

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

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

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

Dartmouth College, Hanover, NH 03755, USA

<sup>2</sup>Akili Interactive Labs

Boston, MA 02110, USA

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

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

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

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

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

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

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

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

45 acquired information to the scaffolding of one’s existing knowledge or experience [4, 8, 10, 11, 53]?  
46 Or weaving a lecture’s atomic elements (e.g., its component words) into a structured network that  
47 describes how those individual elements are related [33]? Conceptual understanding could also  
48 involve building a mental model that transcends the meanings of those individual atomic elements  
49 by reflecting the deeper meaning underlying the gestalt whole [30, 34, 50].

50 The difference between “understanding” and “memorizing,” as framed by researchers in ed-  
51 ucation, cognitive psychology, and cognitive neuroscience (e.g., 19, 22, 26, 34, 50), 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 models  
56 (e.g., 5–7, 9, 12, 32, 41) also attempt to capture the deeper meaning *underlying* those atomic elements.  
57 These models consider not only the co-occurrences of those elements within and across documents,  
58 but (in many cases) also patterns in how those elements appear across different scales (e.g., sen-  
59 tences, paragraphs, chapters, etc.), the temporal and grammatical properties of the elements, and  
60 other high-level characteristics of how they are used [35, 36]. According to these models, the deep  
61 conceptual meaning of a document may be captured by a feature vector in a high-dimensional  
62 representation space, wherein nearby vectors reflect conceptually related documents. A model  
63 that succeeds at capturing an analogue of “understanding” is able to assign nearby feature vectors  
64 to two conceptually related documents, *even when the specific words contained in those documents have*  
65 *very little overlap.*

66 Given these insights, what form might a representation of the sum total of a person’s knowledge  
67 take (speculatively)? First, we might require a means of systematically describing or representing  
68 the nearly infinite set of possible things a person could know. Second, we might want to account  
69 for potential associations between different concepts. For example, the concepts of “fish” and  
70 “water” might be associated in the sense that fish live in water. Third, knowledge may have  
71 a critical dependency structure, such that knowing about a particular concept might require first  
72 knowing about a set of other concepts. For example, understanding the concept of a fish swimming

73 in water first requires understanding what fish and water *are*. Fourth, as we learn, our “current  
74 state of knowledge” should change accordingly. Learning new concepts should both update our  
75 characterizations of “what is known” and also unlock any now-satisfied dependencies of those  
76 newly learned concepts so that they are “tagged” as available for future learning.

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

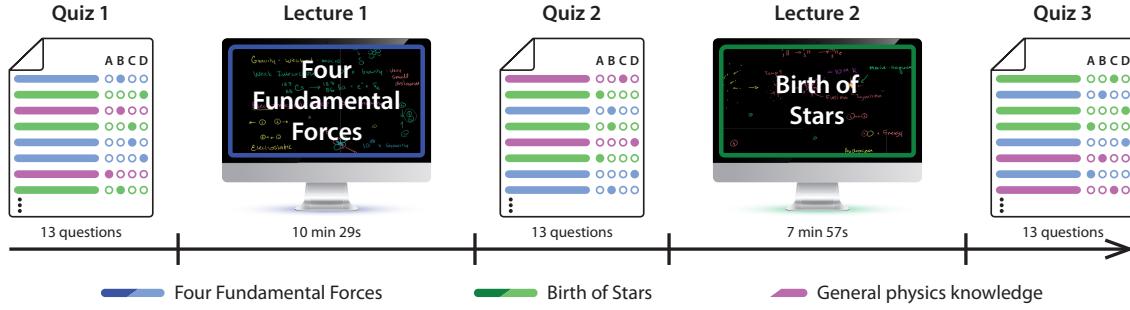
85 Our primary research goal is to advance our understanding of what it means to acquire deep,  
86 real-world conceptual knowledge. Traditional laboratory approaches to studying learning and  
87 memory (e.g., list-learning studies) often draw little distinction between memorization and under-  
88 standing. Instead, these studies typically focus on whether information is effectively encoded or  
89 retrieved, rather than whether the information is *understood*. Approaches to studying conceptual  
90 learning, such as category learning experiments, can begin to investigate the distinction between  
91 memorization and understanding, often by training participants to distinguish arbitrary or random  
92 features in otherwise meaningless categorized stimuli [1, 16, 17, 20, 24, 48]. However the objective  
93 of real-world training, or learning from life experiences more generally, is often to develop new  
94 knowledge that may be applied in *useful* ways in the future. In this sense, the gap between modern  
95 learning theories and modern pedagogical approaches that inform classroom learning strategies is  
96 enormous: most of our theories about *how* people learn are inspired by experimental paradigms  
97 and models that have only peripheral relevance to the kinds of learning that students and teachers  
98 actually seek [22, 34]. To help bridge this gap, our study uses course materials from real on-  
99 line courses to inform, fit, and test models of real-world conceptual learning. We also provide a  
100 demonstration of how our models can be used to construct “maps” of what students know, and

101 how their knowledge changes with training. In addition to helping to visually capture knowledge  
102 (and changes in knowledge), we hope that such maps might lead to real-world tools for improving  
103 how we educate. Taken together, our work shows that existing course materials and evaluative  
104 tools like short multiple-choice quizzes may be leveraged to gain highly detailed insights into what  
105 students know and how they learn.

## 106 Results

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

121 We asked participants in our study to complete brief multiple-choice quizzes before, between,  
122 and after watching two lecture videos from the Khan Academy [29] platform (Fig. 1). The first  
123 lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics:  
124 gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*,  
125 provided an overview of our current understanding of how stars form. We selected these particular  
126 lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad



**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.

audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training on participants' abilities to learn from the lectures. To this end, we selected two introductory videos that were intended to be viewed at the start of students' training in their respective content areas. Second, we wanted the two lectures to have some related content, so that we could test our approach's ability to distinguish similar conceptual content. To this end, we chose two videos from the same Khan Academy course domain, "Cosmology and Astronomy." Third, we sought to minimize dependencies and specific overlap between the videos. For example, we did not want participants' abilities to understand one video to (directly) influence their abilities to understand the other. To satisfy this last criterion, we chose videos from two different lecture series (Lectures 1 and 2 were from the "Scale of the Universe" and "Stars, Black Holes, and Galaxies" series, respectively).

We also wrote a set of multiple-choice quiz questions that we hoped would enable us to evaluate participants' knowledge about each individual lecture, along with related knowledge 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).



**Figure 2: Modeling course content.** **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The full set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding model. **B. Constructing lecture content trajectories.** After training the model on the sliding windows from both lectures, we transform each lecture into a “trajectory” through text embedding space by joining the embedding coordinates of successive sliding windows parsed from its transcript. **C. Embedding multiple lectures and questions in a shared space.** We apply the same model (trained on the two lectures’ windows) to both lectures, along with the text of each question in our pool (Supp. Tab. 1), to project them into a shared text embedding space. This results in one trajectory per lecture and one coordinate for each question. Here, we have projected the 15-dimensional embeddings onto their first 3 principal components for visualization.

