

- 1 Text embedding models yield high-resolution insights
- 2 into conceptual knowledge from short multiple-choice

3 quizzes

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

## Abstract

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

**Keywords:** education, learning, knowledge, concepts, natural language processing

<sup>19</sup> **Introduction**

<sup>20</sup> Suppose that a teacher had access to a complete, tangible “map” of everything a student knows.  
<sup>21</sup> Defining what such a map might even look like, let alone how it might be constructed or filled in, is  
<sup>22</sup> itself a non-trivial problem. But if a teacher *were* to gain access to such a map, how might it change  
<sup>23</sup> their ability to teach that student? Perhaps they might start by checking how well the student  
<sup>24</sup> knows the to-be-learned information already, or how much they know about related concepts.  
<sup>25</sup> For some students, they could potentially optimize their teaching efforts to maximize efficiency  
<sup>26</sup> by focusing primarily on not-yet-known content. For other students (or other content areas), it  
<sup>27</sup> might be more effective to optimize for direct connections between already known content and  
<sup>28</sup> new material. Observing how the student’s knowledge changed over time, in response to their  
<sup>29</sup> teaching, could also help to guide the teacher towards the most effective strategy for that individual  
<sup>30</sup> student.

<sup>31</sup> A common approach to assessing a student’s knowledge is to present them with a set of quiz  
<sup>32</sup> questions, calculate the proportion they answer correctly, and provide them with feedback in the  
<sup>33</sup> form of a simple numeric or letter grade. While such a grade can provide *some* indication of whether  
<sup>34</sup> the student has mastered the to-be-learned material, any univariate measure of performance on a  
<sup>35</sup> complex task sacrifices certain relevant information, risks conflating underlying factors, and so on.  
<sup>36</sup> For example, consider the relative utility of the theoretical map described above that characterizes  
<sup>37</sup> a student’s knowledge in detail, versus a single annotation saying that the student answered 85%  
<sup>38</sup> of their quiz questions correctly, or that they received a ‘B’. Here ~~we~~ show that the same quiz  
<sup>39</sup> data required to compute proportion-correct scores or letter grades can instead be used to obtain  
<sup>40</sup> far more detailed insights into what a student knew at the time they took the quiz.

<sup>41</sup> Designing and building procedures and tools for mapping out knowledge touches on deep  
<sup>42</sup> questions about what it means to learn. For example, how do we acquire conceptual knowledge?  
<sup>43</sup> Memorizing course lectures or textbook chapters by rote can lead to the superficial *appearance*  
<sup>44</sup> of understanding the underlying content, but achieving true conceptual understanding seems to  
<sup>45</sup> require something deeper and richer. Does conceptual understanding entail connecting newly

46 acquired information to the scaffolding of one’s existing knowledge or experience [6, 11, 13, 15, 31,  
47 66]? Or weaving a lecture’s atomic elements (e.g., its component words) into a structured network  
48 that describes how those individual elements are related [41, 71]? Conceptual understanding  
49 could also involve building a mental model that transcends the meanings of those individual  
50 atomic elements by reflecting the deeper meaning underlying the gestalt whole [38, 42, 63, 70].

51 The difference between “understanding” and “memorizing,” as framed by researchers in ed-  
52 ucation, cognitive psychology, and cognitive neuroscience [e.g., 24, 29, 34, 42, 63], has profound  
53 analogs in the fields of natural language processing and natural language understanding. For  
54 example, considering the raw contents of a document (e.g., its constituent symbols, letters, and  
55 words) might provide some clues as to what the document is about, just as memorizing a passage  
56 might provide some ability to answer simple questions about it. However, text embedding mod-  
57 els [e.g., 7, 8, 10, 12, 16, 40, 52, 72] also attempt to capture the deeper meaning *underlying* those  
58 atomic elements. These models consider not only the co-occurrences of those elements within  
59 and across documents, but (in many cases) also patterns in how those elements appear across  
60 different scales (e.g., sentences, paragraphs, chapters, etc.), ~~the~~their temporal and grammatical  
61 properties~~of the elements~~, and other high-level characteristics of how they are used [43, 44]. To  
62 be clear, this is not to say that text embedding models themselves are capable of “understanding”  
63 deep conceptual meaning in any traditional sense. But rather, their ability to capture the under-  
64 lying *structure* of text documents beyond their surface-level contents provides a computational  
65 framework through which those documents’ deeper conceptual meanings may be quantified, ex-  
66 plored, and understood. According to these models, the deep conceptual meaning of a document  
67 may be captured by a feature vector in a high-dimensional representation space, wherein nearby  
68 vectors reflect conceptually related documents. A model that succeeds at capturing an analogue  
69 of “understanding” is able to assign nearby feature vectors to two conceptually related documents  
70 ~~r~~*even when the specific words contained in those documents have limited overlap*. In this way, “concepts”  
71 are defined implicitly by the model’s geometry [e.g., how the embedding coordinate of a given  
72 word or document relates to the coordinates of other text embeddings; 57].

73 Given these insights, what form might a representation of the sum total of a person’s knowledge

74 take? First, we might require a means of systematically describing or representing (at least some  
75 subset of) the nearly infinite set of possible things a person could know. Second, we might want to  
76 account for potential associations between different concepts. For example, the concepts of “fish”  
77 and “water” might be associated in the sense that fish live in water. Third, knowledge may have  
78 a critical dependency structure, such that knowing about a particular concept might require first  
79 knowing about a set of other concepts. For example, understanding the concept of a fish swimming  
80 in water first requires understanding what fish and water *are*. Fourth, as we learn, our “current  
81 state of knowledge” should change accordingly. Learning new concepts should both update our  
82 characterizations of “what is known” and also unlock any now-satisfied dependencies of those  
83 newly learned concepts so that they are “tagged” as available for future learning.

