once fit, a topic model can be used to transform arbitrary (potentially new) documents into sets of “topic proportions,” describing the weighted blend of learned topics reflected in their texts. We parsed automatically generated transcripts of the two lectures into overlapping sliding windows, where each window contained the text of the lecture transcript from a particular time

span. We treated the set of text snippets (across all of these windows) as documents to fit the model (Fig. 2A; see *Constructing text embeddings of multiple lectures and questions*). Transforming the text from every sliding window with the model yielded a number-of-windows by number-of-topics (15) topic-proportions matrix describing the unique mixture of broad themes from both lectures reflected in each window’s text. Each window’s “topic vector” (i.e., column of the topic-proportions matrix) is analogous to a coordinate in a 15-dimensional space whose axes are topics discovered by the model. Within this space, each lecture’s sequence of topic vectors (i.e., corresponding to its transcript’s overlapping text snippets across sliding windows) forms a *trajectory* that captures how its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution of one topic vector for each second of video (i.e., 1 Hz).

We hypothesized that a topic model trained on transcripts of the two lectures should also capture the conceptual knowledge probed by each quiz question. If indeed the topic model could capture information about the deeper conceptual content of the lectures (i.e., beyond surface-level details such as particular word choices), then we should be able to recover a correspondence between each lecture and questions *about* each lecture. Importantly, such a correspondence could not solely arise from superficial text matching between lecture transcripts and questions, since the lectures and questions often used different words (Supp. Fig. 5) and phrasings. Simply comparing the average topic weights from each lecture and question set (averaging across time and questions, respectively) reveals a striking correspondence (Supp. Fig. 2). Specifically, the average topic weights from Lecture 1 are strongly correlated with the average topic weights from Lecture 1 questions ( $r(13) = 0.809$ ,  $p < 0.001$ , 95% confidence interval (CI) = [0.633, 0.962]), and the average topic weights from Lecture 2 are strongly correlated with the average topic weights from Lecture 2 questions ( $r(13) = 0.728$ ,  $p = 0.002$ , 95% CI = [0.456, 0.920]). At the same time, the average topic weights from the two lectures are *negatively* correlated with the average topic weights from their non-matching question sets (Lecture 1 video vs. Lecture 2 questions:  $r(13) = -0.547$ ,  $p = 0.035$ , 95% CI = [-0.812, -0.231]; Lecture 2 video vs. Lecture 1 questions:  $r(13) = -0.612$ ,  $p = 0.015$ , 95% CI = [-0.874, -0.281]), indicating that the topic model also exhibits some degree of specificity. The full set of pairwise comparisons between average topic weights for the lectures and question sets



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

188 is reported in Supplementary Figure 2.

189 Another, more sensitive, way of summarizing the conceptual content of the lectures and questions is to look at *variability* in how topics are weighted over time and across different questions 190 (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “information” [23] the lecture (or question set) reflects about that topic. For example, suppose a given 191 topic is weighted on heavily throughout a lecture. That topic might be characteristic of some 192 aspect or property of the lecture *overall* (conceptual or otherwise), but unless the topic’s weights 193 changed in meaningful ways over time, the topic would be a poor indicator of any *specific* conceptual 194 content in the lecture. We therefore also compared the variances in topic weights (across time 195 or questions) between the lectures and questions. The variability in topic expression (over time 196 or questions) was similar for the Lecture 1 video and questions ( $r(13) = 0.824, p < 0.001$ , 197 95% CI = [0.696, 0.973]) and the Lecture 2 video and questions ( $r(13) = 0.801, p < 0.001$ , 95% 198 CI = [0.539, 0.958]). Simultaneously, as reported in Figure 3B, the variabilities in topic expression 199 across *different* videos and lecture-specific questions (i.e., Lecture 1 video vs. Lecture 2 questions; 200 201