145 To study in detail how participants’ conceptual knowledge changed over the course of the  
 146 experiment, we first sought to model the conceptual content presented to them at each moment  
 147 throughout each of the two lectures. We adapted an approach we developed in prior work [23]  
 148 to identify the latent themes in the lectures using a topic model [6]. Briefly, topic models take  
 149 as input a collection of text documents, and learn a set of “topics” (i.e., latent themes) from their  
 150 contents. Once fit, a topic model can be used to transform arbitrary (potentially new) documents  
 151 into sets of “topic proportions,” describing the weighted blend of learned topics reflected in their  
 152 texts. We parsed automatically generated transcripts of the two lectures into overlapping sliding  
 153 windows, where each window contained the text of the lecture transcript from a particular time  
 154 span. We treated the set of text snippets (across all of these windows) as documents to fit the  
 155 model (Fig. 2A; see *Constructing text embeddings of multiple lectures and questions*). Transforming the  
 156 text from every sliding window with the model yielded a number-of-windows by number-of-topics  
 157 (15) topic-proportions matrix describing the unique mixture of broad themes from both lectures  
 158 reflected in each window’s text. Each window’s “topic vector” (i.e., column of the topic-proportions  
 159 matrix) is analogous to a coordinate in a 15-dimensional space whose axes are topics discovered

160 by the model. Within this space, each lecture's sequence of topic vectors (i.e., corresponding to its  
161 transcript's overlapping text snippets across sliding windows) forms a *trajectory* that captures how  
162 its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution  
163 of one topic vector for each second of video (i.e., 1 Hz).

164 We hypothesized that a topic model trained on transcripts of the two lectures should also capture  
165 the conceptual knowledge probed by each quiz question. If indeed the topic model could capture  
166 information about the deeper conceptual content of the lectures (i.e., beyond surface-level details  
167 such as particular word choices), then we should be able to recover a correspondence between each  
168 lecture and questions *about* each lecture. Importantly, such a correspondence could not solely arise  
169 from superficial text matching between lecture transcripts and questions, since the lectures and  
170 questions used different words. Simply comparing the average topic weights from each lecture and  
171 question set (averaging across time and questions, respectively) reveals a striking correspondence  
172 (Supp. Fig. 2). Specifically, the average topic weights from Lecture 1 are strongly correlated with the  
173 average topic weights from Lecture 1 questions ( $r(13) = 0.809, p < 0.001, 95\% \text{ CI} = [0.633, 0.962]$ ), and the average topic weights from Lecture 2 are strongly correlated with the  
174 average topic weights from Lecture 2 questions ( $r(13) = 0.728, p = 0.002, 95\% \text{ CI} = [0.456, 0.920]$ ).  
175 At the same time, the average topic weights from the two lectures are *negatively* correlated with  
176 their non-matching question sets (Lecture 1 video vs. Lecture 2 questions:  $r(13) = -0.547, p = 0.035,$   
177  $95\% \text{ CI} = [-0.812, -0.231]$ ; Lecture 2 video vs. Lecture 1 questions:  $r(13) = -0.612, p = 0.015, 95\%$   
178  $\text{CI} = [-0.874, -0.281]$ ), indicating that the topic model also exhibits some degree of specificity. The  
179 full set of pairwise comparisons between average topic weights for the lectures and question sets  
180 is reported in Supplementary Figure 2.  
181

182 It is important to clarify that although we use topic model-derived embeddings to *characterize*  
183 the conceptual content of the lectures and questions, we do not claim that the topic model itself  
184 *understands* the conceptual content of the lectures or questions. Rather, we view the topic model as  
185 a tool for capturing the *structure* of the conceptual content of the lectures and questions in a way  
186 that enables us to capture, quantify, and track and predict participants' knowledge.

187 Another, more sensitive, way of summarizing the conceptual content of the lectures and ques-



**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.

tions is to look at *variability* in how topics are weighted over time and across different questions (Fig. 3). Intuitively, the *variability* in the expression of a given topic relates to how much “information” [18] the lecture (or question set) reflects about that topic. For example, suppose a given topic is weighted on heavily throughout a lecture. That topic might be characteristic of some aspect or property of the lecture *overall* (conceptual or otherwise), but unless the topic’s weights changed in meaningful ways over time, the topic would be a poor indicator of any *specific* conceptual content in the lecture. We therefore also compared the variances in topic weights (across time or questions) between the lectures and questions. The variability in topic expression (over time and across questions) was similar for the Lecture 1 video and questions ( $r(13) = 0.824, p < 0.001$ , 95% CI = [0.696, 0.973]) and the Lecture 2 video and questions ( $r(13) = 0.801, p < 0.001$ , 95% CI = [0.539, 0.958]). Simultaneously, as reported in Figure 3B, the variability in topic expression across *different* videos and lecture-specific questions (i.e., Lecture 1 video vs. Lecture 2 questions; Lecture 2 video vs. Lecture 1 questions) were negatively correlated, and neither video’s topic variability was reliably correlated with the topic variability across general physics knowledge

202 questions. Taken together, the analyses reported in Figure 3 and Supplementary Figure 2 indicate  
203 that a topic model fit to the videos’ transcripts can also reveal correspondences (at a coarse scale)  
204 between the lectures and questions.

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

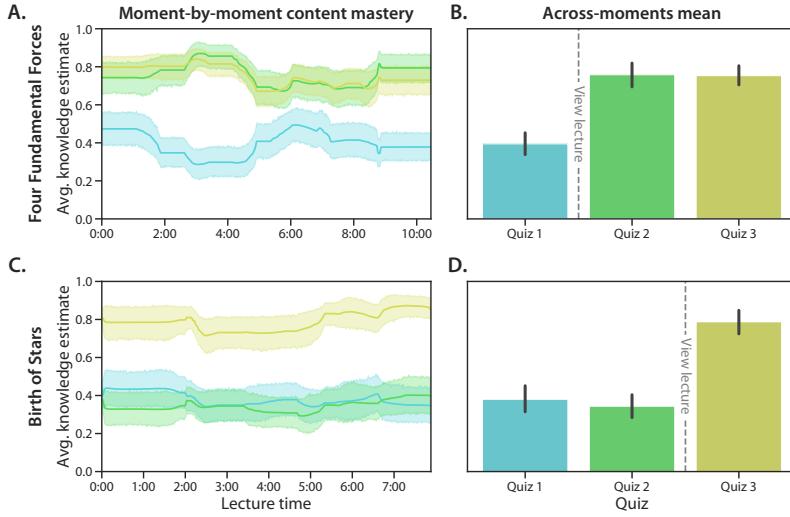
223 The ability to quantify how much each question is “asking about” the content from each moment  
224 of the lectures could enable high-resolution insights into participants’ knowledge. Traditional  
225 approaches to estimating how much a student “knows” about the content of a given lecture entail  
226 computing the proportion of correctly answered questions. But if two students receive identical  
227 scores on an exam, might our modeling framework help us to gain more nuanced insights into the  
228 *specific* content that each student has mastered (or failed to master)? For example, a student who  
229 misses three questions that were all about the same concept (e.g., concept *A*) will have gotten the



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

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

238 We developed a simple formula (Eqn. 1) for using a participant’s responses to a small set of  
 239 multiple-choice questions to estimate how much the participant “knows” about the concept re-  
 240 flected by any arbitrary coordinate,  $x$ , in text embedding space (e.g., the content reflected by any  
 241 moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially, the  
 242 estimated knowledge at coordinate  $x$  is given by the weighted average proportion of quiz questions



**Figure 5: Estimating moment-by-moment knowledge acquisition.** **A. Moment-by-moment knowledge about the *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’s color). The traces are averaged across participants. **B. Average estimated knowledge about the *Four Fundamental Forces*.** Each bar displays the across-timepoint average knowledge, estimated using the responses to one quiz’s questions. **C. Moment-by-moment knowledge about the *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 the *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.