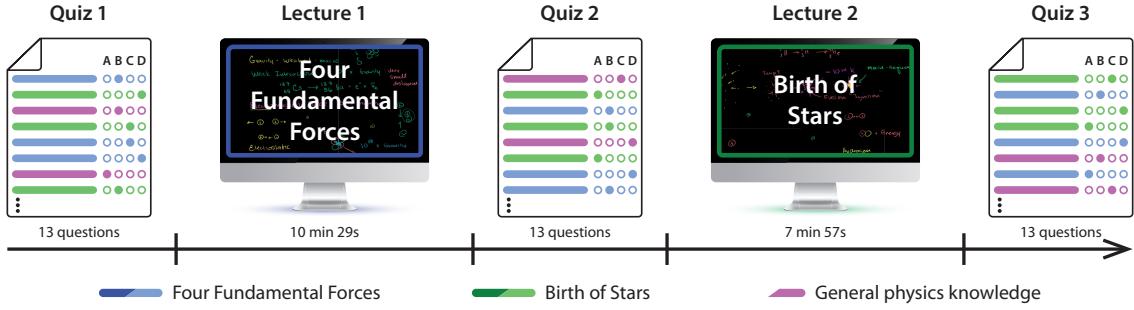
84 Here we develop a framework for modeling how conceptual knowledge is acquired during  
85 learning. The central idea behind our framework is to use text embedding models to define the  
86 coordinate systems of two maps: a *knowledge map* that describes the extent to which each concept is  
87 currently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

92 Our primary research goal is to advance our understanding of what it means to acquire deep,  
93 real-world conceptual knowledge. Traditional laboratory approaches to studying learning and  
94 memory (e.g., list-learning studies) often draw little distinction between memorization and under-  
95 standing. Instead, these studies typically focus on whether information is effectively encoded or  
96 retrieved, rather than whether the information is *understood*. Approaches to studying conceptual  
97 learning, such as category learning experiments, can begin to investigate the distinction between  
98 memorization and understanding, often by training participants to distinguish arbitrary or random  
99 features in otherwise meaningless categorized stimuli [1, 20, 21, 25, 32, 60]. However, the objective  
100 of real-world training, or learning from life experiences more generally, is often to develop new  
101 knowledge that may be applied in *useful* ways in the future. In this sense, the gap between modern

learning theories and modern pedagogical approaches that inform classroom learning strategies is enormous: most of our theories about *how* people learn are inspired by experimental paradigms and models that have only peripheral relevance to the kinds of learning that students and teachers actually seek [29, 42]. To help bridge this gap, our study uses course materials from real online courses to inform, fit, and test models of real-world conceptual learning. We show that these models recover meaningful relationships between concepts presented during course lectures and tested by assessments, and that these relationships can be leveraged to predict students' success on individual quiz questions. We also provide a demonstration of how our models can be used to construct “maps” of what students know, and how their knowledge changes with training. In addition to helping to visually capture knowledge (and changes in knowledge), we hope that such maps might lead to real-world tools for improving how we educate. Taken together, our work shows that existing course materials and evaluative tools like short multiple-choice quizzes may be leveraged to gain highly detailed insights into what students know and how they learn.

## Results

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



**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 appearing on each quiz, and the orders of each quiz's questions, were randomized across participants.

128 that the conceptual content of interest be discovered *automatically*, rather than relying on manually  
 129 produced ratings or labels.

130 We asked participants in our study to complete brief multiple-choice quizzes before, between,  
 131 and after watching two lecture videos from the Khan Academy [37] platform (Fig. 1). The first  
 132 lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics:  
 133 gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*,  
 134 provided an overview of our current understanding of how stars form. We selected these particular  
 135 lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad  
 136 audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training  
 137 on participants' abilities to learn from the lectures. To this end, we selected two introductory  
 138 videos that were intended to be viewed at the start of students' training in their respective content  
 139 areas. Second, we wanted the two lectures to have some related content – so that we could test  
 140 our approach's ability to distinguish similar conceptual content. To this end, we chose two videos  
 141 from the same Khan Academy course domain, “Cosmology and Astronomy.” Third, we sought to  
 142 minimize dependencies and specific overlap between the videos. For example, we did not want  
 143 participants' abilities to understand one video to (directly) influence their abilities to understand the  
 144 other. To satisfy this last criterion, we chose videos from two different lecture series (Lectures 1 and  
 145 2 were from the “Scale of the Universe” and “Stars, Black Holes, and Galaxies” series, respectively).



**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.

146 We also wrote a set of multiple-choice quiz questions that we hoped would enable us to  
 147 evaluate participants’ knowledge about each individual lecture, along with related knowledge  
 148 about physics concepts not specifically presented in either video (see Supp. Tab. 1 for the full list  
 149 of questions in our stimulus pool). Participants answered questions randomly drawn from each  
 150 content area (Lecture 1, Lecture 2, and general physics knowledge) on each of the three quizzes.  
 151 Quiz 1 was intended to assess participants’ “baseline” knowledge before training, Quiz 2 assessed  
 152 knowledge after watching the *Four Fundamental Forces* video (i.e., Lecture 1), and Quiz 3 assessed  
 153 knowledge after watching the *Birth of Stars* video (i.e., Lecture 2).

154 To study in detail how participants’ conceptual knowledge changed over the course of the  
 155 experiment, we first sought to model the conceptual content presented to them at each moment  
 156 throughout each of the two lectures. We adapted an approach we developed in prior work [30]  
 157 to identify the latent themes in the lectures using a topic model [8]. Briefly, topic models take  
 158 as input a collection of text documents, and learn a set of “topics” (i.e., latent themes) from their  
 159 contents. Once fit, a topic model can be used to transform arbitrary (potentially new) documents  
 160 into sets of “topic proportions,” describing the weighted blend of learned topics reflected in their

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

173 We hypothesized that a topic model trained on transcripts of the two lectures should also capture  
174 the conceptual knowledge probed by each quiz question. If indeed the topic model could capture  
175 information about the deeper conceptual content of the lectures (i.e., beyond surface-level details  
176 such as particular word choices), then we should be able to recover a correspondence between  
177 each lecture and questions *about* each lecture. Importantly, such a correspondence could not  
178 ~~solely arise~~ arise solely from superficial text matching between lecture transcripts and questions,  
179 since the lectures and questions often used different words (Supp. Fig. 10) and phrasings. Simply  
180 comparing the average topic weights from each lecture and question set (averaging across time and  
181 questions, respectively) reveals a striking correspondence (Supp. Fig. 2). Specifically, the average  
182 topic weights from Lecture 1 are strongly correlated with the average topic weights from questions  
183 about Lecture 1 questions ( $r(13) = 0.809$ ,  $p < 0.001$ , 95% confidence interval (CI) = [0.633, 0.962]),  
184 and the average topic weights from Lecture 2 are strongly correlated with the average topic weights  
185 from questions about Lecture 2 questions ( $r(13) = 0.728$ ,  $p = 0.002$ , 95% CI = [0.456, 0.920]). At  
186 the same time, the average topic weights from the two lectures are *negatively* correlated with  
187 the average topic weights from their non-matching question sets (Lecture 1 video vs. Lecture 2  
188 questions:  $r(13) = -0.547$ ,  $p = 0.035$ , 95% CI = [-0.812, -0.231]; Lecture 2 video vs. Lecture 1



**Figure 3: Lecture and question topic overlap. A. Topic weight variability.** The bar plots display the variance ~~ef~~ in 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.