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

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

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

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

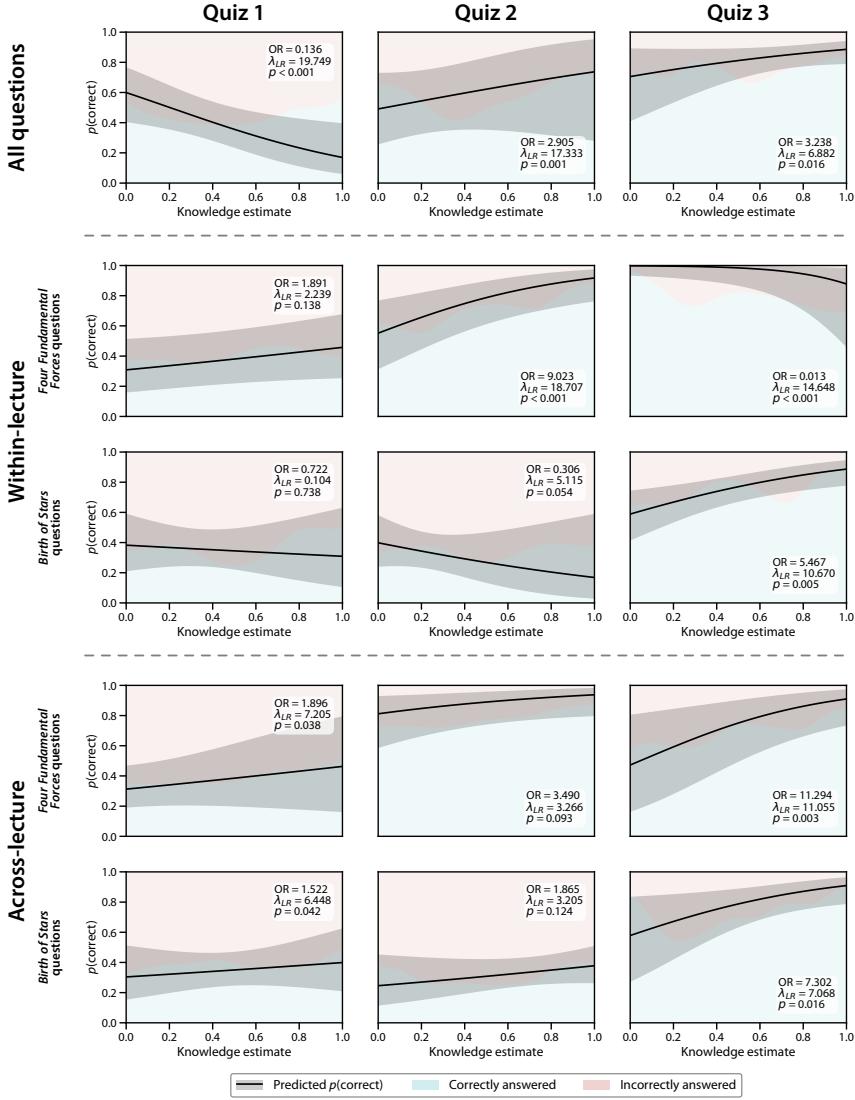


**Figure 5: Estimating knowledge about the content presented at each moment of each lecture.** **A. Knowledge about the time-varying content of *Four Fundamental Forces*.** Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (see *Estimating dynamic knowledge traces*), using responses from a single quiz (color). The traces are averaged across participants. **B. Average estimated knowledge about *Four Fundamental Forces*.** Each bar displays the across-timepoint average knowledge, estimated using the responses to one quiz's questions. **C. Knowledge about the time-varying content of *Birth of Stars*.** The panel is in the same format as Panel A, but here the knowledge estimates are for the moment-by-moment content of the *Birth of Stars* lecture. **D. Average estimated knowledge about *Birth of Stars*.** The panel is in the same format as Panel B, but here the knowledge estimates are for the content of the *Birth of Stars* lecture. All panels: error ribbons and error bars denote 95% confidence intervals, estimated across participants.

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

276 If we are able to accurately estimate a participant’s knowledge about the content tested by a  
277 given question, our estimates of their knowledge should carry some predictive information about  
278 whether they are likely to answer that question correctly or incorrectly. We developed a statistical  
279 approach to test this claim. For each quiz question a participant answered, in turn, we used  
280 Equation 1 to estimate their knowledge at the given question’s embedding space coordinate based  
281 on other questions that participant answered on the same quiz. We repeated this for all participants,  
282 and for each of the three quizzes. Then, separately for each quiz, we fit a generalized linear mixed  
283 model (GLMM) with a logistic link function to explain the likelihood of correctly answering a  
284 question as a function of estimated knowledge for its embedding coordinate, while accounting  
285 for random variation among participants and questions (see *Generalized linear mixed models*). To  
286 assess the predictive value of the knowledge estimates, we compared each GLMM to an analogous  
287 (i.e., nested) “null” model that did not consider estimated knowledge using parametric bootstrap  
288 likelihood-ratio tests.

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



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

299 about the *other* lecture (“Across-lecture”; Fig. 6, bottom rows). This test was intended to assess the  
300 *generalizability* of our approach by asking whether our predictions held across the content areas of  
301 the two lectures.

302 In performing these analyses, our null hypothesis is that the knowledge estimates we compute  
303 based on the quiz questions’ embedding coordinates do *not* provide useful information about  
304 participants’ abilities to answer those questions. What result might we expect to see if this is the  
305 case? To gain an intuition for this scenario, consider the expected outcome if we carried out these  
306 same analyses using a simple proportion-correct measure in lieu of our knowledge estimates.  
307 Suppose a participant correctly answered  $n$  out of  $q$  questions on a given quiz. If we hold out  
308 a single *correctly* answered question, the proportion of remaining questions answered correctly  
309 would be  $\frac{n-1}{q-1}$ . Whereas if we hold out a single *incorrectly* answered question, the proportion of  
310 remaining questions answered correctly would be  $\frac{n}{q-1}$ . In this way, the proportion of correctly  
311 answered remaining questions is always *lower* when the held-out question was answered correctly  
312 than when it was answered incorrectly. Because our knowledge estimates are computed as a  
313 weighted version of this same proportion-correct score (where each held-in question’s weight  
314 reflects its embedding-space distance from the held-out question; see Eqn. 1), if these weights  
315 are uninformative (e.g., randomly distributed), then we should expect to see this same inverse  
316 relationship between estimated knowledge and performance, on average. On the other hand,  
317 if the spatial relationships among the quiz questions’ embeddings *are* predictive of participants’  
318 knowledge about the questions’ content, then we would expect *higher* estimated knowledge for  
319 held-out correctly (versus incorrectly) answered questions.

320 Before presenting our results, it is worth considering three possible explanations of why a  
321 participant might answer a given question correctly or incorrectly. One possibility is that the  
322 participant simply *guessed* the answer. A second is that they selected the incorrect answer by  
323 mistake, despite “knowing” the correct answer (or vice versa). In both of these scenarios, the  
324 participant’s knowledge about the question’s content should be uninformative about their observed  
325 response. A third possibility is that the participant’s response reflects their *actual* knowledge about  
326 the question’s content. In this case, we *might* expect to see a positive relationship between the

327 participant's knowledge and their likelihood of answering the question correctly. However, in  
328 order to see this positive relationship, the participant's knowledge must be structured in a way  
329 that is reflected (at least partially) by the embedding space. In other words, if the participant's  
330 performance reflects their true knowledge, but our text embedding space does not sufficiently  
331 capture the structure of that knowledge, then the knowledge estimates we generate will not be  
332 predictive of the participant's performance. In the extreme, if the embedding space is completely  
333 unstructured with respect to the content of the quiz questions, then we would expect to see the  
334 negative relationship between estimated knowledge and performance that we described above.

