

<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, 9, 11, 12, 57]  
46 [4, 9, 11, 12, 25, 57]? Or weaving a lecture's atomic elements (e.g., its component words) into a  
47 structured network that describes how those individual elements are related [35][35, 61]? Con-  
48 ceptual understanding could also involve building a mental model that transcends the mean-  
49 ings of those individual atomic elements by reflecting the deeper meaning underlying the gestalt  
50 whole [32, 36, 54][32, 36, 54, 60].

51 The difference between "understanding" and "memorizing," as framed by researchers in educa-  
52 tion, cognitive psychology, and cognitive neuroscience(e.g., 20, 23, 28, 36, 54)[e.g., 20, 23, 28, 36, 54]  
53 , has profound analogs in the fields of natural language processing and natural language under-  
54 standing. For example, considering the raw contents of a document (e.g., its constituent sym-  
55 bols, letters, and words) might provide some clues as to what the document is about, just as  
56 memorizing a passage might provide some ability to answer simple questions about it. How-  
57 ever, text embedding models(e.g., 5, 6, 8, 10, 13, 34, 44)[e.g., 5, 6, 8, 10, 13, 34, 44, 62] also attempt  
58 to capture the deeper meaning *underlying* those atomic elements. These models consider not  
59 only the co-occurrences of those elements within and across documents, but (in many cases)  
60 also patterns in how those elements appear across different scales (e.g., sentences, paragraphs,  
61 chapters, etc.), the temporal and grammatical properties of the elements, and other high-level  
62 characteristics of how they are used [37, 38]. According to these models, the deep concep-  
63 tual meaning of a document may be captured by a feature vector in a high-dimensional repre-  
64 sentation space, wherein nearby vectors reflect conceptually related documents. A model that  
65 succeeds at capturing an analogue of "understanding" is able to assign nearby feature vectors  
66 to two conceptually related documents, *even when the specific words contained in those documents*  
67 *have very little do not fully overlap*. *In this way, "concepts" are defined implicitly by the model's*  
68 *geometry*[e.g., how the embedding coordinate of a given word or document relates to the coordinates of other text em

69 :

70 Given these insights, what form might a representation of the sum total of a person's knowledge  
71 take (speculatively)? First, we might require a means of systematically describing or representing  
72 the nearly infinite set of possible things a person could know. Second, we might want to account

73 for potential associations between different concepts. For example, the concepts of “fish” and  
74 “water” might be associated in the sense that fish live in water. Third, knowledge may have  
75 a critical dependency structure, such that knowing about a particular concept might require first  
76 knowing about a set of other concepts. For example, understanding the concept of a fish swimming  
77 in water first requires understanding what fish and water *are*. Fourth, as we learn, our “current  
78 state of knowledge” should change accordingly. Learning new concepts should both update our  
79 characterizations of “what is known” and also unlock any now-satisfied dependencies of those  
80 newly learned concepts so that they are “tagged” as available for future learning.

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

89 Our primary research goal is to advance our understanding of what it means to acquire deep,  
90 real-world conceptual knowledge. Traditional laboratory approaches to studying learning and  
91 memory (e.g., list-learning studies) often draw little distinction between memorization and under-  
92 standing. Instead, these studies typically focus on whether information is effectively encoded or  
93 retrieved, rather than whether the information is *understood*. Approaches to studying conceptual  
94 learning, such as category learning experiments, can begin to investigate the distinction between  
95 memorization and understanding, often by training participants to distinguish arbitrary or random  
96 features in otherwise meaningless categorized stimuli [1, 17, 18, 21, 26, 52]. However the objective  
97 of real-world training, or learning from life experiences more generally, is often to develop new  
98 knowledge that may be applied in *useful* ways in the future. In this sense, the gap between modern  
99 learning theories and modern pedagogical approaches that inform classroom learning strategies is  
100 enormous: most of our theories about *how* people learn are inspired by experimental paradigms

101 and models that have only peripheral relevance to the kinds of learning that students and teachers  
102 actually seek [23, 36]. To help bridge this gap, our study uses course materials from real on-  
103 line courses to inform, fit, and test models of real-world conceptual learning. We also provide a  
104 demonstration of how our models can be used to construct “maps” of what students know, and  
105 how their knowledge changes with training. In addition to helping to visually capture knowledge  
106 (and changes in knowledge), we hope that such maps might lead to real-world tools for improving  
107 how we educate. Taken together, our work shows that existing course materials and evaluative  
108 tools like short multiple-choice quizzes may be leveraged to gain highly detailed insights into what  
109 students know and how they learn.

## 110 Results

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

125 We asked participants in our study to complete brief multiple-choice quizzes before, between,  
126 and after watching two lecture videos from the Khan Academy [31] platform (Fig. 1). The first



**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.

lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics: gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*, provided an overview of our current understanding of how stars form. We selected these particular lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad 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



**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 content area (Lecture 1, Lecture 2, and general physics knowledge) on each of the three quizzes.  
 146 Quiz 1 was intended to assess participants’ “baseline” knowledge before training, Quiz 2 assessed  
 147 knowledge after watching the *Four Fundamental Forces* video (i.e., Lecture 1), and Quiz 3 assessed  
 148 knowledge after watching the *Birth of Stars* video (i.e., Lecture 2).

149 To study in detail how participants’ conceptual knowledge changed over the course of the  
 150 experiment, we first sought to model the conceptual content presented to them at each moment  
 151 throughout each of the two lectures. We adapted an approach we developed in prior work [24]  
 152 to identify the latent themes in the lectures using a topic model [6]. Briefly, topic models take  
 153 as input a collection of text documents, and learn a set of “topics” (i.e., latent themes) from their  
 154 contents. Once fit, a topic model can be used to transform arbitrary (potentially new) documents  
 155 into sets of “topic proportions,” describing the weighted blend of learned topics reflected in their  
 156 texts. We parsed automatically generated transcripts of the two lectures into overlapping sliding  
 157 windows, where each window contained the text of the lecture transcript from a particular time  
 158 span. We treated the set of text snippets (across all of these windows) as documents to fit the  
 159 model (Fig. 2A; see *Constructing text embeddings of multiple lectures and questions*). Transforming the

160 text from every sliding window with the model yielded a number-of-windows by number-of-topics  
161 (15) topic-proportions matrix describing the unique mixture of broad themes from both lectures  
162 reflected in each window’s text. Each window’s “topic vector” (i.e., column of the topic-proportions  
163 matrix) is analogous to a coordinate in a 15-dimensional space whose axes are topics discovered  
164 by the model. Within this space, each lecture’s sequence of topic vectors (i.e., corresponding to its  
165 transcript’s overlapping text snippets across sliding windows) forms a *trajectory* that captures how  
166 its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution  
167 of one topic vector for each second of video (i.e., 1 Hz).

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

187 It is important to clarify that although we use topic model-derived embeddings to characterize



**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.

the conceptual content of the lectures and questions, we do not claim that the topic model itself understands the conceptual content of the lectures or questions. Rather, we view the topic model as a tool for capturing the *structure* of the conceptual content of the lectures and questions in a way that enables us to capture, quantify, and track and predict participants' knowledge.