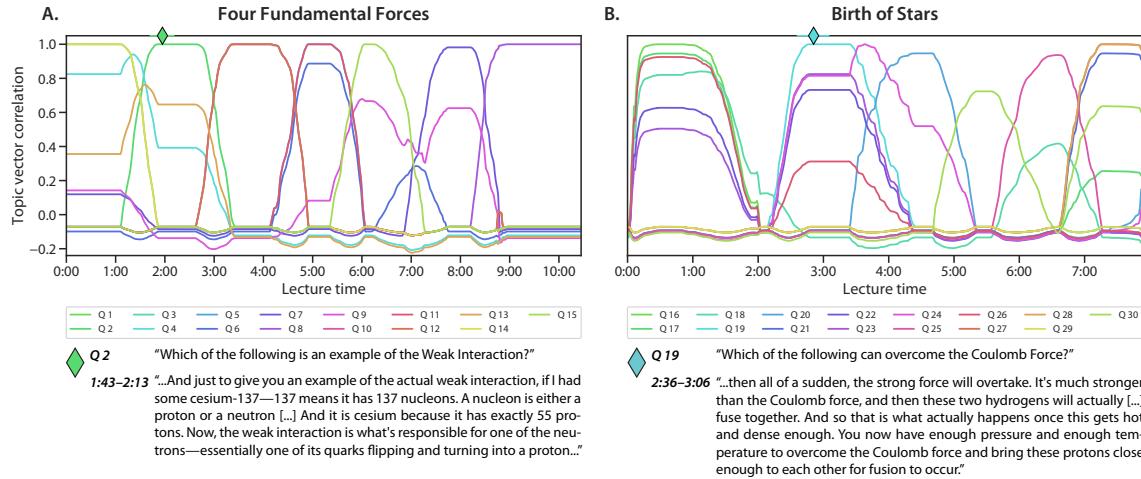
189 questions:  $r(13) = -0.612$ ,  $p = 0.015$ , 95% CI =  $[-0.874, -0.281]$ ), indicating that the topic model  
190 also exhibits some degree of specificity. The full set of pairwise comparisons between average  
191 topic weights for the lectures and question sets is reported in Supplementary Figure 2.

192 Another, more sensitive, way of summarizing the conceptual content of the lectures and ques-  
193 tions is to look at *variability* in how topics are weighted over time and across different questions  
194 (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “in-  
195 formation” [23] the lecture (or question set) reflects about that topic. For example, suppose a  
196 given topic is weighted on heavily throughout a lecture. That topic might be characteristic of  
197 some aspect or property of the lecture *overall* (conceptual or otherwise), but unless the topic’s  
198 weights ~~changed change~~ in meaningful ways over time, ~~the topic it~~ would be a poor indicator of  
199 any *specific* conceptual content in the lecture. We therefore also compared the variances in topic  
200 weights (~~across time or over time and across~~ questions) between the lectures and questions. The  
201 variability in topic expression (~~over time and across questions~~) was similar for the Lecture 1 video  
202 and questions ( $r(13) = 0.824$ ,  $p < 0.001$ , 95% CI =  $[0.696, 0.973]$ )~~and, and for~~ 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 variabilities 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 An~~ individual lecture may be organized around a single broad theme at a coarse scale, ~~but~~ at a finer scale, each moment of a lecture typically covers a narrower range of content. Given the correspondence we found between the variabilities 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 outside of which~~ 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 questions’~~ text to the transcribed text from 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 time series plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel B.: correlations for the *Birth of Stars* lecture and associated questions). The colors denote question identities. The diamonds in each panel denote the moment of peak correlation between the indicated question and the lecture trajectory. The associated questions’ text and snippets of the lectures’ transcripts from the surrounding 30 seconds, are displayed at the bottom of the figure.

231 approaches to estimating how much a student “knows” about the content of a given lecture entail  
 232 administering some form of assessment (e.g., a quiz) and computing the proportion of **correctly**  
 233 **answered questions**—**questions the student answered correctly**. But if two students receive identical  
 234 scores on such an **exam****assessment**, might our modeling framework help us to gain more nuanced  
 235 insights into the *specific* content that each student has mastered (or failed to master)? For example,  
 236 a student who misses three questions that were all about the same concept (e.g., concept *A*) will  
 237 have gotten the same *proportion* of questions correct as another student who missed three questions  
 238 about three *different* concepts (e.g., *A*, *B*, and *C*). But if we wanted to help these two students fill  
 239 in the “gaps” in their understandings, we might do well to focus specifically on concept *A* for the  
 240 first student, but to also add in materials pertaining to concepts *B* and *C* for the second student. In  
 241 other words, raw “proportion-correct” measures may capture *how much* a student knows, but not  
 242 *what* they know. We wondered whether our modeling framework might enable us to (formally and  
 243 automatically) infer participants’ knowledge at the scale of individual concepts (e.g., as captured

244 by a single moment of a lecture).