335 When we fit a GLMM to estimates of participants' knowledge for each Quiz 1 question based  
336 on all other Quiz 1 questions, we observed an outcome consistent with our null hypothesis: higher  
337 estimated knowledge at the embedding coordinate of a held-out question was associated with a  
338 lower likelihood of answering the question correctly (odds ratio ( $OR$ ) = 0.136, likelihood-ratio test  
339 statistic ( $\lambda_{LR}$ ) = 19.749, 95% CI = [14.352, 26.545],  $p < 0.001$ ). This outcome suggests that our  
340 knowledge estimates do *not* provide useful information about participants' Quiz 1 performance  
341 when we aggregated across all question content areas. We speculated that this might either  
342 indicate that the knowledge estimates are uninformative in general, or about Quiz 1 performance  
343 in particular. This would be expected, for example, if participants were guessing about the answers  
344 to the Quiz 1 questions (prior to having watched either lecture). When we repeated this analysis  
345 for Quizzes 2 and 3, we found that higher estimated knowledge for a given question predicted  
346 a *greater* likelihood of answering it correctly (Quiz 2:  $OR$  = 2.905,  $\lambda_{LR}$  = 17.333, 95% CI =  
347 [14.966, 29.309],  $p = 0.001$ ; Quiz 3:  $OR$  = 3.238,  $\lambda_{LR}$  = 6.882, 95% CI = [6.228, 8.184],  $p = 0.016$ ).  
348 Taken together, these results suggest that our knowledge estimates reliably predict participants'  
349 performance on individual held-out quiz questions, but only after participants have received at  
350 least some training.

351 We observed a similar pattern of results when used this approach to estimate participants'  
352 knowledge about held-out questions from one lecture using their performance on other questions  
353 from the *same* lecture. Specifically, for Quiz 1 questions (i.e., prior to watching either), participants'  
354 estimated knowledge for the embedding coordinates of held-out *Four Fundamental Forces*-related

355 questions estimated using other *Four Fundamental Forces*-related questions did not reliably pre-  
356 dict whether those questions were answered correctly ( $OR = 1.891$ ,  $\lambda_{LR} = 2.293$ , 95% CI =  
357 [2.091, 2.622],  $p = 0.138$ ). The same was true of knowledge estimates for held-out *Birth of Stars*-  
358 related questions based on other *Birth of Stars*-related questions ( $OR = 0.722$ ,  $\lambda_{LR} = 5.115$ , 95% CI =  
359 [0.094, 0.146],  $p = 0.738$ ). As in our analysis that included all questions, we speculate that  
360 these “null” results might reflect some degree of random guessing on Quiz 1. When we re-  
361 peated these within-lecture analyses using questions from Quiz 2 (which participants took im-  
362 mediately after viewing *Four Fundamental Forces* but prior to viewing *Birth of Stars*), we found  
363 that they now reliably predicted success on *Four Fundamental Forces*-related questions ( $OR =$   
364  $9.023$ ,  $\lambda_{LR} = 18.707$ , 95% CI = [10.877, 22.222],  $p < 0.001$ ) but not on *Birth of Stars*-related  
365 questions ( $OR = 0.306$ ,  $\lambda_{LR} = 5.115$ , 95% CI = [4.624, 5.655],  $p = 0.054$ ). Here, we speculate  
366 that participants might have been guessing about the *Birth of Stars* content (e.g., prior to having  
367 watched it), whereas they might have been drawing on some structured knowledge about the *Four*  
368 *Fundamental Forces* content (e.g., from having just watched it). When we applied this approach to  
369 Quiz 3 responses (given immediately after viewing *Birth of Stars*), we found that within-lecture  
370 knowledge estimates for *Birth of Stars*-related questions could now reliably predict success on those  
371 questions ( $OR = 5.467$ ,  $\lambda_{LR} = 10.670$ , 95% CI = [7.998, 12.532],  $p = 0.005$ ). However, within-lecture  
372 knowledge estimates for *Four Fundamental Forces* questions answered on Quiz 3 were no longer  
373 directly related to the likelihood of successfully answering them and instead exhibited the inverse  
374 relationship we would expect to arise from unstructured knowledge (with respect to the embed-  
375 ding space;  $OR = 0.013$ ,  $\lambda_{LR} = 14.648$ , 95% CI = [10.695, 23.096],  $p < 0.001$ ). Speculatively, we  
376 suggest that this may reflect participants forgetting some of the *Four Fundamental Forces* content  
377 (e.g., perhaps in favor of prioritizing encoding the just-watched *Birth of Stars* content in preparation  
378 for the third quiz). If this forgetting happens in a relatively “random” way (with respect to spatial  
379 distance within the embedding space), then it could explain why some held-out questions about  
380 *Four Fundamental Forces* were answered incorrectly, even if questions at nearby coordinates (i.e.,  
381 about similar content) were answered correctly. This might lead our approach to over-estimate  
382 knowledge for held-out questions about “forgotten” knowledge that participants answered in-

383 correctly. Taken together, these within-lecture results suggest that our approach can distinguish  
384 between questions about different content covered by a single lecture when participants have suf-  
385 ficiently structured knowledge about its contents, though this specificity may decrease with time  
386 since the relevant material was learned.

387 Finally, we used this approach to estimate participants' knowledge about held-out questions  
388 from one lecture using their performance on questions from the *other* lecture. Here we again  
389 observed a similar pattern of results, though with some notable differences. On Quiz 1, we found  
390 that participants' abilities to correctly answer questions about *Four Fundamental Forces* could be  
391 predicted from their responses to questions about *Birth of Stars* ( $OR = 1.896, \lambda_{LR} = 7.205, 95\% CI =$   
392  $[6.224, 7.524], p = 0.038$ ) and similarly, that their ability to correctly answer *Birth of Stars*-related  
393 questions could be predicted from their responses to *Four Fundamental Forces*-related questions  
394 ( $OR = 1.522, \lambda_{LR} = 6.448, 95\% CI = [5.656, 6.843], p = 0.042$ ). Given the results from our analyses  
395 that included all questions and within-lecture predictions, we were surprised to find that the  
396 knowledge estimates could reliably (if weakly) predict participants' performance across content  
397 from different lectures. It is possible that this result reflects a combination of random guessing  
398 prior to training (leading to a weak effect size), alongside some coarse-scale structured knowledge  
399 that participants had about the content prior to watching either lecture. When we repeated  
400 this analysis using questions from Quiz 2, we found participants' responses to *Four Fundamental*  
401 *Forces*-related questions did *not* reliably predict their success on *Birth of Stars*-related questions  
402 ( $OR = 1.865, \lambda_{LR} = 3.205, 95\% CI = [3.027, 3.600], p = 0.124$ ), nor did their responses to *Birth of*  
403 *Stars*-related questions reliably predict their success on *Four Fundamental Forces*-related questions  
404 ( $OR = 3.490, \lambda_{LR} = 3.266, 95\% CI = [3.033, 3.866], p = 0.093$ ). These "prediction failures" appear  
405 to come from the fact that any signal derived from participants' knowledge about the content of  
406 the *Birth of Stars* lecture (prior to watching it) is overwhelmed by the much more dramatic increase  
407 in their knowledge about the content of the *Four Fundamental Forces* (which they watched just  
408 prior to taking Quiz 2). This is reflected in their Quiz 2 performance for questions about each  
409 lecture (mean proportion correct for *Four Fundamental Forces*-related questions on Quiz 2: 0.77;  
410 mean proportion correct for *Birth of Stars*-related questions on Quiz 2: 0.36). When we carried out