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 (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “information” [19] 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 questions. Taken together, the analyses reported in Figure 3 and Supplementary Figure 2 indicate that a topic model fit to the videos' transcripts can also reveal correspondences (at a coarse scale) between the lectures and questions.

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

The ability to quantify how much each question is “asking about” the content from each moment of the lectures could enable high-resolution insights into participants’ knowledge. Traditional

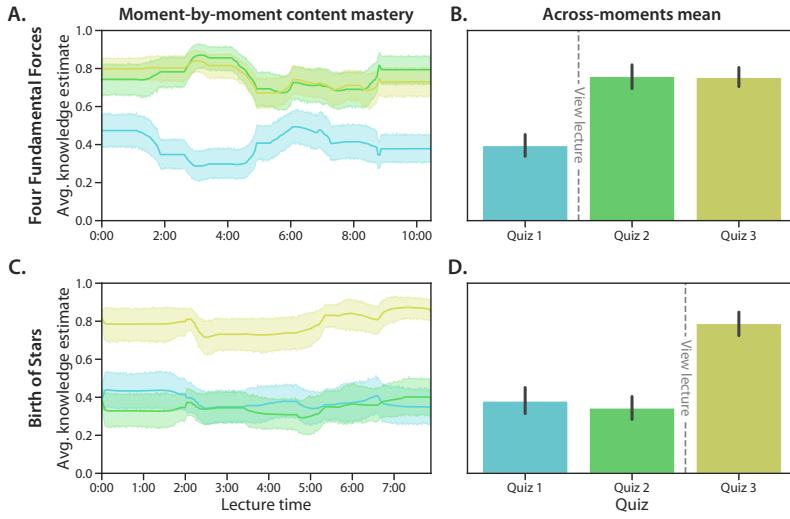


**Figure 4: Which parts of each lecture are captured by each question?** Each panel displays timeseries plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel 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.

approaches to estimating how much a student “knows” about the content of a given lecture entail computing the proportion of correctly answered questions. But if two students receive identical scores on an exam, might our 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).

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

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

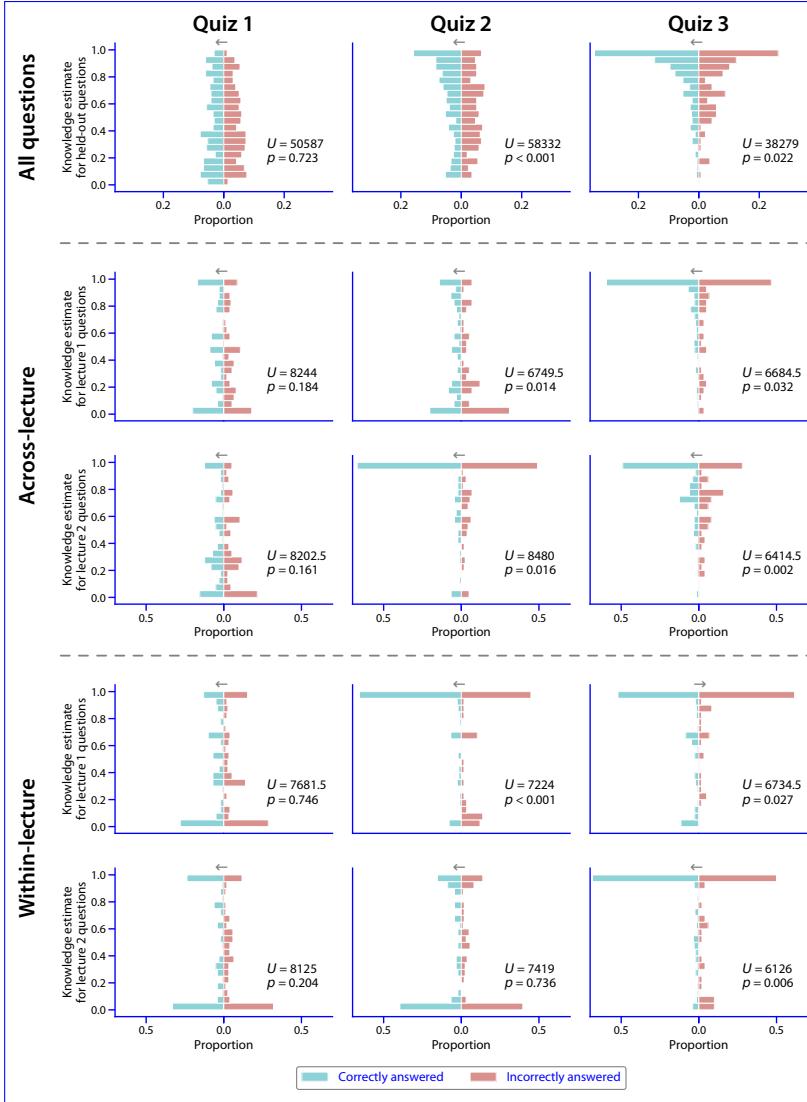


**Figure 5: Estimating moment-by-moment knowledge acquisition.** **A. Moment-by-moment knowledge about the Four Fundamental Forces** **Estimating knowledge about the content presented at each moment of each lecture.** **A. Knowledge about the time-varying content of Four Fundamental Forces.** Estimating dynamic knowledge traces. 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 the Four Fundamental Forces.** **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. Moment-by-moment knowledge about the Birth of Stars** **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 the Birth of Stars.** **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 subsequently confirmed) that participants should show greater estimated knowledge about the  
272 content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 5D). Specifically, since  
273 participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their  
274 knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a “boost” on  
275 Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge  
276 about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ( $t(49) = 1.013, p = 0.316$ ), but the  
277 estimated knowledge was substantially higher on Quiz 3 versus 2 ( $t(49) = 10.561, p < 0.001$ ) and  
278 Quiz 3 versus 1 ( $t(49) = 8.969, p < 0.001$ ).