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

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

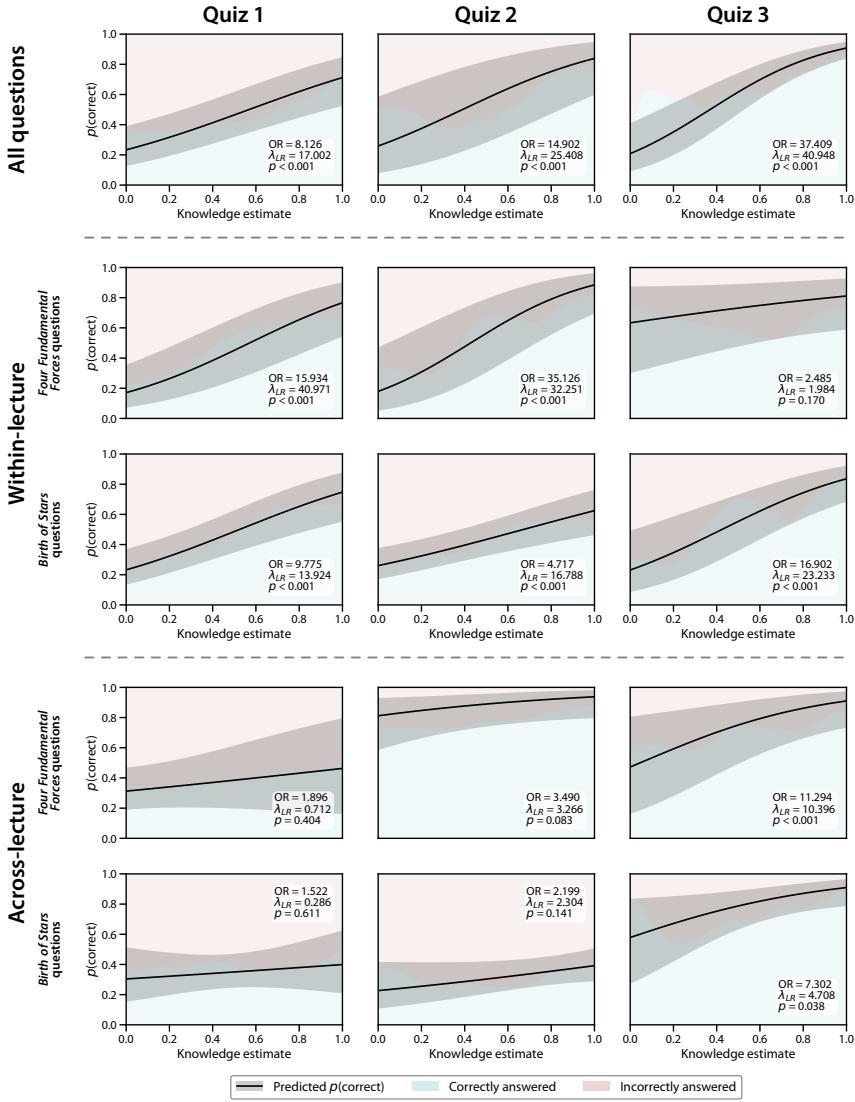


**Figure 5: Estimating knowledge about the content presented at each moment of each lecture.** **A. Knowledge about the time-varying content of *Four Fundamental Forces*.** Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture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

versus 3 ( $t(49) = 0.160$ ,  $p = 0.874$ ). Similarly, we hypothesized (and subsequently confirmed) that participants should show greater estimated knowledge about the content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 5D). Specifically, since participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a “boost” on Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge about the *Birth of Stars* lecture content on ~~Quizzes~~Quiz 1 versus 2 ( $t(49) = 1.013$ ,  $p = 0.316$ ), but ~~the~~-estimated knowledge was substantially higher on Quiz 3 versus 2 ( $t(49) = 10.561$ ,  $p < 0.001$ ) and Quiz 3 versus 1 ( $t(49) = 8.969$ ,  $p < 0.001$ ).

If we are able to accurately estimate a participant’s knowledge about the content tested by a given question, our estimates of their knowledge should carry some predictive information about whether they are likely to answer that question correctly or incorrectly. We developed a statistical approach to test this claim. For each quiz question a participant answered, in turn, we used Equation 1 to estimate their knowledge at ~~the given that~~ question’s ~~embedding space~~ embedding-space coordinate based on other questions that participant answered on the same quiz. We repeated this for all participants, and for each of the three quizzes. Then, separately for each quiz, we fit a generalized linear mixed model (GLMM) with a logistic link function to explain the ~~likelihood~~probability of correctly answering a question as a function of estimated knowledge ~~for at~~ its embedding coordinate, while accounting for ~~random variation among varied effects of individual~~ participants and questions (see *Generalized linear mixed models*). To assess the predictive value of the knowledge estimates, we compared each GLMM to an analogous (i.e., nested) “null” model that ~~did not consider estimated knowledge assumed these estimates carried no predictive information~~ using parametric bootstrap likelihood-ratio tests.

We carried out three different versions of the analyses described above, wherein we considered different sources of information in our estimates of participants’ knowledge for each quiz question. First, we estimated knowledge at each held-out question’s embedding coordinate using *all* other questions answered by the same participant on the same quiz (“All questions”; Fig. 6, top row). This test was intended to assess the overall predictive power of our approach. Second, we estimated



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

knowledge for each question about a given lecture using only the other questions (from the same participant and quiz) about that *same* lecture (“Within-lecture”; Fig. 6, middle rows). This test was intended to assess the *specificity* of our approach by asking whether our predictions could distinguish between questions about different content covered by the same lecture. Third, we estimated knowledge for each question about one lecture using only the questions (from the same participant and quiz) about the *other* lecture (“Across-lecture”; Fig. 6, bottom rows). This test was intended to assess the *generalizability* of our approach by asking whether our predictions ~~held~~could extend across the content areas of the two lectures.

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

328 ~~answered questions~~ questions (see Generalized linear mixed models).

329 Before presenting our results, it is worth considering three possible explanations of why a  
330 participant might answer a given question correctly or incorrectly. One possibility is that the  
331 participant simply *guessed* the answer. A second is that they selected the incorrect answer by  
332 mistake, despite “knowing” the correct answer (or vice versa). In both of these scenarios, the  
333 participant’s knowledge about the question’s content should be uninformative about their observed  
334 response. A third possibility is that the participant’s response reflects their *actual* knowledge about  
335 the question’s content. In this case, we *might* expect to see a positive relationship between the  
336 participant’s knowledge and their likelihood of answering the question correctly. However, in  
337 order to see this positive relationship, the participant’s knowledge must be structured in a way  
338 that is reflected (at least partially) by the embedding space. In other words, if the participant’s  
339 performance reflects their true knowledge, but our text embedding space does not sufficiently  
340 capture the structure of that knowledge, then the knowledge estimates we generate will not be  
341 predictive of the participant’s performance. In the extreme, if the embedding space is completely  
342 unstructured with respect to the content of the quiz questions, then we would expect to see the  
343 negative relationship between estimated knowledge and performance that we described above.