411 these across-lecture knowledge predictions using questions from Quiz 3 (when participants had  
412 now viewed *both* lectures), we could again reliably predict success on questions about both *Four*  
413 *Fundamental Forces* ( $OR = 11.294$ ,  $\lambda_{LR} = 11.055$ , 95% CI = [9.126, 18.476],  $p = 0.003$ ) and *Birth of*  
414 *Stars* ( $OR = 7.302$ ,  $\lambda_{LR} = 7.068$ , 95% CI = [6.490, 8.584],  $p = 0.016$ ) using responses to questions  
415 about the other lecture’s content. Across all three versions of these analyses, our results suggest  
416 that (by and large) our knowledge estimates can reliably predict participants’ abilities to answer  
417 individual quiz questions, distinguish between questions about similar content, and generalize  
418 across content areas, provided that participants’ quiz responses reflect a minimum level of “real”  
419 knowledge about both content on which these predictions are based and that for which they are  
420 made.

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

439 measure that does not consider embedding space distance.

440 With these ideas in mind, we asked: conditioned on answering a question correctly, what  
 441 proportion of all questions (within some radius,  $r$ , of that question’s embedding coordinate)  
 442 were answered correctly? We plotted this proportion as a function of  $r$ . Similarly, we could  
 443 ask, conditioned on answering a question incorrectly, how the proportion of correct responses  
 444 changed with  $r$ . As shown in Figure 7, we found that quiz performance falls off smoothly with  
 445 distance, and the “rate” of the falloff does not appear to change across the different quizzes, as  
 446 measured by the distance at which performance becomes statistically indistinguishable from a  
 447 simple proportion correct score (see *Estimating the “smoothness” of knowledge*). This suggests that,

448 at least within the region of text embedding space covered by the questions our participants  
449 answered (and as characterized using our topic model), the rate at which knowledge changes  
450 with distance is relatively constant, even as participants' overall level of knowledge varies across  
451 quizzes or regions of the embedding space.

452 Knowledge estimates need not be limited to the content of the lectures. As illustrated in  
453 Figure 8, our general approach to estimating knowledge from a small number of quiz questions  
454 may be extended to *any* content, given its text embedding coordinate. To visualize how knowledge  
455 “spreads” through text embedding space to content beyond the lectures participants watched, we  
456 first fit a new topic model to the lectures’ sliding windows with  $k = 100$  topics. Conceptually,  
457 increasing the number of topics used by the model functions to increase the “resolution” of the  
458 embedding space, providing a greater ability to estimate knowledge for content that is highly  
459 similar to (but not precisely the same as) that contained in the two lectures. We note that we  
460 used these 2D maps solely for visualization; all relevant comparisons, distance computations, and  
461 statistical tests we report above were carried out in the original 15-dimensional space, using the  
462 15-topic model. Aside from increasing the number of topics from 15 to 100, all other procedures  
463 and model parameters were carried over from the preceding analyses. As in our other analyses,  
464 we resampled each lecture’s topic trajectory to 1 Hz and projected each question into a shared text  
465 embedding space.

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

475 Several features of the resulting knowledge maps are worth noting. The average knowledge



**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.

476 map estimated from Quiz 1 responses (Fig. 8A, leftmost map) shows that participants tended to  
477 have relatively little knowledge about any parts of the text embedding space (i.e., the shading is  
478 relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked  
479 increase in knowledge on the left side of the map (around roughly the same range of coordinates  
480 traversed by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words,  
481 participants' estimated increase in knowledge is localized to conceptual content that is nearby (i.e.,  
482 related to) the content from the lecture they watched prior to taking Quiz 2. This localization is  
483 non-trivial: these knowledge estimates are informed only by the embedded coordinates of the  
484 *quiz questions*, not by the embeddings of either lecture (see Eqn. 4). Finally, the knowledge map  
485 estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region  
486 surrounding the embedding of the *Birth of Stars* lecture participants watched immediately prior to  
487 taking Quiz 3.

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

494 Because the 2D projection we used to construct the knowledge and learning maps is invertible,  
495 we may gain additional insights into these maps' meanings by reconstructing the original high-  
496 dimensional topic vector for any location on the map we are interested in. For example, this could  
497 serve as a useful tool for an instructor looking to better understand which content areas a student  
498 (or a group of students) knows well (or poorly). As a demonstration, we show the top-weighted  
499 words from the blends of topics reconstructed from three example locations on the maps (Fig. 8C):  
500 one point near the *Four Fundamental Forces* embedding (yellow), a second point near the *Birth of*  
501 *Stars* embedding (orange), and a third point between the two lectures' embeddings (pink). As  
502 shown in the word clouds in the panel, the top-weighted words at the example coordinate near the  
503 *Four Fundamental Forces* embedding tended to be weighted more heavily by the topics expressed

504 in that lecture. Similarly, the top-weighted words at the example coordinate near the *Birth of Stars*  
505 embedding tended to be weighted more heavily by the topics expressed in *that* lecture. And the  
506 top-weighted words at the example coordinate between the two lectures' embeddings show a  
507 roughly even mix of words most strongly associated with each lecture.