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

289 For the initial quizzes participants took (prior to watching either lecture), participants’ esti-  
290 mated knowledge tended to be low overall, and relatively unstructured (Fig. 6, left ~~panel~~column).  
291 When we held out individual questions and estimated their knowledge at the held-out questions’  
292 embedding coordinates, we found no reliable differences in the estimates when the held-out ques-  
293 tion had been correctly versus incorrectly answered ( $\text{t}(633) = 0.577, p = 0.564$ ). ~~This~~ This “null”  
294 ~~effect persisted when we used all of the Quiz 1 questions from a given participant to predict a~~  
295 ~~held-out question (“All questions”;  $U = 50587, p = 0.723$ ), when we used questions from one lecture~~  
296 ~~to predict knowledge at the embedding coordinate of a held-out question about the other lecture~~  
297 ~~(“Across-lecture”; predicting knowledge for held-out *Four Fundamental Forces Questions* using *Birth*~~  
298 ~~of Stars~~ questions:  $U = 8244, p = 0.184$ ; predicting knowledge for held-out *Birth of Stars* questions:



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

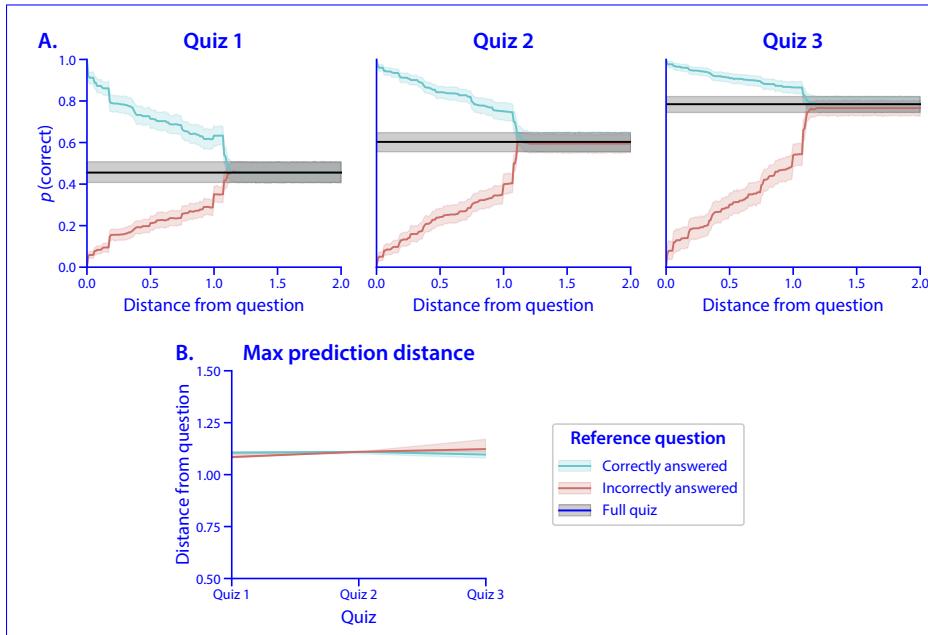
299  $U = 8202.5, p = 0.161$ ), and when we used questions from one lecture to predict knowledge at  
300 the embedding coordinate of a held-out question about the *same* lecture (“Within-lecture”; *Four*  
301 *Fundamental Forces*:  $U = 7681.5, p = 0.746$ ; *Birth of Stars*:  $U = 8125, p = 0.204$ ). We believe that this  
302 reflects a floor effect: when knowledge is low everywhere, there is little signal to differentiate  
303 between what is known versus unknown. **After watching the first lecture**

304 After watching *Four Fundamental Forces*, estimated knowledge for held-out correctly answered  
305 questions (from the second quiz; Fig. 6, middle **panel**) exhibited a positive shift relative to  
306 held-out incorrectly answered questions. This held when we included all questions in the analysis  
307 ( $U = 58332, p < 0.001$ ), when we predicted knowledge across-lectures ( $t(633) = 3.961, p < 0.001$ ).  
308 **This second quiz provides the maximally sensitive test for our knowledge predictions, since (if**  
309 **knowledge is estimated accurately) participants’ Quiz 2 responses should demonstrate specific**  
310 **knowledge about Lecture 1 content** *Four Fundamental Forces*:  $U = 6749.5, p = 0.014$ ; *Birth of Stars*:  
311  $U = 8480, p = 0.016$ ), and when we predicted knowledge at the embedding coordinates of held-out  
312 *Four Fundamental Forces* questions using other *Four Fundamental Forces* questions from the same quiz  
313 and participant ( $U = 7224, p < 0.001$ ). This difference did *not* hold for within-lecture predictions of  
314 *Birth of Stars* knowledge ( $U = 7419, p = 0.739$ ). Again, we suggest that this might reflect a floor effect  
315 whereby knowledge about the content of the *Birth of Stars* material is relatively low everywhere in  
316 that region of text embedding space.

317 Finally, after watching *Birth of Stars*, ~~but knowledge about Lecture 2 and general physics~~  
318 ~~concepts should be roughly unchanged from before they watched Lecture 1. After watching the~~  
319 ~~second lecture,~~ estimated knowledge ~~for held-out correctly answered questions~~ (from the third  
320 quiz; Fig. 6, right **panel**) ~~for all questions exhibited a positive shift. However, the estimated~~  
321 ~~knowledge for column~~ was higher ~~for~~ held-out ~~correctly answered questions remained greater~~  
322 ~~than that~~ ~~correctly answered questions than for held-out incorrectly answered questions. This~~  
323 ~~held when we included all questions in the analysis ( $U = 38279, p = 0.022$ ), when we carried out~~  
324 ~~across-lecture predictions (*Four Fundamental Forces*:  $U = 6684.5, p = 0.032$ ; *Birth of Stars*:  $U = 6414.5, p = 0.002$ ),~~  
325 ~~and and when we carried out within-lecture predictions of held-out *Birth of Stars* questions using~~  
326 ~~other *Birth of Stars* questions from the same quiz and participant ( $U = 6126, p = 0.006$ ).~~ However,

327 we found the *opposite* effect when we carried out within-lecture predictions of held-out *Four*  
328 *Fundamental Forces* questions using other *Four Fundamental Forces* questions from the same quiz and  
329 participant ( $U = 6734, p = 0.027$ ). Specifically, held-out correctly answered Quiz 3 questions about  
330 *Four Fundamental Forces* had reliably *lower* estimated knowledge than held-out incorrectly answered  
331 questions. Speculatively, we suggest that this may reflect participants forgetting some of the *Four*  
332 *Fundamental Forces* content. If this forgetting happens in a relatively “random” way (with respect  
333 to spatial distance within the text embedding space), then it could explain why some held-out  
334 questions about *Four Fundamental Forces* were answered incorrectly, even if questions at nearby  
335 coordinates (i.e., about similar content) were answered correctly. This might lead our approach  
336 to over-estimate knowledge for held-out *incorrectly* questions about “forgotten” knowledge that  
337 participants answered incorrectly.

338 That the knowledge estimates derived from the text embedding space reliably distinguish  
339 between held-out correctly versus incorrectly answered questions ( $t(628) = 2.045, p = 0.041$ ) Fig. 6)  
340 suggests that the text embedding space bears at least some relationship to participants’ knowledge.  
341 But what does that relationship look like as we move through the embedding space? For example,  
342 suppose we know that a participant answers a question (at embedding coordinate  $X$ ) correctly.  
343 As we move away from  $X$  in the embedding space, how does quiz performance “fall off” with  
344 distance? Or, suppose the participant instead answered that same question *incorrectly*. Again, as  
345 we move away from  $X$  in the embedding space, how do the chances that the participant does *not*  
346 know about the content change with distance? We reasoned that, assuming our space is capturing  
347 something about how participants actually organize their knowledge, conceptual knowledge right  
348 around  $X$  should be similar to the participant’s knowledge of the content at  $X$ . And at another  
349 extreme, at some distance (after moving sufficiently far away from  $X$ ), our guesses about what  
350 participants know (based on their response to the question at location  $X$ ) should be no better than  
351 guessing based on their overall proportion of correctly answered questions—i.e., if  $Y$  is very far  
352 away from  $X$ , all we can do with the participant’s response to  $X$  is guess that “their performance  
353 on quiz questions about  $Y$  is about equal to their average performance on quiz questions about  
354 any material.”