344 When we fit a GLMM to estimates of participants’ knowledge for each Quiz 1 question based  
345 on all other Quiz 1 questions, we ~~observed an outcome consistent with our null hypothesis:~~  
346 found that higher estimated knowledge ~~at the embedding coordinate of a held-out question~~  
347 ~~was associated with a lower~~ for a given question predicted a greater likelihood of answering  
348 ~~the question~~ it correctly (odds ratio ( $OR$ ) = 0.136, likelihood-ratio test statistic ( $\lambda_{LR}$ ) = 19.749,  
349 ( $OR$ ) = 8.126, 95% CI = [14.352, 26.545],  $p < 0.001$  CI = [3.116, 20.123], likelihood-ratio test statistic  
350 ( $\lambda_{LR}$ ) = 17.002,  $p < 0.001$ ). This ~~outcome suggests that our knowledge estimates do not provide~~  
351 ~~useful information about participants’ Quiz 1 performance when we aggregated across all question~~  
352 ~~content areas. We speculated that this might either indicate that the knowledge estimates are~~  
353 ~~uninformative in general, or about Quiz 1 performance in particular. This would be expected,~~  
354 ~~for example, if participants were guessing about the answers to the Quiz 1 questions (prior~~  
355 ~~to having watched either lecture). When we~~ held when we repeated this analysis

356 for QuizzesQuiz 2 and ( $OR = 14.902$ , 95% CI = [4.976, 39.807],  $\lambda_{LR} = 25.408$ ,  $p < 0.001$ ) and again  
357 for Quiz 3, we found that higher estimated knowledge for a given question predicted a greater  
358 likelihood of answering it correctly (Quiz 2:  $OR = 2.905$ ,  $\lambda_{LR} = 17.333$ , 95% CI = [14.966, 29.309],  $p = 0.001$ ;  
359 Quiz 3:  $OR = 3.238$ ,  $\lambda_{LR} = 6.882$ , 95% CI = [6.228, 8.184],  $p = 0.016$  ( $OR = 37.409$ , 95% CI = [10.425, 107.145],  $\lambda_{LR} = 40$ ).  
360 Taken together, these results suggest that our knowledge estimates can reliably predict participants'  
361 performance on individual held-out quiz questions, but only after participants have received at  
362 least some training questions when they incorporate information from all (other) quiz content.  
363 We observed a similar pattern set of results when used this approach to estimate we restricted  
364 our estimates of participants' knowledge about held-out questions from one lecture using to  
365 consider only their performance on other questions from about the same lecture. Specifically,  
366 for Quiz 1 questions (i.e., prior to watching either), participants' estimated knowledge for the  
367 embedding coordinates of held-out participants' knowledge of Four Fundamental Forces-related  
368 questions estimated using estimated from their performance on other Four Fundamental Forces-  
369 related questions did not reliably predict whether those questions were answered correctly ( $OR = 1.891$ ,  $\lambda_{LR} = 2.293$ , 95%  
370 was predictive of their ability to answer those questions correctly ( $OR = 15.934$ , 95% CI = [5.173, 38.005],  $\lambda_{LR} = 40.971$ ,  $p < 0.001$ ).  
371 The same was true of knowledge estimates for held-out participants' estimated knowledge for  
372 Birth of Stars-related questions based on their performance on other Birth of Stars-related questions  
373 ( $OR = 0.722$ ,  $\lambda_{LR} = 5.115$ , 95% CI = [0.094, 0.146],  $p = 0.738$ ). As in our analysis that included all  
374 questions, we speculate that these "null" results might reflect some degree of random guessing on  
375 Quiz 1. When we repeated these within-lecture analyses using questions from Quiz  $OR = 9.775$ , 95% CI = [2.93, 25.08],  $\lambda_{LR} = 32.251$ ,  $p < 0.001$ .  
376 These results also held for participants' Quiz 2 (which participants took immediately after viewing  
377 responses (Four Fundamental Forces but prior to viewing:  $OR = 35.126$ , 95% CI = [5.113, 123.868],  $\lambda_{LR} = 32.251$ ,  $p < 0.001$ ).  
378 Birth of Stars), we found that they now reliably predicted success on Four Fundamental Forces-related  
379 questions ( $OR = 9.023$ ,  $\lambda_{LR} = 18.707$ , 95% CI = [10.877, 22.222],  $p < 0.001$ ) but not on:  $OR = 4.717$ , 95% CI = [2.021, 9.805],  $\lambda_{LR} = 16.902$ , 95% CI = [3.353, 53.265],  $\lambda_{LR} = 23.233$ ,  $p < 0.001$ ; Four Fun-  
380 and partially for their Quiz 3 responses (Birth of Stars-related questions ( $OR = 0.306$ ,  $\lambda_{LR} = 5.115$ , 95% CI = [4.624, 5.655],  $\lambda_{LR} = 16.902$ , 95% CI = [3.353, 53.265],  $\lambda_{LR} = 23.233$ ,  $p < 0.001$ ); Four Fun-  
381 Here, we speculate that participants might have been guessing about the Birth of Stars content  
382 (e.g., prior to having watched it), whereas they might have been drawing on some structured  
383 knowledge about the:  $OR = 16.902$ , 95% CI = [3.353, 53.265],  $\lambda_{LR} = 23.233$ ,  $p < 0.001$ ; Four Fun-