## 508 Discussion

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

520 We view our work as making several contributions to the study of how people acquire con-  
521 ceptual knowledge. First, from a methodological standpoint, our modeling framework provides  
522 a systematic means of mapping out and characterizing knowledge in maps that have infinite (ar-  
523 bitrarily many) numbers of coordinates, and of “filling out” those maps using relatively small  
524 numbers of multiple choice quiz questions. Our experimental finding that we can use these maps  
525 to predict responses to held-out questions has several psychological implications as well. For ex-  
526 ample, concepts that are assigned to nearby coordinates by the text embedding model also appear  
527 to be “known to a similar extent” (as reflected by participants’ responses to held-out questions;  
528 Fig. 6). This suggests that participants also *conceptualize* similarly the content reflected by nearby  
529 embedding coordinates. How participants’ knowledge falls off with spatial distance is captured

530 by the knowledge maps we infer from their quiz responses (e.g., Figs. 7, 8). In other words, our  
531 study shows that knowledge about a given concept implies knowledge about related concepts,  
532 and we also show how estimated knowledge falls off with distance in text embedding space.

533 In our study, we characterize the “coordinates” of participants’ knowledge using a relatively  
534 simple “bag of words” text embedding model [LDA; 8]. More sophisticated text embedding mod-  
535 els, such as transformer-based models [18, 55, 68, 71] can learn complex grammatical and semantic  
536 relationships between words, higher-order syntactic structures, stylistic features, and more. We  
537 considered using transformer-based models in our study, but we found that the text embeddings  
538 derived from these models were surprisingly uninformative with respect to differentiating or oth-  
539 erwise characterizing the conceptual content of the lectures and questions we used. We suspect  
540 that this reflects a broader challenge in constructing models that are high-resolution within a given  
541 domain (e.g., the domain of physics lectures and questions) *and* sufficiently broad so as to enable  
542 them to cover a wide range of domains. For example, we found that the embeddings derived even  
543 from much larger and more modern models like BERT [18], GPT [71], LLaMa [68], and others that  
544 are trained on enormous text corpora, end up yielding poor resolution within the content space  
545 spanned by individual course videos (Supp. Fig. 6). Whereas the LDA embeddings of the lectures  
546 and questions are “near” each other (i.e., the convex hull enclosing the two lectures’ trajectories is  
547 highly overlapping with the convex hull enclosing the questions’ embeddings), the BERT embed-  
548 dings of the lectures and questions are instead largely distinct (top row of Supp. Fig. 6). The LDA  
549 embeddings of the questions for each lecture and the corresponding lecture’s trajectory are also  
550 similar. For example, as shown in Fig. 2C, the LDA embeddings for *Four Fundamental Forces* ques-  
551 tions (blue dots) appear closer to the *Four Fundamental Forces* lecture trajectory (blue line), whereas  
552 the LDA embeddings for *Birth of Stars* questions (green dots) appear closer to the *Birth of Stars*  
553 lecture trajectory (green line). The BERT embeddings of the lectures and questions do not show  
554 this property (Supp. Fig. 6). We also examined per-question “content matches” between individual  
555 questions and individual moments of each lecture (Figs. 4, 6). The time series plot of individual  
556 questions’ correlations are different from each other when computed using LDA (e.g., the traces  
557 can be clearly visually separated), whereas the correlations computed from BERT embeddings of

558 different questions all look very similar. This tells us that LDA is capturing some differences in  
559 content between the questions, whereas BERT is not. The time series plots of individual ques-  
560 tions’ correlations have clear “peaks” when computed using LDA, but not when computed using  
561 BERT. This tells us that LDA is capturing a “match” between the content of each question and a  
562 relatively well-defined time window of the corresponding lectures. The BERT embeddings appear  
563 to blur together the content of the questions versus specific moments of each lecture. Finally, we  
564 also compared the pairwise correlations between embeddings of questions within versus across  
565 content areas (i.e., content covered by the individual lectures, lecture-specific questions, and by the  
566 “general physics knowledge” questions). The LDA embeddings show a strong contrast between  
567 same-content embeddings versus across-content embeddings. In other words, the embeddings of  
568 questions about the *Four Fundamental Forces* material are highly correlated with the embeddings of  
569 the *Four Fundamental Forces* lecture, but not with the embeddings of *Birth of Stars*, questions about  
570 *Birth of Stars*, or general physics knowledge questions. We see a similar pattern with the LDA  
571 embeddings of the *Birth of Stars* questions (Fig. 3, Supp. Fig. 2). In contrast, the BERT embeddings  
572 are all highly correlated with each other (Supp. Fig. 6). Taken together, these comparisons illus-  
573 trate how LDA (trained on the specific content in question) provides both coverage of the requisite  
574 material and specificity at the level of the content covered by individual questions. BERT, on the  
575 other hand, essentially assigns both lectures and all of the questions (which are all broadly about  
576 “physics”) into a tiny region of its embedding space, thereby blurring out meaningful distinctions  
577 between different specific concepts covered by the lectures and questions. We note that these are  
578 not criticisms of BERT (or other large language models trained on large and diverse corpora).  
579 Rather, our point is that simple fine-tuned models trained on a relatively small but specialized  
580 corpus can outperform much more complicated models trained on much larger corpora, when we  
581 are specifically interested in capturing subtle conceptual differences at the level of a single course  
582 lecture or question. Of course if our goal had been to find a model that generalized to many  
583 different content areas, we would expect our approach to perform comparatively poorly relative to  
584 BERT or other much larger models. We suggest that bridging the tradeoff between high resolution  
585 within each content area versus the ability to generalize to many different content areas will be an

586 important challenge for future work in this domain.

587 Another application for large language models that does *not* require explicitly modeling the  
588 content of individual lectures or questions is to leverage the models' abilities to generate text. For  
589 example, generative text models like ChatGPT [55] and LLaMa [68] are already being used to build  
590 a new generation of interactive tutoring systems [e.g., 45]. Unlike the approach we have taken here,  
591 these generative text model-based systems do not explicitly model what learners know, or how  
592 their knowledge changes over time with training. One could imagine building a hybrid system  
593 that combines the best of both worlds: a large language model that can *generate* text, combined  
594 with a smaller model that can *infer* what learners know and how their knowledge changes over  
595 time. Such a hybrid system could potentially be used to build the next generation of interactive  
596 tutoring systems that are able to adapt to learners' needs in real time, and that are able to provide  
597 more nuanced feedback about what learners know and what they do not know.