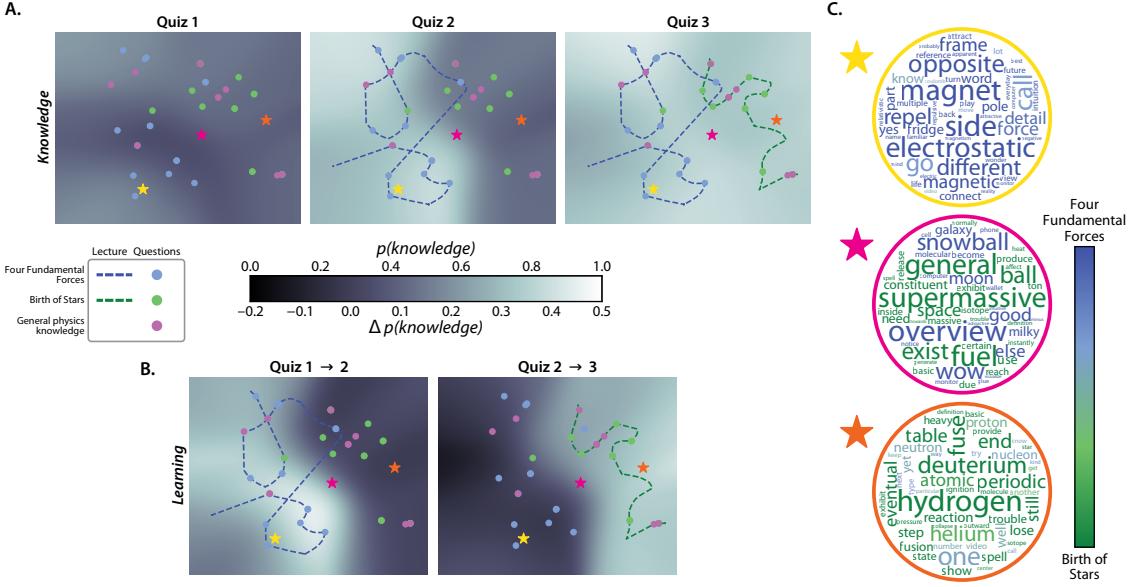


**Figure 7: Quiz performance falls off gradually in text embedding space. A. Performance versus distance.** For each participant, for each correctly answered question (blue) or incorrectly answered question (red), we computed the proportion of correctly answered questions within a given distance of that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

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

364 (and as characterized using our topic model), the rate at which knowledge changes with distance  
365 is relatively constant, even as participants' overall level of knowledge varies across quizzes or  
366 regions of the embedding space. This third contrast reflects a ceiling effect: when knowledge is  
367 relatively high everywhere, the signal differentiating what is known versus unknown is relatively  
368 weak. Taken together, this set of analyses demonstrates that our knowledge prediction framework  
369 is most informative when participants exhibit variability in their knowledge of the content captured  
370 by the text embedding model.

371 Knowledge estimates need not be limited to the content of the lectures. As illustrated in  
372 Figure 8, our general approach to estimating knowledge from a small number of quiz questions  
373 may be extended to *any* content, given its text embedding coordinate. To visualize how knowledge  
374 "spreads" through text embedding space to content beyond the lectures participants watched, we  
375 first fit a new topic model to the lectures' sliding windows with (up to)  $k = 100$  topics. Conceptually,  
376 increasing the number of topics used by the model functions to increase the "resolution" of the  
377 embedding space, providing a greater ability to estimate knowledge for content that is highly  
378 similar to (but not precisely the same as) that contained in the two lectures. This change in the  
379 number of topics overcame an undesirable behavior in the UMAP embedding procedure [40],  
380 whereby embedding coordinates for the 15-topic model tended to be "clumped" into separated  
381 clusters, rather than forming a smooth trajectory through the 2D space. When we increased the  
382 number of topics to 100, the embedding coordinates in the 2D space formed a smooth trajectory  
383 through the space, with substantially less clumping (Fig. 8). We note that we used these 2D maps  
384 solely for visualization; all relevant comparisons, distance computations, and statistical tests we  
385 report above were carried out in the original 15-dimensional space, using the 15-topic model. Aside  
386 from increasing the number of topics from 15 to 100, all other procedures and model parameters  
387 were carried over from the preceding analyses. As in our other analyses, we resampled each  
388 lecture's topic trajectory to 1 Hz and projected each question into a shared text embedding space.  
389 We projected the resulting 100-dimensional topic vectors (for each second of video and each quiz  
390 question) onto a shared 2-dimensional plane (see *Creating knowledge and learning map visualizations*).  
391 Next, we sampled points from a  $100 \times 100$  grid of coordinates that evenly tiled a rectangle enclos-



**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 the *Four Fundamental Forces* (blue) versus *Birth of Stars* (green) lectures, on average, across all timepoints’ topic vectors.

392 ing the 2D projections of the videos and questions. We used Equation 4 to estimate participants'  
393 knowledge at each of these 10,000 sampled locations, and averaged these estimates across par-  
394 ticipants to obtain an estimated average *knowledge map* (Fig. 8A). Intuitively, the knowledge map  
395 constructed from a given quiz's responses provides a visualization of how "much" participants  
396 knew about any content expressible by the fitted text embedding model at the point in time when  
397 they completed that quiz.

398 Several features of the resulting knowledge maps are worth noting. The average knowledge  
399 map estimated from Quiz 1 responses (Fig. 8A, leftmost map) shows that participants tended to  
400 have relatively little knowledge about any parts of the text embedding space (i.e., the shading is  
401 relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked  
402 increase in knowledge on the left side of the map (around roughly the same range of coordinates  
403 traversed by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words,  
404 participants' estimated increase in knowledge is localized to conceptual content that is nearby (i.e.,  
405 related to) the content from the lecture they watched prior to taking Quiz 2. This localization is  
406 non-trivial: these knowledge estimates are informed only by the embedded coordinates of the  
407 *quiz questions*, not by the embeddings of either lecture (see Eqn. 4). Finally, the knowledge map  
408 estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region  
409 surrounding the embedding of the *Birth of Stars* lecture participants watched immediately prior to  
410 taking Quiz 3.