384 ~~damental Forces~~content (e.g., from having just watched it). When we applied this approach to  
385 :  $OR = 2.485$ , 95% CI = [0.724, 8.366],  $\lambda_{LR} = 1.984$ ,  $p = 0.170$ ). Speculatively, the Quiz 3 results  
386 suggest that the within-lecture knowledge estimates may be susceptible to ceiling effects in  
387 participants' quiz performance. On Quiz 3 responses (given immediately after viewing *Birth of*  
388 *Stars*), we found that within-lecture knowledge estimates for ~~after viewing both lectures, no~~ participant answered more than three *Birth of Stars*~~Four Fundamental Forces~~-related questions ~~could~~  
389 now reliably predict success on those questions ( $OR = 5.467$ ,  $\lambda_{LR} = 10.670$ , 95% CI = [7.998, 12.532],  $p = 0.005$ )  
390 . However, incorrectly, and all but five participants (out of 50) answered two or fewer incorrectly.  
391 (This was the only subset of questions about either lecture, across all three quizzes, for which this  
392 was true.) Consequently, for 90% of participants, our within-lecture knowledge estimates estimates  
393 of their knowledge for *Four Fundamental Forces*questions answered on Quiz 3 were no longer  
394 directly related to ~~related questions that they answered incorrectly leveraged information from at~~  
395 most a single other question they were *not* able to correctly answer. This likely hampered our ability  
396 to accurately characterize the specific (and by the time they took Quiz 3, relatively few) aspects  
397 of the lecture content these participants did *not* know about, and successfully distinguish them  
398 from the far more numerous aspects of the likelihood of successfully answering them and instead  
399 exhibited the inverse relationship we would expect to arise from unstructured knowledge (with  
400 respect to the embedding space;  $OR = 0.013$ ,  $\lambda_{LR} = 14.648$ , 95% CI = [10.695, 23.096],  $p < 0.001$ ).  
401 Speculatively, we suggest that this may reflect participants forgetting some of the *Four Fundamental*  
402 *Forces* content (e.g., perhaps in favor of prioritizing encoding the just-watched *Birth of Stars* content  
403 in preparation for the third quiz). If this forgetting happens in a relatively "random" way (with  
404 respect to spatial distance within the embedding space), then it could explain why some held-out  
405 questions about *Four Fundamental Forces* were answered incorrectly, even if questions at nearby  
406 coordinates (i.e., about similar content) were answered correctly. This might lead our approach  
407 to over-estimate knowledge for held-out questions about "forgotten" knowledge that participants  
408 answered incorrectly. lecture content they now *did* know about. Taken together, these within-  
409 lecture results suggest that our approach can knowledge estimates can reliably distinguish between  
410 questions about different content covered by a single lecture when participants have sufficiently

412 structured knowledge about its contents, though this specificity may decrease with time since the  
413 relevant material was learned, provided there is sufficient diversity in participants' quiz responses  
414 to extract meaningful information about both what they know and what they do not know.

415 Finally, we used this approach to estimate estimated participants' knowledge about held-out  
416 questions from one lecture using for each question about each lecture using only their performance  
417 on questions from the (from the same quiz) about the other lecture. Here we again observed a  
418 similar pattern of results, though with some notable differences. On Quiz This is an especially  
419 stringent test of our approach. Our primary assumption in constructing our knowledge estimates is  
420 that knowledge about a given concept is similar to knowledge about other concepts that are nearby  
421 in the embedding space. However, our analyses in Figure 3 and Supplementary Figure 2 show  
422 that the embeddings of content from the two lectures (and of their associated quiz questions) are  
423 largely distinct from each other. Therefore, any predictive power of these across-lecture knowledge  
424 estimates must overcome large distances in the embedding space. To put this in concrete terms,  
425 this test requires predicting participants' performance on individual, highly specific questions  
426 about the formation of stars from their responses to just five multiple-choice questions about the  
427 fundamental forces of the universe (and vice versa).

428 We found that, before viewing either lecture (i.e., on Quiz 1, we found that participants' abilities  
429 to correctly answer questions about), participants' abilities to answer Four Fundamental Forces could  
430 be predicted from their responses to questions about Birth of Stars ( $OR = 1.896$ ,  $\lambda_{LR} = 7.205$ , 95% CI = [6.224, 7.524],  $p = 0$ )  
431 and similarly, that their ability to correctly answer Birth of Stars-related questions could be predicted  
432 from their responses to Four Fundamental Forces-related questions ( $OR = 1.522$ ,  $\lambda_{LR} = 6.448$ , 95% CI = [5.656, 6.843],  $p = 0$ ).  
433 Given the results from our analyses that included all questions and within-lecture predictions, we  
434 were surprised to find that the knowledge estimates could reliably (if weakly) predict participants'  
435 performance across content from different lectures. It is possible that this result reflects a  
436 combination of random guessing prior to training (leading to a weak effect size), alongside  
437 some coarse-scale structured knowledge that participants had about the content prior to watching  
438 either lecture. When we repeated this analysis using questions from Quiz 2, we found participants'  
439 responses to Four Fundamental Forces-related questions did not reliably predict their success on could

440 not be predicted from their responses to *Birth of Stars*-related questions ( $OR = 1.865, \lambda_{LR} = 3.205, 95\% CI = [3.027, 3.600]$ ),  
441 nor did their responses to could their abilities to answer *Birth of Stars*-related questions reliably  
442 predict their success on be predicted from their responses to *Four Fundamental Forces*-related ques-  
443 tions ( $OR = 3.490, \lambda_{LR} = 3.266, 95\% CI = [3.033, 3.866], p = 0.093$ ). These “prediction failures” appear  
444 to come from the fact that any signal derived from participants’ knowledge about the content of the  
445 *Birth of Stars* lecture (prior to watching it) is overwhelmed by the much more dramatic increase in  
446 their knowledge about the content of the *Four Fundamental Forces* (which they watched just prior to  
447 taking Quiz 2). This is reflected in their Quiz 2 performance for questions about each lecture (mean  
448 proportion correct for *Four Fundamental Forces*-related  $OR = 1.522, 95\% CI = [0.332, 6.835], \lambda_{LR} = 0.286, p = 0.611$ ).  
449 Similarly, we found that participants’ performance on questions about either lecture could not be  
450 predicted given their responses to questions about the other lecture after viewing *Four Fundamental*  
451 *Forces* but before viewing *Birth of Stars* (i.e., on Quiz 2: 0.77; mean proportion correct for *Four*  
452 *Fundamental Forces* questions given *Birth of Stars* -related questions on Quiz 2: 0.36). When we  
453 carried out these across-lecture knowledge predictions using questions from questions:  $OR = 3.49, 95\% CI = [0.739, 12.8]$   
454 *Birth of Stars* questions given *Four Fundamental Forces* questions:  $OR = 2.199, 95\% CI = [0.711, 5.623], \lambda_{LR} = 2.304, p = 0.$   
455 Only after viewing both lectures (i.e., on Quiz 3 (when participants had now viewed both lectures)),  
456 we could again reliably predict success on ) did these across-lecture knowledge estimates reliably  
457 predict participants’ success on individual quiz questions (*Four Fundamental Forces* questions given  
458 *Birth of Stars* questions about both *Four Fundamental Forces* ( $OR = 11.294, \lambda_{LR} = 11.055, 95\% CI = [9.126, 18.476], p = 0.00$ )  
459 and:  $OR = 11.294, 95\% CI = [1.375, 47.744], \lambda_{LR} = 10.396, p < 0.001$ ; *Birth of Stars* ( $OR = 7.302, \lambda_{LR} = 7.068, 95\% CI = [$   
460 using responses to questions about the other lecture’s content. Across all three versions of these  
461 analyses, our questions given *Four Fundamental Forces* questions:  $OR = 7.302, 95\% CI = [1.077, 44.879], \lambda_{LR} = 4.708, p =$   
462 Taken together, these results suggest that (by and large) our knowledge estimates can reliably  
463 predict participants’ abilities to answer individual quiz questions, distinguish between questions about  
464 similar content , and generalize across our ability to form estimates solely across different content  
465 areas is more limited than our ability to form estimates that incorporate responses to questions  
466 from both content areas (as in Fig. 6, “All questions”) or within a single content area (as in Fig. 6,  
467 “Within-lecture”). However, if participants have recently received some training on both content