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

614 on the other hand, are strong when *any* of these words are mentioned, even if they do not occur  
615 together.

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

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

628 Within the past several years, the global pandemic forced many educators to suddenly adapt to  
629 teaching remotely [36, 52, 64, 72]. This change in world circumstances is happening alongside (and  
630 perhaps accelerating) geometric growth in the availability of high-quality online courses from plat-  
631 forms such as Khan Academy [37], Coursera [73], EdX [39], and others [60]. Continued expansion  
632 of the global internet backbone and improvements in computing hardware have also facilitated  
633 improvements in video streaming, enabling videos to be easily shared and viewed by increasingly  
634 large segments of the world’s population. This exciting time for online course instruction provides  
635 an opportunity to re-evaluate how we, as a global community, educate ourselves and each other.  
636 For example, we can ask: what defines an effective course or training program? Which aspects of  
637 teaching might be optimized and/or augmented by automated tools? How and why do learning  
638 needs and goals vary across people? How might we lower barriers to receiving a high-quality  
639 education?

640 Alongside these questions, there is a growing desire to extend existing theories beyond the  
641 domain of lab testing rooms and into real classrooms [35]. In part, this has led to a recent

642 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better  
643 reflect more ethologically valid phenomena that are more directly relevant to real-world situations  
644 and behaviors [53]. In turn, this has brought new challenges in data analysis and interpretation. A  
645 key step towards solving these challenges will be to build explicit models of real-world scenarios  
646 and how people behave in them (e.g., models of how people learn conceptual content from real-  
647 world courses, as in our current study). A second key step will be to understand which sorts  
648 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 4,  
649 19, 50, 54, 57] might help to inform these models. A third major step will be to develop and  
650 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic  
651 paradigms.

652 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also  
653 relate to the notion of “theory of mind” of other individuals [27, 33, 49]. Considering others’ unique  
654 perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and  
655 communicate [58, 63, 67]. One could imagine future extensions of our work (e.g., analogous to  
656 the knowledge and learning maps shown in Fig. 8), that attempt to characterize how well-aligned  
657 different people’s knowledge bases or backgrounds are. In turn, this might be used to model how  
658 knowledge (or other forms of communicable information) flows not just between teachers and  
659 students, but between friends having a conversation, individuals on a first date, participants at  
660 a business meeting, doctors and patients, experts and non-experts, political allies or adversaries,  
661 and more. For example, the extent to which two people’s knowledge maps “match” or “align” in  
662 a given region of text embedding space might serve as a predictor of how effectively they will be  
663 able to communicate about the corresponding conceptual content.

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

670 **Materials and methods**

671 **Participants**

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

677 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09  
678 years). A total of 15 participants reported their gender as male and 35 participants reported their  
679 gender as female. A total of 49 participants reported their native language as "English" and 1  
680 reported having another native language. A total of 47 participants reported their ethnicity as  
681 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants  
682 reported their races as White (32 participants), Asian (14 participants), Black or African American  
683 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other  
684 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

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

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

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

## 731 Experiment

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

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

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

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

## 762 **Analysis**

### 763 **Statistics**

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

### 767 **Constructing text embeddings of multiple lectures and questions**

768 We adapted an approach we developed in prior work [30] to embed each moment of the two  
769 lectures and each question in our pool in a common representational space. Briefly, our approach  
770 uses a topic model [Latent Dirichlet Allocation; 8] trained on a set of documents, to discover a set  
771 of  $k$  “topics” or “themes.” Formally, each topic is defined as a distribution of weights over words  
772 in the model’s vocabulary (i.e., the union of all unique words, across all documents, excluding  
773 “stop words.”). Conceptually, each topic is intended to give larger weights to words that are  
774 semantically related (as inferred from their tendency to co-occur in the same document). After  
775 fitting a topic model, each document in the training set, or any *new* document that contains at  
776 least some of the words in the model’s vocabulary, may be represented as a  $k$ -dimensional vector

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

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

795 We fetched these automated transcripts using the `youtube-transcript-api` Python pack-  
796 age [17]. The transcripts consisted of one timestamped line of text for every few seconds (mean:  
797 2.34 s; standard deviation: 0.83 s) of spoken content in the video (i.e., corresponding to each indi-  
798 vidual caption that would appear on-screen if viewing the lecture via YouTube, and when those  
799 lines would appear). We defined a sliding window length of (up to)  $w = 30$  transcript lines, and  
800 assigned each window a timestamp corresponding to the midpoint between the timestamps for its  
801 first and last lines. This  $w$  parameter was chosen to match the same number of words per sliding  
802 window (rounded to the nearest whole word, and before preprocessing) as the sliding windows  
803 we defined in our prior work [30; i.e., 185 words per sliding window].

804 These sliding windows ramped up and down in length at the beginning and end of each

805 transcript, respectively. In other words, each transcript’s first sliding window covered only its first  
806 line, the second sliding window covered the first two lines, and so on. This ensured that each line  
807 from the transcripts appeared in the same number ( $w$ ) of sliding windows. We next performed a  
808 series of standard text preprocessing steps: normalizing case, lemmatizing, removing punctuation  
809 and removing stop-words. We constructed our corpus of stop words by augmenting the Natural  
810 Language Toolkit [NLTK; 5] English stop word list with the following additional words, selected  
811 using one of the approaches suggested by Boyd-Graber et al. [9]: “actual,” “actually,” “also,” “bit,”  
812 “could,” “e,” “even,” “first,” “follow,” “following,” “four,” “let,” “like,” “mc,” “really,”, “saw,”  
813 “see,” “seen,” “thing,” and “two.” This yielded sliding windows with an average of 73.8 remaining  
814 words, and lasting for an average of 62.22 seconds. We treated the text from each sliding window  
815 as a single “document,” and combined these documents across the two videos’ windows to create  
816 a single training corpus for the topic model.