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 timecourse describing how much “knowledge” the participant has about 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 timecourses 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

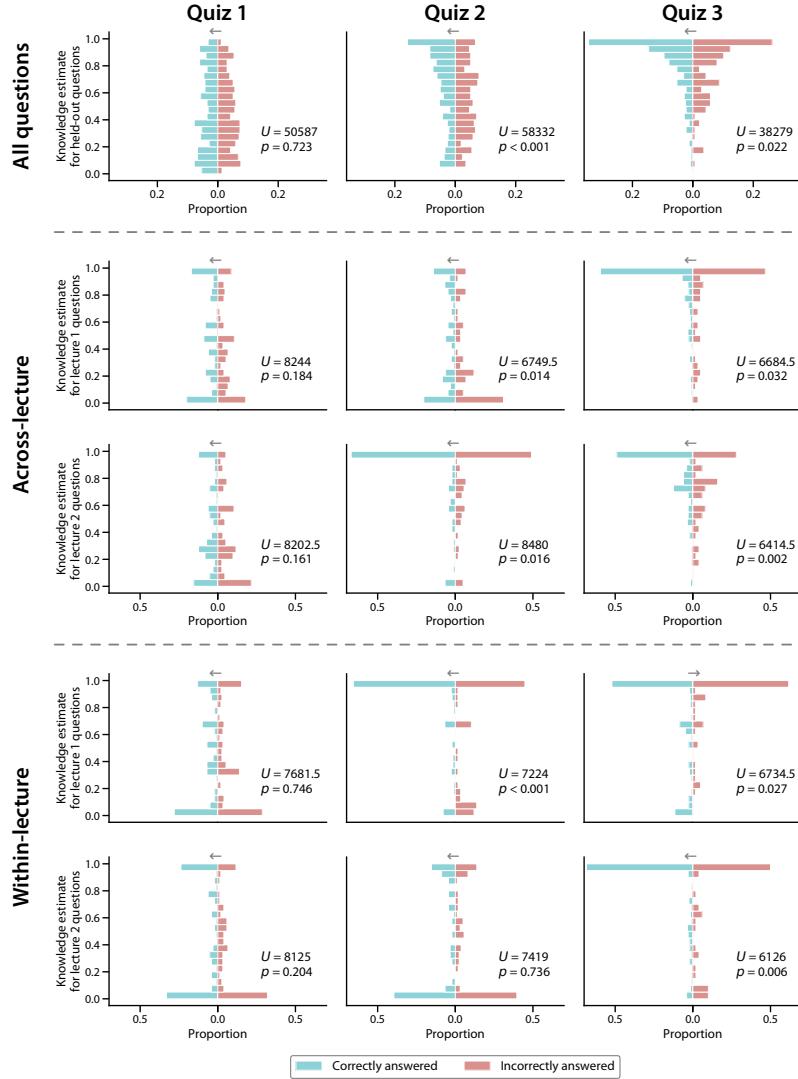
254 participants actually know. As one sanity check, we anticipated that the knowledge estimates  
255 should reflect a content-specific “boost” in participants’ knowledge after watching each lecture.  
256 In other words, if participants learn about each lecture’s content when they watch each lecture,  
257 the knowledge estimates should capture that. After watching the *Four Fundamental Forces* lecture,  
258 participants should exhibit more knowledge for the content of that lecture than they had before,  
259 and that knowledge should persist for the remainder of the experiment. Specifically, knowledge  
260 about that lecture’s content should be relatively low when estimated using Quiz 1 responses, but  
261 should increase when estimated using Quiz 2 or 3 responses (Fig. 5B). Indeed, we found that  
262 participants’ estimated knowledge about the content of the *Four Fundamental Forces* was substan-  
263 tially higher on Quiz 2 versus Quiz 1 ( $t(49) = 8.764, p < 0.001$ ) and on Quiz 3 versus Quiz 1  
264 ( $t(49) = 10.519, p < 0.001$ ). We found no reliable differences in estimated knowledge about that  
265 lecture’s content on Quiz 2 versus 3 ( $t(49) = 0.160, p = 0.874$ ). Similarly, we hypothesized (and  
266 subsequently confirmed) that participants should show greater estimated knowledge about the  
267 content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 5D). Specifically, since  
268 participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their  
269 knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a “boost” on  
270 Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge  
271 about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ( $t(49) = 1.013, p = 0.316$ ), but the  
272 estimated knowledge was substantially higher on Quiz 3 versus 2 ( $t(49) = 10.561, p < 0.001$ ) and  
273 Quiz 3 versus 1 ( $t(49) = 8.969, p < 0.001$ ).

274 If we are able to accurately estimate a participant’s knowledge about the content tested by a  
275 given question, our estimates of their knowledge should carry some predictive information about  
276 whether the participant is likely to answer that question correctly or incorrectly. We developed a  
277 statistical approach to test this claim. For each question, in turn, we used Equation 1 to estimate  
278 each participant’s knowledge at the given question’s embedding space coordinate, using all *other*  
279 questions that participant answered on the same quiz. For each quiz, we grouped these estimates  
280 into two distributions: one for the estimated knowledge at the coordinates of *correctly* answered  
questions, and another for the estimated knowledge at the coordinates of *incorrectly* answered