468 areas, ~~provided that participants' quiz responses reflect a minimum level of "real" knowledge~~  
469 ~~about both content on which these predictions are based and that for which they are made~~ the  
470 knowledge estimates appear to be informative even across content areas.

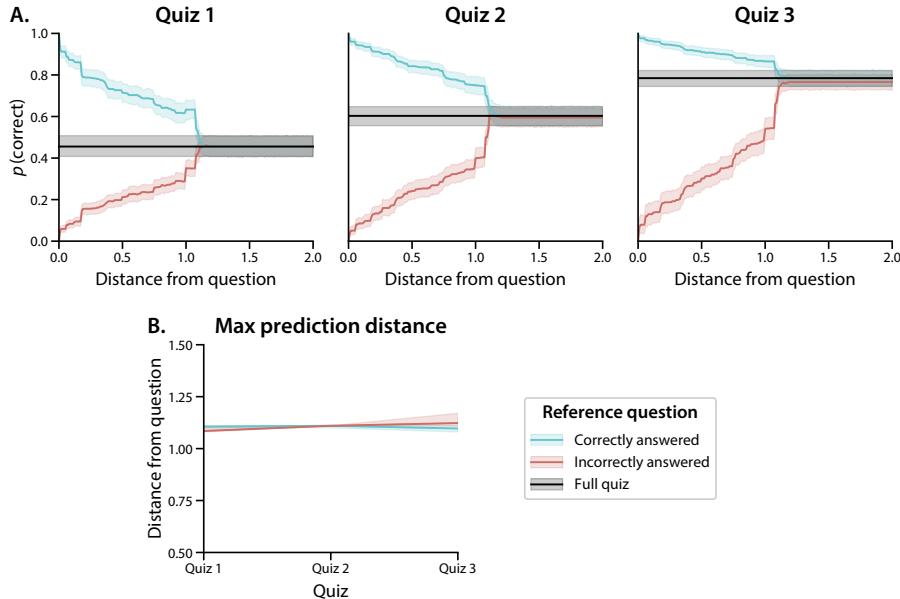
471 We speculate that these "Across-lecture" results might relate to some of our earlier work on the  
472 nature of semantic representations [45]. In that work, we asked whether semantic similarities could  
473 be captured through behavioral measures, even if participants' "true" internal representations  
474 differed from the embeddings used to *characterize* their behaviors. We found that mismatches  
475 between an individual's internal representation of a set of concepts and the representation used  
476 to characterize their behaviors can lead to underestimates of how semantically driven those  
477 behaviors are. Along similar lines, we suspect that in our current study, participants' conceptual  
478 representations may initially differ from the representations learned by our topic model. (Although  
479 the topic model's representations are still *related* to participants' initial internal representations;  
480 otherwise we would have found that knowledge estimates derived from Quizzes 1 and 2 had  
481 no predictive power in the other tests we conducted.) After watching both lectures, however,  
482 participants' internal representations may become more aligned with the embeddings used to  
483 estimate their knowledge (since those embeddings were trained on the lectures' transcripts). This  
484 could help explain why the knowledge estimates derived from Quizzes 1 and 2 (before both lectures  
485 had been watched) do not reliably predict performance across content areas, whereas estimates  
486 derived from Quiz 3 do.

487 That the knowledge predictions derived from the text embedding space reliably distinguish be-  
488 tween ~~held-out correctly versus incorrectly answered~~ correctly and incorrectly answered held-out  
489 questions (Fig. 6) suggests that ~~spatial geometric~~ relationships within this space can help explain  
490 what participants know. But how far does this explanatory power extend? For example, suppose  
491 we know that a participant correctly answered a question at embedding coordinate  $x$ . As we move  
492 farther away from  $x$  in the embedding space, how does the likelihood that the participant knows  
493 about the content at a given location "fall off" with distance? Conversely, suppose the participant  
494 instead answered that same question ~~in~~incorrectly. Again, as we move farther away from  
495  $x$  in the embedding space, how does the likelihood that the participant does *not* know about a

496 coordinate's content the content at a given coordinate change with distance? We reasoned that,  
497 assuming our embedding space is capturing something about how individuals actually organize  
498 their knowledge, a participant's ability to answer questions embedded very close to  $x$  should tend  
499 to be similar to their ability to answer the question embedded at  $x$ . Whereas at another extreme,  
500 But once we reach some sufficiently large distance from  $x$ , our ability to infer whether or not a par-  
501 ticipant will correctly answer a question based on their ability to answer the question at  $x$  should be  
502 no better than guessing based on their *overall* proportion of correctly answered questions. In other  
503 words, beyond the maximum distance at which the a participant's ability to answer the question  
504 at  $x$  is informative of their ability to answer a second question at location  $y$ , then guessing the  
505 outcome at  $y$  based on the outcome at  $x$  should be no more successful than guessing based on a  
506 measure that does not consider embedding space embedding-space distance.