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

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

832 **Estimating dynamic knowledge traces**

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

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

835 where

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

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

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

850 **Generalized linear mixed models**

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

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

861 In each version of this analysis (i.e., row in Fig. 6), and separately for each of the three quizzes  
862 (i.e., column in Fig. 6), we then fit a generalized linear mixed model (GLMM) with a logistic link  
863 function to the set of knowledge estimates for all questions that participants answered on the  
864 given quiz. We implemented these models in R using the `lme4` package [3] and fit them following  
865 guidance from Bates et al. [2] and Matuschek et al. [46]. Specifically, we initially fit each model  
866 with the maximal random effects structure afforded by our design, which we identified as:

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

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

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

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

883 where "accuracy", "participant", and "question" are as defined above. As with our full models,  
884 the null models we fit for the "All questions" version of the analysis for each quiz contained an  
885 additional term,  $(1 | \text{lecture})$ , where "lecture" is as defined above. We then compared each  
886 full model to its reduced (null) equivalent using a likelihood-ratio test (LRT). Because the typical  
887 asymptotic  $\chi_d^2$  approximation of the null distribution for the LRT statistic ( $\lambda_{LR}$ ) is anti-conservative  
888 for models that differ in their random slope terms [26, 61, 66], we computed  $p$ -values for these  
889 tests using a parametric bootstrap procedure [14, 28]. For each of 1,000 bootstraps, we used the  
890 fitted null model to simulate a sample of observations of equal size to our original sample. We  
891 then re-fit both the null and full models to this simulated sample and compared them via an LRT.  
892 This yielded a distribution of  $\lambda_{LR}$  statistics we may expect to observe under our null hypothesis.  
893 Following Ewens [22], we computed a corrected  $p$ -value for our observed  $\lambda_{LR}$  as  $\frac{r}{n}$ , where  $r$  is the  
894 number of simulated model comparisons that yielded a  $\lambda_{LR}$  greater than or equal to our observed  
895 value and  $n$  is the number of simulations we ran (1,000).

896 **Estimating the "smoothness" of knowledge**

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

902 For each of 10,000 iterations, we drew a random subsample (with replacement) of 50 partic-

903 ipants from our dataset. Within each iteration, we first computed the 95% confidence interval  
904 (CI) of the across-subsample-participants mean proportion correct on each of the three quizzes,  
905 separately. To compute this interval for each quiz, we repeatedly (1,000 times) subsampled par-  
906 ticipants (with replacement, from the outer subsample for the current iteration) and computed  
907 the mean proportion correct of each of these inner subsamples. We then identified the 2.5<sup>th</sup> and  
908 97.5<sup>th</sup> percentiles of the resulting distributions of 1,000 means. These three intervals (one for each  
909 quiz) served as our thresholds for confidence that the proportion correct within a given distance  
910 from a reference question was reliably different (at the  $p < 0.05$  significance level) from the average  
911 proportion correct across all questions on the given quiz.

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

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

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

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

934 **Creating knowledge and learning map visualizations**

935 An important feature of our approach is that, given a trained text embedding model and partic-  
936 ipants' quiz performance on each question, we can estimate their knowledge about *any* content  
937 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-  
938 tions, or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 7, 8, 9, 10,  
939 and 11), we used Uniform Manifold Approximation and Projection [UMAP; 47, 48] to construct a  
940 2D projection of the text embedding space. Whereas our main analyses used a 15-topic embedding  
941 space, we used a 100-topic embedding space for these visualizations. This change in the number  
942 of topics overcame an undesirable behavior in the UMAP embedding procedure, whereby embed-  
943 ding coordinates for the 15-topic model tended to be “clumped” into separated clusters, rather  
944 than forming a smooth trajectory through the 2D space. When we increased the number of topics  
945 to 100, the embedding coordinates in the 2D space formed a smooth trajectory through the space,  
946 with substantially less clumping (Fig. 8). Creating a “map” by sampling this 100-dimensional  
947 space at high resolution to obtain an adequate set of topic vectors spanning the embedding space  
948 would be computationally intractable. However, sampling a 2D grid is trivial.

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

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

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

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

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

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

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

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

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

982 **Author contributions**

983 Conceptualization: P.C.F., A.C.H., and J.R.M. Methodology: P.C.F., A.C.H., and J.R.M. Software:  
984 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:  
985 P.C.F. Writing (original draft): J.R.M. Writing (review and editing): P.C.F., A.C.H., and J.R.M. Visu-  
986 alization: P.C.F. and J.R.M. Supervision: J.R.M. Project administration: P.C.F. Funding acquisition:  
987 J.R.M.

988 **Data availability**

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

991 **Code availability**

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

994 **Acknowledgements**

995 We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of  
996 this study, and assistance with data collection efforts from Will Baxley, Max Bluestone, Daniel  
997 Carstensen, Kunal Jha, Caroline Lee, Lucy Owen, Xinming Xu, and Kirsten Ziman. Our work was  
998 supported in part by NSF CAREER Award Number 2145172 to J.R.M. The content is solely the  
999 responsibility of the authors and does not necessarily represent the official views of our supporting  
1000 organizations. The funders had no role in study design, data collection and analysis, decision to  
1001 publish, or preparation of the manuscript.

## 1002 References

- 1003 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,  
1004 56:149–178.
- 1005 [2] Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015a). Parsimonious mixed models. *arXiv*,  
1006 1506.04967.
- 1007 [3] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015b). Fitting linear mixed-effects models  
1008 using lme4. *Journal of Statistical Software*, 67(1):1–48.
- 1009 [4] Bevilacqua, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and  
1010 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom  
1011 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 1012 [5] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text  
1013 with the natural language toolkit*. Reilly Media, Inc.
- 1014 [6] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-  
1015 dren: distinguishing response flexibility from conceptual flexibility; the protracted development  
1016 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 1017 [7] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International  
1018 Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing  
1019 Machinery.
- 1020 [8] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine  
1021 Learning Research*, 3:993–1022.
- 1022 [9] Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models:  
1023 problems, diagnostics, and improvements. In Airolidi, E. M., Blei, D. M., Erosheva, E. A., and  
1024 Fienberg, S. E., editors, *Handbook of Mixed Membership Models and Their Applications*. CRC Press.