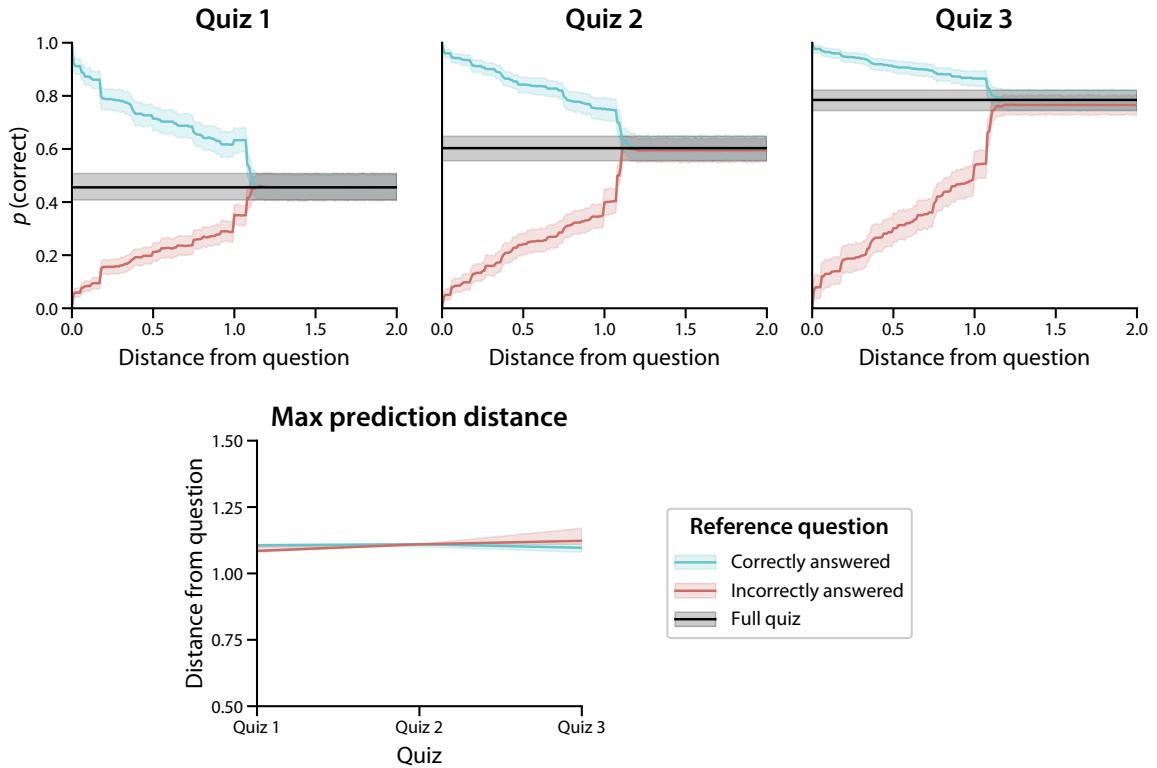
282 questions (Fig. 6). We then used independent samples *t*-tests to compare the means of these  
283 distributions of estimated knowledge.

284 For the initial quizzes participants took (prior to watching either lecture), participants' estimated  
285 knowledge tended to be low overall, and relatively unstructured (Fig. 6, left panel). When we held  
286 out individual questions and estimated their knowledge at the held-out questions' embedding  
287 coordinates, we found no reliable differences in the estimates when the held-out question had been  
288 correctly versus incorrectly answered ( $t(633) = 0.577, p = 0.564$ ). This reflects a floor effect: when  
289 knowledge is low everywhere, there is little signal to differentiate between what is known versus  
290 unknown. After watching the first lecture, estimated knowledge for held-out correctly answered  
291 questions (from the second quiz; Fig. 6, middle panel) exhibited a positive shift relative to held-  
292 out incorrectly answered questions ( $t(633) = 3.961, p < 0.001$ ). This second quiz provides the  
293 maximally sensitive test for our knowledge predictions, since (if knowledge is estimated accurately)  
294 participants' Quiz 2 responses should demonstrate specific knowledge about Lecture 1 content,  
295 but knowledge about Lecture 2 and general physics concepts should be roughly unchanged from  
296 before they watched Lecture 1. After watching the second lecture, estimated knowledge (from the  
297 third quiz; Fig. 6, right panel) for *all* questions exhibited a positive shift. However, the estimated  
298 knowledge for held-out correctly answered questions remained greater than that for held-out  
299 incorrectly answered questions ( $t(628) = 2.045, p = 0.041$ ). This third contrast reflects a ceiling  
300 effect: when knowledge is relatively high everywhere, the signal differentiating what is known  
301 versus unknown is relatively weak. Taken together, this set of analyses demonstrates that our  
302 knowledge prediction framework is most informative when participants exhibit variability in their  
303 knowledge of the content captured by the text embedding model.

304 Knowledge estimates need not be limited to the content of the lectures. As illustrated in  
305 Figure 8, our general approach to estimating knowledge from a small number of quiz questions  
306 may be extended to *any* content, given its text embedding coordinate. To visualize how knowledge  
307 "spreads" through text embedding space to content beyond the lectures participants watched, we  
308 first fit a new topic model to the lectures' sliding windows with (up to)  $k = 100$  topics. Conceptually,  
309 increasing the number of topics used by the model functions to increase the "resolution" of the



**Figure 6: Estimating knowledge at the embedding coordinates of held-out questions.** Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held-out correctly (blue) or incorrectly (red) answered question. The *t*-tests reported in each panel are between the distributions of estimated knowledge at the coordinates of correctly versus incorrectly answered held-out questions.



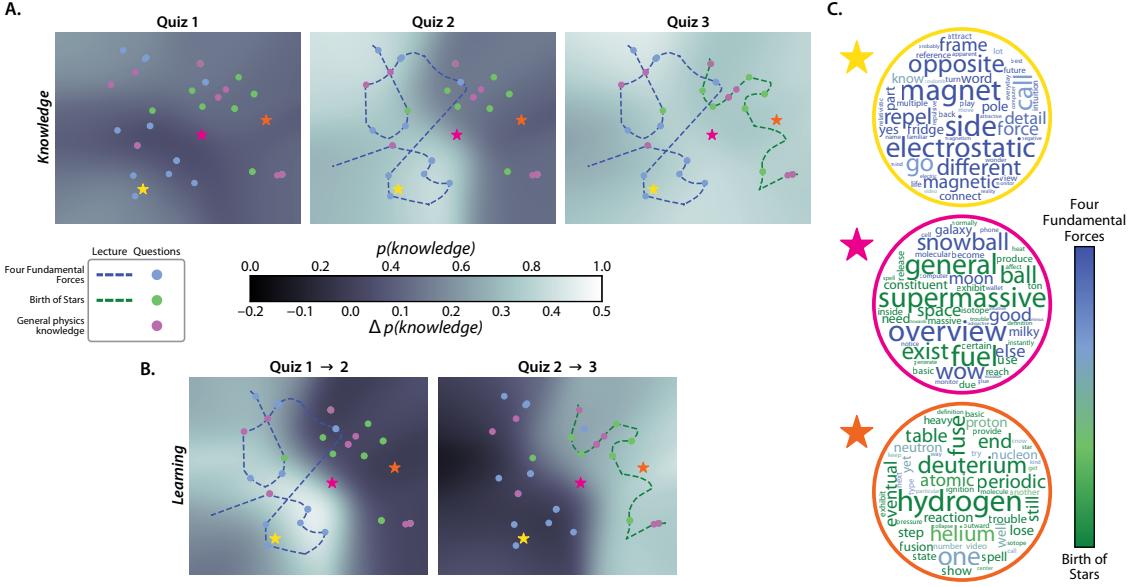
**Figure 7: Caption title.** Caption content.

embedding space, providing a greater ability to estimate knowledge for content that is highly similar to (but not precisely the same as) that contained in the two lectures. This change in the number of topics overcame an undesirable behavior in the UMAP embedding procedure [?], whereby embedding coordinates for the 15-topic model tended to be “clumped” into separated clusters, rather than forming a smooth trajectory through the 2D space. When we increased the number of topics to 100, the embedding coordinates in the 2D space formed a smooth trajectory through the space, with substantially less clumping (Fig. 8). We note that we used these 2D maps solely for visualization; all relevant comparisons, distance computations, and statistical tests we report above were carried out in the original 15-dimensional space, using the 15-topic model. Aside from increasing the number of topics from 15 to 100, all other procedures and model parameters were carried over from the preceding analyses. As in our other analyses, we resampled each