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

417 Because the 2D projection we used to construct the knowledge and learning maps is invertible,  
418 we may gain additional insights into these maps' meaning-meanings by reconstructing the original  
419 high-dimensional topic vector for any location on the map we are interested in. For example, this

420 could serve as a useful tool for an instructor looking to better understand which content areas a  
421 student (or a group of students) knows well (or poorly). As a demonstration, we show the top-  
422 weighted words from the blends of topics reconstructed from three example locations on the maps  
423 (Fig. 8C): one point near the *Four Fundamental Forces* embedding (yellow), a second point near the  
424 *Birth of Stars* embedding (orange), and a third point between the two lectures' embeddings (pink).  
425 As shown in the word clouds in the panel, the top-weighted words at the example coordinate near  
426 the *Four Fundamental Forces* embedding tended to be weighted more heavily by the topics expressed  
427 in that lecture. Similarly, the top-weighted words at the example coordinate near the *Birth of Stars*  
428 embedding tended to be weighted more heavily by the topics expressed in *that* lecture. And the  
429 top-weighted words at the example coordinate between the two lectures' embeddings show a  
430 roughly even mix of words most strongly associated with each lecture.

## 431 Discussion

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

443 ~~Over the~~ We view our work as making several contributions to the study of how people  
444 acquire conceptual knowledge. First, from a methodological standpoint, our modeling framework  
445 provides a systematic means of mapping out and characterizing knowledge in maps that have

446 infinite (arbitrarily many) numbers of coordinates, and of “filling out” those maps using relatively  
447 small numbers of multiple choice quiz questions. Our experimental finding that we can use these  
448 maps to predict responses to held-out questions has several psychological implications as well. For  
449 example, concepts that are assigned to nearby coordinates by the text embedding model also appear  
450 to be “known to a similar extent” (as reflected by participants’ responses to held-out questions;  
451 Fig. 6). This suggests that participants also *conceptualize* similarly the content reflected by nearby  
452 embedding coordinates. The “spatial smoothness” of participants’ knowledge (as estimated using  
453 quiz performance) is being captured by the knowledge maps we are inferring from their quiz  
454 responses (e.g., Figs. 7, 8). In other words, our study shows that knowledge about a given concept  
455 implies knowledge about related concepts, and we also show how estimated knowledge falls off  
456 with distance in text embedding space.

457 In our study, we characterize the “coordinates” of participants’ knowledge using a relatively  
458 simple “bag of words” text embedding model [LDA; 6]. More sophisticated text embedding  
459 models, such as transformer-based models [15, 48, 59, 62] can learn complex grammatical and  
460 semantic relationships between words, higher-order syntactic structures, stylistic features, and  
461 more. We considered using transformer-based models in our study, but we found that the  
462 text embeddings derived from these models were surprisingly uninformative with respect to  
463 differentiating or otherwise characterizing the conceptual content of the lectures and questions  
464 we used. We suspect that this reflects a broader challenge in constructing models that are  
465 high-resolution within a given domain (e.g., the domain of physics lectures and questions) *and*  
466 sufficiently broad so as to enable them to cover a wide range of domains. For example, we found  
467 that the embeddings derived even from much larger and more modern models like BERT [15]  
468 , GPT [62], LLaMa [59], and others that are trained on enormous text corpora, end up yielding  
469 poor resolution within the content space spanned by individual course videos (Supp. Fig. 6).  
470 Whereas the LDA embeddings of the lectures and questions are “near” each other (i.e., the  
471 convex hull enclosing the two lectures’ trajectories is highly overlapping with the convex hull  
472 enclosing the questions’ embeddings), the BERT embeddings of the lectures and questions are  
473 instead largely distinct (top row of Supp. Fig. 6). 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. 6). We also examined per-question “content matches” between individual questions and individual moments of each lecture (Figs. 4, 6). 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. 6). Taken together, these comparisons illustrate how LDA (trained on the specific content in question) provides both coverage of the requisite material and specificity at the level of the content covered by individual questions. BERT, on the other hand, essentially assigns both lectures and all of the questions (which are all broadly about “physics”) into a tiny region of its embedding space, thereby blurring out meaningful distinctions

502 between different specific concepts covered by the lectures and questions. We note that these are  
503 not criticisms of BERT (or other large language models trained on large and diverse corpora).  
504 Rather, our point is that simple fine-tuned models trained on a relatively small but specialized  
505 corpus can outperform much more complicated models trained on much larger corpora, when we  
506 are specifically interested in capturing subtle conceptual differences at the level of a single course  
507 lecture or question. Of course if our goal had been to find a model that generalized to many  
508 different content areas, we would expect our approach to perform comparatively poorly relative to  
509 BERT or other much larger models. We suggest that bridging the tradeoff between high resolution  
510 within each content area versus the ability to generalize to many different content areas will be an  
511 important challenge for future work in this domain.

512 Another application for large language models that does *not* require explicitly modeling the  
513 content of individual lectures or questions is to leverage the models' ability to generate text. For  
514 example, generative text models like ChatGPT [48] and LLaMa [59] are already being used to build  
515 a new generation of interactive tutoring systems [e.g., 39]. Unlike the approach we have taken here,  
516 these generative text model-based systems do not explicitly model what learners know, or how  
517 their knowledge changes over time with training. One could imagine building a hybrid system  
518 that combines the best of both worlds: a large language model that can *generate* text, combined  
519 with a smaller model that can *infer* what learners know and how their knowledge changes over  
520 time. Such a hybrid system could potentially be used to build the next generation of interactive  
521 tutoring systems that are able to adapt to learners' needs in real time, and that are able to provide  
522 more nuanced feedback about what learners know and what they do not know.

523 At the opposite end of the spectrum from large language models, one could also imagine  
524 simplifying some aspects of our LDA-based approach by computing simple word overlap metrics.  
525 For example, the Jaccard similarity between text *A* and *B* is computed as the number of unique  
526 words in the intersection of words from *A* and *B* divided by the number of unique words in  
527 the union of words from *A* and *B*. In a supplemental analysis (Supp. Fig. 5), we compared the  
528 LDA-based question-lecture matches we reported in Figure 4 with the Jaccard similarities between  
529 each question and each sliding window of text from the corresponding lecture. As shown in

530 Supplementary Figure 5, this simple word-matching approach does not appear to capture the same  
531 level of specificity as the LDA-based approach. Whereas the LDA-based approach often yields a  
532 clear peak in the timeseries of correlations between each question and the corresponding lecture,  
533 the Jaccard similarity-based approach does not. Furthermore, these LDA-based matches appear  
534 to capture conceptual overlaps between the questions and lectures (Supp. Tab. 3), whereas simple  
535 word matching does not. For example, one of the example questions examined in Supplementary  
536 Figure 5 asks “Which of the following occurs as a cloud of atoms gets more dense?”. The LDA-based  
537 matches identify lecture timepoints where the relevant *topics* are discussed (e.g., when words like  
538 “cloud,” “atom,” “dense,” etc., are mentioned *together*). The Jaccard similarity-based matches,  
539 on the other hand, are strong when *any* of these words are mentioned, even if they do not occur  
540 together.