- 1025 [10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,  
1026 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,  
1027 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,  
1028 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,  
1029 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 1030 [11] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the  
1031 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 1032 [12] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-  
1033 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal  
1034 sentence encoder. *arXiv*, 1803.11175.
- 1035 [13] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual  
1036 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 1037 [14] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge  
1038 Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- 1039 [15] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).  
1040 Evidence for a new conceptualization of semantic representation in the left and right cerebral  
1041 hemispheres. *Cortex*, 40(3):467–478.
- 1042 [16] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).  
1043 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,  
1044 41(6):391–407.
- 1045 [17] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 1046 [18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep  
1047 bidirectional transformers for language understanding. *arXiv*, 1810.04805.

- 1049 [19] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,  
1050 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony  
1051 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 1052 [20] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 1053 [21] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of*  
1054 *Experimental Psychology: General*, 115:155–174.
- 1055 [22] Ewens, W. J. (2003). On Estimating  $P$  Values by Monte Carlo Methods. *American Journal of*  
1056 *Human Genetics*, 72(2):496–498.
- 1057 [23] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*  
1058 *Transactions of the Royal Society A*, 222(602):309–368.
- 1059 [24] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.  
1060 *School Science and Mathematics*, 100(6):310–318.
- 1061 [25] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather  
1062 prediction” task? individual variability in strategies for probabilistic category learning. *Learning*  
1063 *and Memory*, 9:408–418.
- 1064 [26] Goldman, N. and Whelan, S. (2000). Statistical Tests of Gamma-Distributed Rate Heterogeneity  
1065 in Models of Sequence Evolution in Phylogenetics. *Molecular Biology and Evolution*, 17(6):975–978.
- 1066 [27] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*  
1067 *Cognition and Development*, 13(1):19–37.
- 1068 [28] Halekoh, U. and Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric  
1069 Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *Journal of*  
1070 *Statistical Software*, 59(9):1–32.
- 1071 [29] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual  
1072 learning, pages 212–221. Sage Publications.

- 1073 [30] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-  
1074 ioral and neural signatures of transforming experiences into memories. *Nature Human Behaviour*,  
1075 5:905–919.
- 1076 [31] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge au-  
1077 tomatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*,  
1078 9:doi.org/10.3389/fpsyg.2018.00133.
- 1079 [32] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-  
1080 peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–  
1081 4008.
- 1082 [33] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating  
1083 empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 1084 [34] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.  
1085 Columbia University Press.
- 1086 [35] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,  
1087 326(7382):213–216.
- 1088 [36] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).  
1089 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International  
1090 Journal of Environmental Research and Public Health*, 18(5):2672.
- 1091 [37] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 1092 [38] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 1093 [39] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.  
1094 *The Chronicle of Higher Education*, 21:1–5.
- 1095 [40] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic  
1096 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,  
1097 104:211–240.

- 1098 [41] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic  
1099 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.
- 1100 [42] Macellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of*  
1101 *Educational Studies*, 53(2):129–147.
- 1102 [43] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”  
1103 function? *Psychological Review*, 128(4):711–725.
- 1104 [44] Manning, J. R. (2023). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,  
1105 *Handbook of Human Memory*. Oxford University Press.
- 1106 [45] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension  
1107 for adding LLM-driven chatbots to interactive notebooks. [https://github.com/ContextLab/  
1108 chatify](https://github.com/ContextLab/chatify).
- 1109 [46] Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type i error  
1110 and power in linear mixed models. *Journal of Memory and Language*, 94:305–315.
- 1111 [47] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and  
1112 projection for dimension reduction. *arXiv*, 1802(03426).
- 1113 [48] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold  
1114 Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 1115 [49] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of  
1116 mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 1117 [50] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,  
1118 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to  
1119 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 1120 [51] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-  
1121 tations in vector space. *arXiv*, 1301.3781.

- 1122 [52] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications  
1123 from a national survey of language educators. *System*, 97:102431.
- 1124 [53] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of  
1125 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1126 [54] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).  
1127 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*  
1128 *Neuroscience*, 17(4):367–376.
- 1129 [55] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 1130 [56] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models.  
1131 *arXiv*, 2208.02957.
- 1132 [57] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG  
1133 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,  
1134 7:43916.
- 1135 [58] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*  
1136 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 1137 [59] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.  
1138 *Biological Cybernetics*, 45(1):35–41.
- 1139 [60] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in  
1140 higher education: unmasking power and raising questions about the movement’s democratic  
1141 potential. *Educational Theory*, 63(1):87–110.
- 1142 [61] Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for a zero random  
1143 effect variance or polynomial regression in additive and linear mixed models. *Computational*  
1144 *Statistics & Data Analysis*, 52(7):3283–3299.
- 1145 [62] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter  
1146 Student conceptions and conceptual learning in science. Routledge.

- 1147 [63] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-  
1148 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*  
1149 *tion in Nursing*, 22:32–42.
- 1150 [64] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching  
1151 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 1152 [65] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for  
1153 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*  
1154 *Mathematics Education*, 35(5):305–329.
- 1155 [66] Snijders, T. A. B. and Bosker, R. (2011). More powerful tests for variance parameters. In  
1156 *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, chapter 6, pages  
1157 94–108. Sage Publications, 2nd edition.
- 1158 [67] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal*  
1159 *Medicine*, 21:524–530.
- 1160 [68] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,  
1161 Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023).  
1162 LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 1163 [69] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Pio-  
1164 ntkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust  
1165 detection of AI-generated texts. *arXiv*, 2306.04723.
- 1166 [70] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold?  
1167 distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of*  
1168 *the Cognitive Science Society*, 43(43).
- 1169 [71] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and  
1170 Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing*  
1171 *Systems*.

<sub>1172</sub> [72] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned  
<sub>1173</sub> from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.

<sub>1174</sub> [73] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from  
<sub>1175</sub> free courses. *The Chronicle of Higher Education*, 19(7):1–4.

<sub>1176</sub> [74] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is  
<sub>1177</sub> automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*  
<sub>1178</sub> *Research Methods*, 50:2597–2605.