321 lecture’s topic trajectory to 1 Hz and projected each question into a shared text embedding space.  
322 We projected the resulting 100-dimensional topic vectors (for each second of video and each quiz  
323 question) onto a shared 2-dimensional plane (see *Creating knowledge and learning map visualizations*).  
324 Next, we sampled points from a  $100 \times 100$  grid of coordinates that evenly tiled a rectangle enclos-  
325 ing the 2D projections of the videos and questions. We used Equation 4 to estimate participants’  
326 knowledge at each of these 10,000 sampled locations, and averaged these estimates across par-  
327 ticipants to obtain an estimated average *knowledge map* (Fig. 8A). Intuitively, the knowledge map  
328 constructed from a given quiz’s responses provides a visualization of how “much” participants  
329 knew about any content expressible by the fitted text embedding model at the point in time when  
330 they completed that quiz.

331 Several features of the resulting knowledge maps are worth noting. The average knowledge  
332 map estimated from Quiz 1 responses (Fig. 8A, leftmost map) shows that participants tended to  
333 have relatively little knowledge about any parts of the text embedding space (i.e., the shading is  
334 relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked  
335 increase in knowledge on the left side of the map (around roughly the same range of coordinates  
336 traversed by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words,  
337 participants’ estimated increase in knowledge is localized to conceptual content that is nearby (i.e.,  
338 related to) the content from the lecture they watched prior to taking Quiz 2. This localization is  
339 non-trivial: these knowledge estimates are informed only by the embedded coordinates of the  
340 *quiz questions*, not by the embeddings of either lecture (see Eqn. 4). Finally, the knowledge map  
341 estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region  
342 surrounding the embedding of the *Birth of Stars* lecture participants watched immediately prior to  
343 taking Quiz 3.

344 Another way of visualizing these content-specific increases in knowledge after participants  
345 viewed each lecture is displayed in Figure 8B. Taking the point-by-point difference between the  
346 knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map*  
347 that describes the *change* in knowledge estimates from one quiz to the next. These learning maps  
348 highlight that the estimated knowledge increases we observed across maps were specific to the



**Figure 8: Mapping out the geometry of knowledge and learning.** **A.** Average “knowledge maps” estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*). The topic trajectories of the two lectures are indicated by dotted lines (blue: Lecture 1; green: Lecture 2), and the coordinates of each question are indicated by dots (light blue: Lecture 1-related; light green: Lecture 2-related; purple: general physics knowledge). Each map reflects an average across all participants. For individual participants’ maps, see Supplementary Figures 3, 4, and 5. **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 6 and 7. **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 the *Four Fundamental Forces* (blue) versus *Birth of Stars* (green) lectures, on average, across all timepoints’ topic vectors.

349 regions around the embeddings of each lecture, in turn.

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

## 364 Discussion

365 We developed a computational framework that uses short multiple-choice quizzes to gain nuanced  
366 insights into what learners know and how their knowledge changes with training. First, we show  
367 that our approach can automatically match the conceptual knowledge probed by individual quiz  
368 questions to the corresponding moments in lecture videos when those concepts were presented  
369 (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment "knowledge traces"  
370 that reflect the degree of knowledge participants have about each video's time-varying content,  
371 and capture temporally specific increases in knowledge after viewing each lecture (Fig. 5). We  
372 also show that these knowledge estimates can generalize to held-out questions (Fig. 6). Finally,  
373 we use our framework to construct visual maps that provide snapshot estimates of how much  
374 participants know about any concept within the scope of our text embedding model, and how

375 much their knowledge of those concepts changes with training (Fig. 8).

376 We view our work as making several contributions to the study of how people acquire con-  
377 ceptual knowledge. First, from a methodological standpoint, our modeling framework provides  
378 a systematic means of mapping out and characterizing knowledge in maps that have infinite (ar-  
379 bitrarily many) numbers of coordinates, and of “filling out” those maps using relatively small  
380 numbers of multiple choice quiz questions. Our experimental finding that we can use these maps  
381 to predict responses to held-out questions (Fig. 6) also has important psychological implications.  
382 One such psychological implication is that concepts that are assigned to nearby coordinates by the  
383 text embedding model also appear to be “known to a similar extent” (as reflected by participants’  
384 responses to held-out questions). This suggests that participants also *conceptualize* similarly the  
385 content reflected by nearby embedding coordinates. A second psychological implication is that  
386 something about how participants’ knowledge “spreads” across concepts is being captured by the  
387 knowledge maps we are inferring from their quiz responses (e.g., Figs. 7, 8). In other words, our  
388 study shows that knowledge about a given concept implies knowledge about related concepts.

389 In our study, we characterize the “coordinates” of participants’ knowledge using a relatively  
390 simple “bag of words” text embedding model [6, LDA; ]. More sophisticated text embedding  
391 models, such as transformer-based models [14, 45, 55, 56] can learn complex grammatical and  
392 semantic relationships between words, higher-order syntactic structures, stylistic features, and  
393 more. We considered using transformer-based models in our study, but we were surprised to  
394 find that the text embeddings derived from these models were surprisingly uninformative with  
395 respect to differentiating or otherwise characterizing the conceptual content of the lectures and  
396 questions we used. We suspect that this reflects a broader challenge in constructing models that  
397 are high-resolution within a given domain (e.g., the domain of physics lectures and questions)  
398 and sufficiently broad so as to enable them to cover a wide range of domains. For example,  
399 we found that the embeddings derived even from much larger and more modern models like  
400 BERT [14], GPT [56], LLaMa [55], and others that are trained on enormous text corpora, end up  
401 yielding poor resolution within the content space spanned by individual course videos (Supp.  
402 Fig. 8). Whereas the LDA embeddings of the lectures and questions are “near” each other (i.e.,