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

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

553 Within the past several years, the global pandemic ~~has~~ forced many educators to suddenly  
554 adapt to teaching remotely [30, 45, 56, 63]. This change in world circumstances is happening  
555 alongside (and perhaps accelerating) geometric growth in the availability of high-quality online  
556 courses from platforms such as Khan Academy [31], Coursera [64], EdX [33], and others [53].  
557 Continued expansion of the global internet backbone and improvements in computing hardware

558 have also facilitated improvements in video streaming, enabling videos to be easily shared and  
559 viewed by increasingly large segments of the world’s population. This exciting time for online  
560 course instruction provides an opportunity to re-evaluate how we, as a global community, educate  
561 ourselves and each other. For example, we can ask: what defines an effective course or training  
562 program? Which aspects of teaching might be optimized and/or augmented by automated tools?  
563 How and why do learning needs and goals vary across people? How might we lower barriers of  
564 access to a high-quality education?

565 Alongside these questions, there is a growing desire to extend existing theories beyond the  
566 domain of lab testing rooms and into real classrooms [29]. In part, this has led to a recent resur-  
567 gence of “naturalistic” or “observational” experimental paradigms that attempt to better reflect  
568 more ethologically valid phenomena that are more directly relevant to real-world situations and  
569 behaviors [46]. In turn, this has brought new challenges in data analysis and interpretation. A key  
570 step towards solving these challenges will be to build explicit models of real-world scenarios and  
571 how people behave in them (e.g., models of how people learn conceptual content from real-world  
572 courses, as in our current study). A second key step will be to understand which sorts of signals  
573 derived from behaviors and/or other measurements(e.g., [neurophysiological data; 2, 16, 43, 47, 50](#))  
574 [\[e.g., neurophysiological data; 2, 16, 43, 47, 50\]](#) might help to inform these models. A third major  
575 step will be to develop and employ reliable ways of evaluating the complex models and data that  
576 are a hallmark of naturalistic paradigms.

577 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also  
578 relate to the notion of “theory of mind” of other individuals [22, 27, 42]. Considering others’ unique  
579 perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and  
580 communicate [51, 55, 58]. One could imagine future extensions of our work (e.g., analogous to  
581 the knowledge and learning maps shown in Fig. 8), that attempt to characterize how well-aligned  
582 different people’s knowledge bases or backgrounds are. In turn, this might be used to model how  
583 knowledge (or other forms of communicable information) flows not just between teachers and  
584 students, but between friends having a conversation, individuals on a first date, participants at  
585 a business meeting, doctors and patients, experts and non-experts, political allies or adversaries,

586 and more. For example, the extent to which two people's knowledge maps "match" or "align" in  
587 a given region of text embedding space might serve as a predictor of how effectively they will be  
588 able to communicate about the corresponding conceptual content.

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

## 595 Materials and methods

### 596 Participants

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

602 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09  
603 years). A total of 15 participants reported their gender as male and 35 participants reported their  
604 gender as female. A total of 49 participants reported their native language as "English" and 1  
605 reported having another native language. A total of 47 participants reported their ethnicity as  
606 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants  
607 reported their races as White (32 participants), Asian (14 participants), Black or African American  
608 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other  
609 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

610 A total of 49 participants reporting having normal hearing and 1 participant reported having

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

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

622 Participants reported their undergraduate major(s) as “social sciences” (28 participants), “nat-  
623 ural sciences” (16 participants), “professional” (e.g., pre-med or pre-law; 8 participants), “mathe-  
624 matics and engineering” (7 participants), “humanities” (4 participants), or “undecided” (3 partici-  
625 pants). Note that some participants selected multiple categories for their undergraduate major(s).  
626 We also asked participants about the courses they had taken. In total, 45 participants reported hav-  
627 ing taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan  
628 Academy courses. Of those who reported having watched at least one Khan Academy course,  
629 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8  
630 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We  
631 also asked participants about the specific courses they had watched, categorized under different  
632 subject areas. In the “Mathematics” area, participants reported having watched videos on AP  
633 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-  
634 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry  
635 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential  
636 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),  
637 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other  
638 videos not listed in our survey (5 participants). In the “Science and engineering” area, participants

639 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 participants);  
640 Physics, AP Physics I, or AP Physics II (18 participants); Biology, AP Biology; or High  
641 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed  
642 in our survey (5 participants). We also asked participants whether they had specifically seen the  
643 videos used in our experiment. Of the 45 participants who reported having taken at least  
644 one Khan Academy course in the past, 44 participants reported that they had not watched the *Four*  
645 *Fundamental Forces* video, and 1 participant reported that they were not sure whether they had  
646 watched it. All participants reported that they had not watched the *Birth of Stars* video. When  
647 we asked participants about non-Khan Academy online courses, they reported having watched  
648 or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test  
649 preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 participants),  
650 Computing (2 participants), and other categories not listed in our survey (17 participants).  
651 Finally, we asked participants about in-person courses they had taken in different subject areas.  
652 They reported taking courses in Mathematics (38 participants), Science and engineering (37 participants),  
653 Arts and humanities (34 participants), Test preparation (27 participants), Economics  
654 and finance (26 participants), Computing (14 participants), College and careers (7 participants), or  
655 other courses not listed in our survey (6 participants).

## 656 Experiment

657 We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces*  
658 (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force;  
659 duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed;  
660 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  
661 of viewing order, nor have we are aware of how or why viewing order might influence our main  
662 findings, we acknowledge that we did not control for potential order effects in our study.

664 We then hand-created 39 multiple-choice questions: 15 about the conceptual content of *Four*  
665 *Fundamental Forces* (i.e., Lecture 1), 15 about the conceptual content of *Birth of Stars* (i.e., Lecture 2),

666 and 9 questions that tested for general conceptual knowledge about basic physics (covering material  
667 that was not presented in either video). ~~The One of our group's undergraduate research assistants~~  
668 ~~worked alongside a rotating Masters student to develop this set of questions (these researchers~~  
669 ~~are acknowledged in our paper for their contribution, although they did not meet the criteria for~~  
670 ~~authorship discussed with all team members at the start of the project, as determined by J.R.M.) The~~  
671 ~~senior author (J.R.M.) tasked the pair of researchers with coming up with "15 conceptual questions~~  
672 ~~about each lecture, along with 9 additional questions about general physics knowledge."~~ To  
673 ~~help broaden the set of lecture-specific questions, the researchers were further instructed to work~~  
674 ~~through each lecture in small segments, identify what each segment was "about" conceptually,~~  
675 ~~and then write a question about that concept. The general physics questions were drawn from the~~  
676 ~~researchers' coursework along with internet searches and brainstorming with the project team and~~  
677 ~~other members of J.R.M.'s lab. The final set of questions (and response options) was reviewed and~~  
678 ~~approved by J.R.M. before we collected or analyzed the text or experimental data.~~

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

683 Over the course of the experiment, participants completed three 13-question multiple-choice  
684 quizzes: the first before viewing Lecture 1, the second between Lectures 1 and 2, and the third after  
685 viewing Lecture 2 (see Fig. 1). The questions appearing on each quiz, for each participant, were  
686 randomly chosen from the full set of 39, with the constraints that (a) each quiz ~~contain~~ contained  
687 exactly 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general  
688 physics knowledge, and (b) each question appear exactly once for each participant. The orders of  
689 questions on each quiz, and the orders of answer options for each question, were also randomized.  
690 Our experimental protocol was approved by the Committee for the Protection of Human Subjects  
691 at Dartmouth College. We used this experiment to develop and test our computational framework  
692 for estimating knowledge and learning.