the convex hull enclosing the two lectures' trajectories is highly overlapping with the convex hull enclosing the questions' embeddings), the BERT embeddings of the lectures and questions are instead largely distinct (top row of Supp. Fig. 8). The LDA embeddings of the questions for each lecture and the corresponding lecture's trajectory are also similar. For example, as shown in Fig. 2C, the LDA embeddings for *Four Fundamental Forces* questions (blue dots) appear closer to the *Four Fundamental Forces* lecture trajectory (blue line), whereas the LDA embeddings for *Birth of Stars* questions (green dots) appear closer to the *Birth of Stars* lecture trajectory (green line). The BERT embeddings of the lectures and questions do not show this property (Supp. Fig. 8). We also examined per-question “content matches” between individual questions and individual moments of each lecture (Figs. 4, 8). The timeseries plot of individual questions’ correlations are different from each other when computed using LDA (e.g., the traces can be clearly visually separated), whereas the correlations computed from BERT embeddings of different questions all look very similar. This tells us that LDA is capturing some differences in content between the questions, whereas BERT is not. The timeseries plots of individual questions’ correlations have clear “peaks” when computed using LDA, but not when computed using BERT. This tells us that LDA is capturing a “match” between the content of each question and a relatively well-defined time window of the corresponding lectures. The BERT embeddings appear to blur together the content of the questions versus specific moments of each lecture. Finally, we also compared the pairwise correlations between embeddings of questions within versus across content areas (i.e., content covered by the individual lectures, lecture-specific questions, and by the “general physics knowledge” questions). The LDA embeddings show a strong contrast between same-content embeddings versus across-content embeddings. In other words, the embeddings of questions about the *Four Fundamental Forces* material are highly correlated with the embeddings of the *Four Fundamental Forces* lecture, but not with the embeddings of *Birth of Stars*, questions about *Birth of Stars*, or general physics knowledge questions. We see a similar pattern with the LDA embeddings of the *Birth of Stars* questions (Fig. 3, Supp. Fig. 2). In contrast, the BERT embeddings are all highly correlated with each other (Supp. Fig. 8). Taken together, these comparisons illustrate how LDA (trained on the specific content in question) provides both coverage of the requisite

431 material and specificity at the level of the content covered by individual questions. BERT, on the  
432 other hand, essentially assigns both lectures and all of the questions (which are all broadly about  
433 “physics”) into a tiny region of its embedding space, thereby blurring out meaningful distinctions  
434 between different specific concepts covered by the lectures and questions. We note that these are  
435 not criticisms of BERT (or other large language models trained on large and diverse corpora).  
436 Rather, our point is that simple fine-tuned models trained on a relatively small but specialized  
437 corpus can outperform much more complicated models trained on much larger corpora, when we  
438 are specifically interested in capturing subtle conceptual differences at the level of a single course  
439 lecture or question. Of course if our goal had been to find a model that generalized to many  
440 different content areas, we would expect our approach to perform comparatively poorly relative to  
441 BERT or other much larger models. We suggest that bridging the tradeoff between high resolution  
442 within each content area versus the ability to generalize to many different content areas will be an  
443 important challenge for future work in this domain.

444 At the opposite end of the spectrum from large language models, one could also imagine  
445 *simplifying* some aspects of our LDA-based approach by computing simple word overlap metrics.  
446 For example, the Jaccard similarity between text  $A$  and  $B$  is computed as the number of unique  
447 words in the intersection of words from  $A$  and  $B$  divided by the number of unique words in  
448 the union of words from  $A$  and  $B$ . In a supplemental analysis (Supp. Fig. 9), we compared the  
449 LDA-based question-lecture matches we reported in Figure 4 with the Jaccard similarities between  
450 each question and each sliding window of text from the corresponding lecture. As shown in  
451 Supplementary Figure 9, this simple word-matching approach does not appear to capture the  
452 same level of specificity as the LDA-based approach. For example, whereas the LDA-based  
453 approach often yields a clear peak in the timeseries of correlations between each question and  
454 the corresponding lecture, the Jaccard similarity-based approach does not. Furthermore, these  
455 LDA-based matches appear to capture conceptual overlaps between the questions and lectures  
456 (Supp. Tab. 3), whereas simple word matching does not. For example, one of the example  
457 questions examined in Supplementary Figure 9 asks “Which of the following occurs as a cloud of  
458 atoms gets more dense?”. The LDA-based matches identify lecture timepoints where the relevant

459 *topics* are discussed (e.g., when words like “cloud,” “atom,” “dense,” etc., are mentioned *together*).  
460 The Jaccard similarity-based matches, on the other hand, are strong when *any* of these words are  
461 mentioned, even if they do not occur together.

462 Over the past several years, the global pandemic has forced many educators to suddenly  
463 adapt to teaching remotely [28, 42, 52, 57]. This change in world circumstances is happening  
464 alongside (and perhaps accelerating) geometric growth in the availability of high-quality online  
465 courses from platforms such as Khan Academy [29], Coursera [58], EdX [31], and others [49].  
466 Continued expansion of the global internet backbone and improvements in computing hardware  
467 have also facilitated improvements in video streaming, enabling videos to be easily shared and  
468 viewed by increasingly large segments of the world’s population. This exciting time for online  
469 course instruction provides an opportunity to re-evaluate how we, as a global community, educate  
470 ourselves and each other. For example, we can ask: what defines an effective course or training  
471 program? Which aspects of teaching might be optimized and/or augmented by automated tools?  
472 How and why do learning needs and goals vary across people? How might we lower barriers of  
473 access to a high-quality education?

474 Alongside these questions, there is a growing desire to extend existing theories beyond the  
475 domain of lab testing rooms and into real classrooms [27]. In part, this has led to a recent  
476 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better  
477 reflect more ethologically valid phenomena that are more directly relevant to real-world situations  
478 and behaviors [43]. In turn, this has brought new challenges in data analysis and interpretation. A  
479 key step towards solving these challenges will be to build explicit models of real-world scenarios  
480 and how people behave in them (e.g., models of how people learn conceptual content from real-  
481 world courses, as in our current study). A second key step will be to understand which sorts of  
482 signals derived from behaviors and/or other measurements (e.g., neurophysiological data; 2, 15, 40,  
483 44, 46) might help to inform these models. A third major step will be to develop and employ reliable  
484 ways of evaluating the complex models and data that are a hallmark of naturalistic paradigms.

485 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also  
486 relate to the notion of “theory of mind” of other individuals [21, 25, 39]. Considering others’ unique

487 perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and  
488 communicate [47, 51, 54]. One could imagine future extensions of our work (e.g., analogous to  
489 the knowledge and learning maps shown in Fig. 8), that attempt to characterize how well-aligned  
490 different people’s knowledge bases or backgrounds are. In turn, this might be used to model how  
491 knowledge (or other forms of communicable information) flows not just between teachers and  
492 students, but between friends having a conversation, individuals on a first date, participants at  
493 a business meeting, doctors and patients, experts and non-experts, political allies or adversaries,  
494 and more. For example, the extent to which two people’s knowledge maps “match” or “align” in  
495 a given region of text embedding space might serve as a predictor of how effectively they will be  
496 able to communicate about the corresponding conceptual content.

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

## 503 Materials and methods

### 504 Participants

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

510 Participants’ ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09  
511 years). A total of 15 participants reported their gender as male and 35 participants reported their

512 gender as female. A total of 49 participants reported their native language as “English” and 1  
513 reported having another native language. A total of 47 participants reported their ethnicity as  
514 “Not Hispanic or Latino” and three reported their ethnicity as “Hispanic or Latino.” Participants  
515 reported their races as White (32 participants), Asian (14 participants), Black or African American  
516 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other  
517 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

530 Participants reported their undergraduate major(s) as “social sciences” (28 participants), “nat-  
531 ural sciences” (16 participants), “professional” (e.g., pre-med or pre-law; 8 participants), “mathe-  
532 matics and engineering” (7 participants), “humanities” (4 participants), or “undecided” (3 partici-  
533 pants). Note that some participants selected multiple categories for their undergraduate major(s).  
534 We also asked participants about the courses they had taken. In total, 45 participants reported hav-  
535 ing taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan  
536 Academy courses. Of those who reported having watched at least one Khan Academy course,  
537 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8  
538 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We  
539 also asked participants about the specific courses they had watched, categorized under different

540 subject areas. In the “Mathematics” area, participants reported having watched videos on AP  
541 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-  
542 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry  
543 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential  
544 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),  
545 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other  
546 videos not listed in our survey (5 participants). In the “Science and engineering” area, participants  
547 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-  
548 ipants); Physics, AP Physics I, or AP Physics II (18 participants); Biology, AP Biology; or High  
549 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed  
550 in our survey (5 participants). We also asked participants whether they had specifically seen the  
551 videos used in our experiment. Of the 45 participants who reported having taken at least  
552 one Khan Academy course in the past, 44 participants reported that they had not watched the *Four*  
553 *Fundamental Forces* video, and 1 participant reported that they were not sure whether they had  
554 watched it. All participants reported that they had not watched the *Birth of Stars* video. When  
555 we asked participants about non-Khan Academy online courses, they reported having watched  
556 or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test  
557 preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 partic-  
558 ipants), Computing (2 participants), and other categories not listed in our survey (17 participants).  
559 Finally, we asked participants about in-person courses they had taken in different subject areas.  
560 They reported taking courses in Mathematics (38 participants), Science and engineering (37 par-  
561 ticipants), Arts and humanities (34 participants), Test preparation (27 participants), Economics  
562 and finance (26 participants), Computing (14 participants), College and careers (7 participants), or  
563 other courses not listed in our survey (6 participants).

## 564 Experiment

565 We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces*  
566 (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force;

duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed; duration: 7 minutes and 57 seconds). All participants viewed the videos in the same order (i.e., *Four Fundamental Forces* followed by *Birth of Stars*). While we are not aware of any specific confounds of viewing order, nor have we been aware of how or why viewing order might influence our main findings, we acknowledge that we did not control for potential order effects in our study.

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

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

Over the course of the experiment, participants completed three 13-question multiple-choice quizzes: the first before viewing Lecture 1, the second between Lectures 1 and 2, and the third after viewing Lecture 2 (see Fig. 1). The questions appearing on each quiz, for each participant, were randomly chosen from the full set of 39, with the constraints that (a) each quiz contain

595 exactly 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general  
596 physics knowledge, and (b) each question appear exactly once for each participant. The orders of  
597 questions on each quiz, and the orders of answer options for each question, were also randomized.  
598 Our experimental protocol was approved by the Committee for the Protection of Human Subjects  
599 at Dartmouth College. We used this experiment to develop and test our computational framework  
600 for estimating knowledge and learning.

## 601 **Analysis**

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

603 We adapted an approach we developed in prior work [23] to embed each moment of the two  
604 lectures and each question in our pool in a common representational space. Briefly, our approach  
605 uses a topic model (Latent Dirichlet Allocation; 6), trained on a set of documents, to discover a set  
606 of (up to)  $k$  “topics” or “themes.” Formally, each topic is defined as a distribution of weights over  
607 each word in the model’s vocabulary (i.e., the union of all unique words, across all documents,  
608 excluding “stop words.”). Conceptually, each topic is intended to give larger weights to words that  
609 are semantically related or tend to co-occur in the same documents. After fitting a topic model,  
610 each document in the training set, or any *new* document that contains at least some of the words  
611 in the model’s vocabulary, may be represented as a  $k$ -dimensional vector describing how much  
612 the document (most probably) reflects each topic. To select an appropriate  $k$  for our model, we  
613 identified the minimum number of topics that yielded at least one “unused” topic (i.e., in which  
614 all words in the vocabulary were assigned uniform weights) after training. This indicated that  
615 the number of topics was sufficient to capture the set of latent themes present in the two lectures  
616 (from which we constructed our document corpus, as described below). We found this value to  
617 be  $k = 15$  topics. The distribution of weights over words in the vocabulary for each discovered  
618 topic is shown in Supplementary Figure 1, and each topic’s top-weighted words may be found in  
619 Supplementary Table 2.

620 As illustrated in Figure 2A, we start by building up a corpus of documents using overlapping

sliding windows that span each video’s transcript. Khan Academy provides professionally created, manual transcriptions of all videos for closed captioning. However, such transcripts would not be readily available in all contexts to which our framework could potentially be applied. Khan Academy videos are hosted on the YouTube platform, which additionally provides automated captions. We opted to use these automated transcripts (which, in prior work, we have found to be of sufficiently near-human quality to yield reliable data in behavioral studies; 59) when developing our framework in order to make it more directly extensible and adaptable by others in the future.

We fetched these automated transcripts using the `youtube-transcript-api` Python package [13]. The transcripts consisted of one timestamped line of text for every few seconds (mean: 2.34 s; standard deviation: 0.83 s) of spoken content in the video (i.e., corresponding to each individual caption that would appear on-screen if viewing the lecture via YouTube, and when those lines would appear). After scraping the transcripts, we carried out several standard preprocessing steps. First, we removed all punctuation, replaced numbers with their word equivalents (e.g., “1” with “one”), and converted all text to lowercase. Next, we removed all stop words, as defined by the Natural Language Toolkit [3, NLTK; ] English stop word list, augmented with the following additional words that we hand-identified as frequently occurring and non-diagnostic: “actual,” “actually,” “also,” “bit,” “could,” “e,” “even,” “first,” “follow,” “following,” “four,” “let,” “like,” “mc,” “really,” “saw,” “see,” “seen,” “thing,” and “two.” We defined a sliding window length of (up to)  $w = 30$  transcript lines, and assigned each window a timestamp corresponding to the midpoint between the timestamps for its first and last lines. This  $w$  parameter was chosen to match the same number of words per sliding window (rounded to the nearest whole word, and after preprocessing) as the sliding windows we defined in our prior work [23]. The sliding windows contained an average of 73.8 words and lasted for an average of 62.22 seconds.

These sliding windows ramped up and down in length at the beginning and end of each transcript, respectively. In other words, each transcript’s first sliding window covered only its first line, the second sliding window covered the first two lines, and so on. This ensured that each line from the transcripts appeared in the same number ( $w$ ) of sliding windows. After performing various standard text preprocessing (e.g., normalizing case, lemmatizing, removing punctuation

649 and stop-words), we treated the text from each sliding window as a single “document,” and  
650 combined these documents across the two videos’ windows to create a single training corpus for  
651 the topic model.

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

659 We transformed each sliding window’s text into a topic vector, and then used linear interpo-  
660 lation (independently for each topic dimension) to resample the resulting timeseries to one vector  
661 per second. We also used the fitted model to obtain topic vectors for each question in our pool (see  
662 Supp. Tab. 1). Taken together, we obtained a *trajectory* for each video, describing its path through  
663 topic space, and a single coordinate for each question (Fig. 2C). Embedding both videos and all of  
664 the questions using a common model enables us to compare the content from different moments  
665 of videos, compare the content across videos, and estimate potential associations between specific  
666 questions and specific moments of video.