693 **Analysis**

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

695 We adapted an approach we developed in prior work [24] to embed each moment of the two  
696 lectures and each question in our pool in a common representational space. Briefly, our approach  
697 uses a topic model([Latent Dirichlet Allocation; 6](#)), [[Latent Dirichlet Allocation; 6](#)] trained on a set  
698 of documents, to discover a set of (up to)  $k$  “topics” or “themes.” Formally, each topic is defined  
699 as a distribution of weights over [each word](#) [words](#) in the model’s vocabulary (i.e., the union of  
700 all unique words, across all documents, excluding “stop words.”). Conceptually, each topic is  
701 intended to give larger weights to words that are semantically related [or tend to co-occur](#), [as](#)  
702 [implied by their co-occurring](#) in the same documents. After fitting a topic model, each document  
703 in the training set, or any *new* document that contains at least some of the words in the model’s  
704 vocabulary, may be represented as a  $k$ -dimensional vector describing how much the document  
705 (most probably) reflects each topic. To select an appropriate  $k$  for our model, we identified the  
706 minimum number of topics that yielded at least one “unused” topic (i.e., in which all words in  
707 the vocabulary were assigned uniform weights) after training. This indicated that the number of  
708 topics was sufficient to capture the set of latent themes present in the two lectures (from which we  
709 constructed our document corpus, as described below). We found this value to be  $k = 15$  topics.  
710 The distribution of weights over words in the vocabulary for each discovered topic is shown in  
711 Supplementary Figure 1, and each topic’s top-weighted words may be found in Supplementary  
712 Table 2.

713 As illustrated in Figure 2A, we start by building up a corpus of documents using overlap-  
714 ping sliding windows that span each video’s transcript. Khan Academy provides professionally  
715 created, manual transcriptions of all videos for closed captioning. However, such transcripts  
716 would not be readily available in all contexts to which our framework could potentially be ap-  
717 plied. Khan Academy videos are hosted on the YouTube platform, which additionally provides  
718 automated captions. We opted to use these automated transcripts([which, in prior work, we](#)  
719 [have found to be of sufficiently near-human quality to yield reliable data in behavioral studies; 65](#))

720 [which, in prior work, we have found to be of sufficiently near-human quality to yield reliable data in behavioral stud  
721 when developing our framework in order to make it more directly extensible and adaptable by  
722 others in the future.

723 We fetched these automated transcripts using the youtube-transcript-api Python pack-  
724 age [14]. The transcripts consisted of one timestamped line of text for every few seconds (mean:  
725 2.34 s; standard deviation: 0.83 s) of spoken content in the video (i.e., corresponding to each indi-  
726 vidual caption that would appear on-screen if viewing the lecture via YouTube, and when those  
727 lines would appear). We defined a sliding window length of (up to)  $w = 30$  transcript lines, and  
728 assigned each window a timestamp corresponding to the midpoint between the timestamps for its  
729 first and last lines. This  $w$  parameter was chosen to match the same number of words per sliding  
730 window (rounded to the nearest whole word, and before preprocessing) as the sliding windows  
731 we defined in our prior work [24] (i.e., 185 words per sliding window).

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

745 After fitting a topic model to the two videos' transcripts, we could use the trained model to  
746 transform arbitrary (potentially new) documents into  $k$ -dimensional topic vectors. A convenient  
747 property of these topic vectors is that documents that reflect similar blends of topics (i.e., documents

748 that reflect similar themes, according to the model) will yield similar coordinates (in terms of  
 749 correlation, cosine similarity, Kullback-Leibler divergence, Euclidean distance, or other geometric  
 750 measures). In general, the similarity between different documents' topic vectors may be used to  
 751 characterize the similarity in conceptual content between the documents.

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

## 760 **Estimating dynamic knowledge traces**

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

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

763 where

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

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

770 Intuitively,  $\text{ncorr}(x, y)$  is the correlation between two topic vectors (e.g., the topic vector from one

771 timepoint in a lecture,  $x$ , and the topic vector for one question,  $y$ ), normalized by the minimum and  
772 maximum correlations (across all timepoints  $t$  and questions  $Q$ ) to range between 0 and 1, inclusive.  
773 Equation 1 then computes the weighted average proportion of correctly answered questions about  
774 the content presented at timepoint  $t$ , where the weights are given by the normalized correlations  
775 between timepoint  $t$ 's topic vector and the topic vectors for each question. The normalization step  
776 (i.e., using `ncorr` instead of the raw correlations) ~~insures~~ ensures that every question contributes  
777 some non-negative amount to the knowledge estimate.

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

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

784 In our bootstrap procedure, we ran 10,000 iterations to estimate the relationship between  
785 participants' performance and the distance to a given reference question. For each of these  
786 iterations, for every individual quiz ( $q$ ), we first determined the across-participants average  
787 “simple” proportion correct and its 95% confidence interval. This interval was established by  
788 repeatedly (1,000 times) subsampling participants with replacement, computing the mean “simple”  
789 proportion correct for each subsample, and then deriving the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles from the  
790 distribution of these subsample means. We used this interval as our benchmark for determining  
791 whether the proportion of correctly answered questions for a given subset of questions was reliably  
792 different (at the  $p < 0.05$  significance level) from the average proportion correct across all questions.

793

794 Next, for each participant, we examined all 15 questions they answered on quiz  $q$ . We treated  
795 each question as the “reference question” in turn. Around this reference, we constructed a series of  
796 15-dimensional spheres (starting with a radius of 0), where each successive sphere had a radius of  
797 0.01 (correlation distance) greater than its predecessor. Within each of these spheres, we calculated

798 the proportion of questions answered correctly by the participant. This yielded two distinct sets  
799 of proportion-correct values for each binned distance (radius) for a specific participant and quiz:  
800 one set of values where the reference questions had been answered correctly, and another set  
801 where the reference questions had been answered incorrectly. From these, we established the  
802 average proportion correct within each radius for both categories of reference questions. Finally,  
803 we identified the minimum binned distance from the correctly answered reference questions for  
804 which the average proportion correct intersected the 95% confidence interval of the simple average  
805 proportion correct computed earlier. We display the resulting distance estimates, for each quiz  
806 and reference question status, in Figure 7B.

807 **Creating knowledge and learning map visualizations**

808 An important feature of our approach is that, given a trained text embedding model and partici-  
809 pants' quiz performance on each question, we can estimate their knowledge about *any* content ex-  
810 pressible by the embedding model—not solely the content explicitly probed by the quiz questions,  
811 or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 7, 8, 9, 10, and 11),  
812 we used Uniform Manifold Approximation and Projection([UMAP; 40, 41](#)) [[UMAP; 40, 41](#)] to con-  
813 struct a 2D projection of the text embedding space. Sampling the original 100-dimensional space  
814 at high resolution to obtain an adequate set of topic vectors spanning the embedding space would  
815 be computationally intractable. However, sampling a 2D grid is trivial.