#### 667 **Estimating dynamic knowledge traces**

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

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

670 where

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

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

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

#### 685 **Creating knowledge and learning map visualizations**

686 An important feature of our approach is that, given a trained text embedding model and partic-  
687 ipants’ quiz performance on each question, we can estimate their knowledge about *any* content  
688 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-  
689 tions, or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 3, 4, 5, 6,  
690 and 7), we used Uniform Manifold Approximation and Projection (UMAP; 37, 38) to construct a  
691 2D projection of the text embedding space. Sampling the original 100-dimensional space at high  
692 resolution to obtain an adequate set of topic vectors spanning the embedding space would be  
693 computationally intractable. However, sampling a 2D grid is trivial.

694 At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing  
695 the cross-entropy between the pairwise (clustered) distances between the observations in their  
696 original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional  
697 embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise

698 distances in the original high-dimensional space were defined as 1 minus the correlation between  
699 each pair of coordinates, and pairwise distances in the low-dimensional embedding space were  
700 defined as the Euclidean distance between each pair of coordinates.

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

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

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

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

720 The  $\lambda$  term in the RBF equation controls the “smoothness” of the function, where larger values  
721 of  $\lambda$  result in smoother maps. In our implementation we used  $\lambda = 50$ . Next, we estimated the

722 “knowledge” at each coordinate,  $x$ , using:

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

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

726 Intuitively, learning maps reflect the *change* in knowledge across two maps.

## 727 **Author contributions**

728 Conceptualization: PCF, ACH, and JRM. Methodology: PCF, ACH, and JRM. Software: PCF.  
729 Validation: PCF. Formal analysis: PCF. Resources: PCF, ACH, and JRM. Data curation: PCF.  
730 Writing (original draft): JRM. Writing (review and editing): PCF, ACH, and JRM. Visualization:  
731 PCF and JRM. Supervision: JRM. Project administration: PCF. Funding acquisition: JRM.

## 732 **Data and code availability**

733 All of the data analyzed in this manuscript, along with all of the code for running our experiment  
734 and carrying out the analyses may be found at <https://github.com/ContextLab/efficient-learning-khan>.

## 736 **Acknowledgements**

737 We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of  
738 this study, and assistance with data collection efforts from Will Baxley, Max Bluestone, Daniel  
739 Carstensen, Kunal Jha, Caroline Lee, Lucy Owen, Xinming Xu, and Kirsten Ziman. Our work  
740 was supported in part by NSF CAREER Award Number 2145172 to JRM. The content is solely the  
741 responsibility of the authors and does not necessarily represent the official views of our supporting

742 organizations. The funders had no role in study design, data collection and analysis, decision to  
743 publish, or preparation of the manuscript.

## 744 References

- 745 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,  
746 56:149–178.
- 747 [2] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and  
748 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom  
749 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 750 [3] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text  
751 with the natural language toolkit*. Reilly Media, Inc.
- 752 [4] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-  
753 dren: distinguishing response flexibility from conceptual flexibility; the protracted development  
754 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 755 [5] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International  
756 Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing  
757 Machinery.
- 758 [6] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine  
759 Learning Research*, 3:993–1022.
- 760 [7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,  
761 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,  
762 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,  
763 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,  
764 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.

- 765 [8] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the  
766 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 767 [9] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-  
768 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal  
769 sentence encoder. *arXiv*, 1803.11175.
- 770 [10] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual  
771 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 772 [11] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).  
773 Evidence for a new conceptualization of semantic representation in the left and right cerebral  
774 hemispheres. *Cortex*, 40(3):467–478.
- 775 [12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).  
776 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,  
777 41(6):391–407.
- 778 [13] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 780 [14] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep  
781 bidirectional transformers for language understanding. *arXiv*, 1810.04805.
- 782 [15] Dikker, S., Wan, L., Davidescu, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,  
783 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony  
784 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 785 [16] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 786 [17] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of  
787 Experimental Psychology: General*, 115:155–174.
- 788 [18] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical  
789 Transactions of the Royal Society A*, 222(602):309–368.

- 790 [19] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.  
791     *School Science and Mathematics*, 100(6):310–318.
- 792 [20] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather  
793     prediction” task? individual variability in strategies for probabilistic category learning. *Learning  
794     and Memory*, 9:408–418.
- 795 [21] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of  
796     Cognition and Development*, 13(1):19–37.
- 797 [22] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual  
798     learning, pages 212–221. Sage Publications.
- 799 [23] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-  
800     ioral and neural signatures of transforming experiences into memories. *Nature Human Behavior*,  
801     5:905–919.
- 802 [24] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-  
803     peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–  
804     4008.
- 805 [25] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating  
806     empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 807 [26] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.  
808     Columbia University Press.
- 809 [27] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,  
810     326(7382):213–216.
- 811 [28] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).  
812     Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International  
813     Journal of Environmental Research and Public Health*, 18(5):2672.

- 814 [29] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 815 [30] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 816 [31] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
- 817 *The Chronicle of Higher Education*, 21:1–5.
- 818 [32] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic  
819 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,  
820 104:211–240.
- 821 [33] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic  
822 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.
- 823 [34] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of  
824 Educational Studies*, 53(2):129–147.
- 825 [35] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,  
826 *Handbook of Human Memory*. Oxford University Press.
- 827 [36] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum "memory wave"  
828 function? *Psychological Review*, 128(4):711–725.
- 829 [37] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and  
830 projection for dimension reduction. *arXiv*, 1802(03426).
- 831 [38] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold  
832 Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 833 [39] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of  
834 mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 835 [40] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,  
836 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to  
837 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.

- 838 [41] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-  
839 tations in vector space. *arXiv*, 1301.3781.
- 840 [42] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications  
841 from a national survey of language educators. *System*, 97:102431.
- 842 [43] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of  
843 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 844 [44] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).  
845 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*  
846 *Neuroscience*, 17(4):367–376.
- 847 [45] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 848 [46] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG  
849 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,  
850 7:43916.
- 851 [47] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*  
852 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 853 [48] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.  
854 *Biological Cybernetics*, 45(1):35–41.
- 855 [49] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in  
856 higher education: unmasking power and raising questions about the movement’s democratic  
857 potential. *Educational Theory*, 63(1):87–110.
- 858 [50] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter  
859 Student conceptions and conceptual learning in science. Routledge.
- 860 [51] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-  
861 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*  
862 *tion in Nursing*, 22:32–42.

- 863 [52] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching  
864 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 865 [53] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for  
866 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in  
867 Mathematics Education*, 35(5):305–329.
- 868 [54] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal  
869 Medicine*, 21:524–530.
- 870 [55] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,  
871 Goyal, N., Hambro, E., Azhar, F., Rodriguz, A., Joulin, A., Grave, E., and Lample, G. (2023).  
872 LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 873 [56] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and  
874 Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing  
875 Systems*.
- 876 [57] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned  
877 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 878 [58] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from  
879 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 880 [59] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is  
881 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior  
882 Research Methods*, 50:2597–2605.