816 At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing  
817 the cross-entropy between the pairwise (clustered) distances between the observations in their  
818 original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional  
819 embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise  
820 distances in the original high-dimensional space were defined as 1 minus the correlation between  
821 each pair of coordinates, and pairwise distances in the low-dimensional embedding space were  
822 defined as the Euclidean distance between each pair of coordinates.

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

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

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

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

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

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

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

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

849 **Author contributions**

850 Conceptualization: PCF, ACH, and JRM. Methodology: PCF, ACH, and JRM. Software: PCF.  
851 Validation: PCF. Formal analysis: PCF. Resources: PCF, ACH, and JRM. Data curation: PCF.  
852 Writing (original draft): JRM. Writing (review and editing): PCF, ACH, and JRM. Visualization:  
853 PCF and JRM. Supervision: JRM. Project administration: PCF. Funding acquisition: JRM.

854 **Data and code availability**

855 All of the data analyzed in this manuscript, along with all of the code for running our experiment  
856 and carrying out the analyses may be found at [https://github.com/ContextLab/efficient-learning-](https://github.com/ContextLab/efficient-learning-khan)  
857 [khan](https://github.com/ContextLab/efficient-learning-khan).

858 **Acknowledgements**

859 We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of  
860 this study, and assistance with data collection efforts from Will Baxley, Max Bluestone, Daniel  
861 Carstensen, Kunal Jha, Caroline Lee, Lucy Owen, Xinming Xu, and Kirsten Ziman. Our work was  
862 supported in part by NSF CAREER Award Number 2145172 to [JRM](#). The content is solely the  
863 responsibility of the authors and does not necessarily represent the official views of our supporting  
864 organizations. The funders had no role in study design, data collection and analysis, decision to  
865 publish, or preparation of the manuscript.

866 **References**

- 867 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,  
868 56:149–178.
- 869 [2] Bevilacque, D., Davidescu, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and

- 870 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom  
871 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 872 [3] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text*  
873 *with the natural language toolkit*. Reilly Media, Inc.
- 874 [4] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-  
875 dren: distinguishing response flexibility from conceptual flexibility; the protracted development  
876 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 877 [5] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*  
878 *Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing  
879 Machinery.
- 880 [6] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*  
881 *Learning Research*, 3:993–1022.
- 882 [7] Boyd-Graber, J. and Mimno, D. (2014). Care and feeding of topic models: problems, diagnostics,  
883 and improvements. In Airolidi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E., editors,  
884 *Handbook of Mixed Membership Models and Their Applications*. CRC Press.
- 885 [8] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,  
886 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,  
887 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,  
888 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,  
889 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 890 [9] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the  
891 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 892 [10] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-  
893 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal  
894 sentence encoder. *arXiv*, 1803.11175.

- 895 [11] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual  
896 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 897 [12] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).  
898 Evidence for a new conceptualization of semantic representation in the left and right cerebral  
899 hemispheres. *Cortex*, 40(3):467–478.
- 900 [13] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).  
901 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,  
902 41(6):391–407.
- 903 [14] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 904
- 905 [15] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep  
906 bidirectional transformers for language understanding. *arXiv*, 1810.04805.
- 907 [16] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,  
908 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony  
909 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 910 [17] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 911 [18] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of*  
912 *Experimental Psychology: General*, 115:155–174.
- 913 [19] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*  
914 *Transactions of the Royal Society A*, 222(602):309–368.
- 915 [20] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.  
916 *School Science and Mathematics*, 100(6):310–318.
- 917 [21] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather  
918 prediction” task? individual variability in strategies for probabilistic category learning. *Learning*  
919 *and Memory*, 9:408–418.

- 920 [22] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*  
921 *Cognition and Development*, 13(1):19–37.
- 922 [23] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual  
923 learning, pages 212–221. Sage Publications.
- 924 [24] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-  
925 ioral and neural signatures of transforming experiences into memories. *Nature Human Behavior*,  
926 5:905–919.
- 927 [25] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge au-  
928 tomatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*,  
929 9:doi.org/10.3389/fpsyg.2018.00133.
- 930 [26] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-  
931 peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–  
932 4008.
- 933 [27] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating  
934 empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 935 [28] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.  
936 Columbia University Press.
- 937 [29] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,  
938 326(7382):213–216.
- 939 [30] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).  
940 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International*  
941 *Journal of Environmental Research and Public Health*, 18(5):2672.
- 942 [31] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 943 [32] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.

- 944 [33] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
- 945       *The Chronicle of Higher Education*, 21:1–5.
- 946 [34] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic  
947       analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,  
948       104:211–240.
- 949 [35] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic  
950       events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.
- 951 [36] MacLellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of  
952       Educational Studies*, 53(2):129–147.
- 953 [37] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,  
954       *Handbook of Human Memory*. Oxford University Press.
- 955 [38] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”  
956       function? *Psychological Review*, 128(4):711–725.
- 957 [39] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension  
958       for adding LLM-driven chatbots to interactive notebooks. [https://github.com/ContextLab/  
959       chatify](https://github.com/ContextLab/chatify).
- 960 [40] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and  
961       projection for dimension reduction. *arXiv*, 1802(03426).
- 962 [41] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold  
963       Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 964 [42] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of  
965       mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 966 [43] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,  
967       U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to  
968       computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.

- 969 [44] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-  
970 tations in vector space. *arXiv*, 1301.3781.
- 971 [45] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications  
972 from a national survey of language educators. *System*, 97:102431.
- 973 [46] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of  
974 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 975 [47] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).  
976 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*  
977 *Neuroscience*, 17(4):367–376.
- 978 [48] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 979 [49] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models.  
980 *arXiv*, 2208.02957.
- 981 [50] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG  
982 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,  
983 7:43916.
- 984 [51] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*  
985 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 986 [52] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.  
987 *Biological Cybernetics*, 45(1):35–41.
- 988 [53] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in  
989 higher education: unmasking power and raising questions about the movement’s democratic  
990 potential. *Educational Theory*, 63(1):87–110.
- 991 [54] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter  
992 Student conceptions and conceptual learning in science. Routledge.

- 993 [55] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-  
994 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*  
995 *tion in Nursing*, 22:32–42.
- 996 [56] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching  
997 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 998 [57] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for  
999 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*  
1000 *Mathematics Education*, 35(5):305–329.
- 1001 [58] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal*  
1002 *Medicine*, 21:524–530.
- 1003 [59] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,  
1004 Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023).  
1005 LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 1006 [60] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Pio-  
1007 ntkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust  
1008 detection of AI-generated texts. *arXiv*, 2306.04723.
- 1009 [61] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold?  
1010 distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of*  
1011 *the Cognitive Science Society*, 43(43).
- 1012 [62] Viswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and  
1013 Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing*  
1014 *Systems*.
- 1015 [63] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned  
1016 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.

- 1017 [64] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from  
1018 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 1019 [65] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is  
1020 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*  
1021 *Research Methods*, 50:2597–2605.