507 With these ideas in mind, we asked: conditioned on answering a a participant's ability to  
508 answer a given question correctly, what proportion of all questions (within some radius  $r$ , of that  
509 question's embedding coordinate) were answered of its embedding coordinate were they able to  
510 answer correctly? We plotted this proportion as a function of  $r$ . Similarly, we could ask, conditioned  
511 on answering a question incorrectly, how the proportion of correct responses changed with r for  
512 questions that participants answered correctly, and for questions they answered incorrectly. As  
513 shown in Figure 7, we found that quiz performance falls off smoothly with distance, and the "rate"  
514 of the falloff at which it falls off does not appear to change across the different differ across quizzes,  
515 as measured by the distance at which performance becomes statistically indistinguishable from a  
516 simple proportion correct proportion-correct score (see Estimating the "smoothness" of knowledge).  
517 This suggests that, at least within the region of text embedding space covered spanned by the  
518 questions our study's participants answered (and as characterized using our topic model), the rate  
519 at which knowledge changes with distance is relatively constant, even as participants' overall level  
520 of knowledge varies across quizzes or and regions of the embedding space.

521 Knowledge estimates need not be limited to the content of the lectures contents of these  
522 particular lectures and quizzes. As illustrated in Figure 8, our general approach to estimating  
523 knowledge from a small number of quiz questions may be extended to *any* content, given its text



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

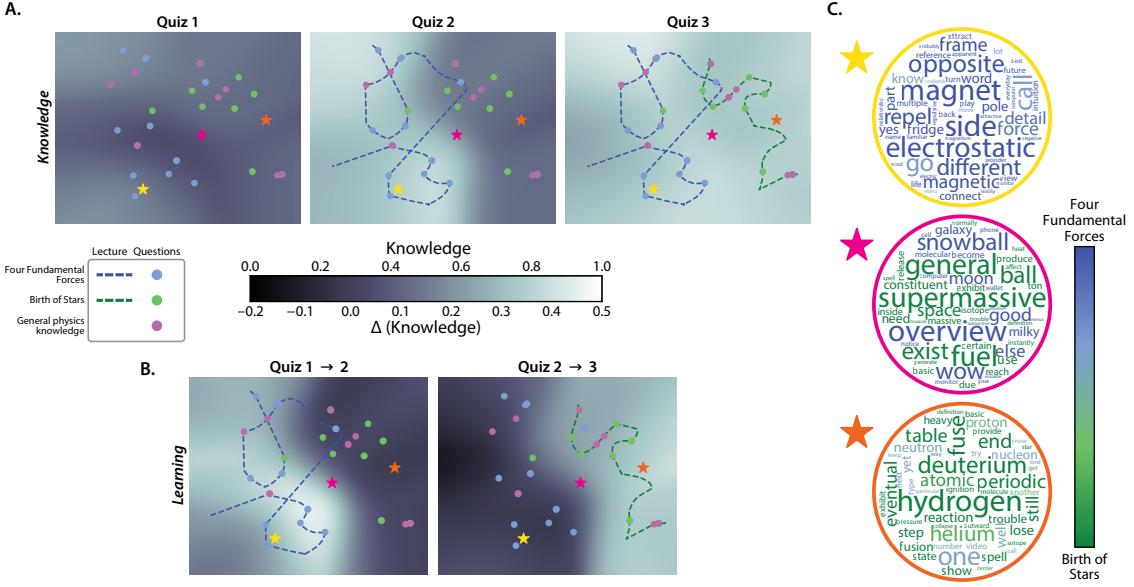
embedding coordinate. To visualize how knowledge “spreads” through text embedding space to content beyond the lectures participants watched *and the questions they answered*, we first fit a new topic model to the lectures’ sliding windows with  $k = 100$  topics. Conceptually, increasing the number of topics used by the model functions to increase the “resolution” of the embedding space, providing a greater ability to estimate knowledge for content that is highly similar to (but not precisely the same as) that contained in the two lectures. *We note that we used these 2D maps solely for visualization; all relevant comparisons, distance computations, and statistical tests we report above were carried out in the original 15-dimensional space, using the 15 topics used to train the model.* Aside from increasing the number of topics from 15 to 100, all other procedures and

533 model parameters were carried over from the preceding analyses. As in our other analyses, we  
534 resampled each lecture's topic trajectory to 1 Hz and projected each question into a shared text  
535 embedding space.

536 We projected the resulting 100-dimensional topic vectors (for each second of ~~video~~ the lectures  
537 and each quiz question) onto a shared 2-dimensional plane (see *Creating knowledge and learning map*  
538 *visualizations*). Next, we sampled points from a  $100 \times 100$  grid of coordinates that evenly tiled a  
539 rectangle enclosing the 2D projections of the ~~videos~~ lectures and questions. We then used Equation 4  
540 to estimate participants' knowledge at each of these 10,000 sampled locations, and averaged these  
541 estimates across participants to obtain an estimated average *knowledge map* (Fig. 8A). Intuitively,  
542 the knowledge map constructed from a given quiz's responses provides a visualization of ~~how~~  
543 “how” much participants knew about any content expressible by the fitted text embedding model  
544 at the point in time when they completed that quiz. We note that we used these 2D maps solely  
545 for visualization; all relevant comparisons, distance computations, and statistical tests we report  
546 above were carried out in the original 15-dimensional space, using the 15-topic model.

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

560 Another way of visualizing these content-specific increases in knowledge after participants



**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 participants’](#) 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 5, 6, and 7. **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 8 and 9. **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.

561 viewed each lecture is displayed in Figure 8B. Taking the point-by-point difference between the  
562 knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map*  
563 that describes the *change* in ~~knowledge estimates~~ ~~estimated knowledge~~ from one quiz to the next.  
564 These learning maps highlight that the estimated knowledge increases we observed across maps  
565 were specific to the regions around the embeddings of each lecture, in turn.

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

## 580 Discussion

581 We developed a computational framework that uses short multiple-choice quizzes to gain nuanced  
582 insights into what learners know and how their knowledge changes with training. First, we show  
583 that our approach can automatically match the conceptual knowledge probed by individual quiz  
584 questions to the corresponding moments in lecture videos when those concepts were presented  
585 (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment "knowledge traces" that  
586 reflect the degree of knowledge participants have about each ~~video~~ lecture's time-varying content,

587 and capture temporally specific increases in knowledge after viewing each lecture (Fig. 5). We  
588 ~~also then~~ show that these knowledge estimates can generalize to held-out questions ~~and predict~~  
589 ~~participants' abilities to answer them correctly~~ (Fig. 6). Finally, we use our framework to construct  
590 visual maps that provide snapshot estimates of how much participants know about any concept  
591 within the scope of our text embedding model, and how much their knowledge of those concepts  
592 changes with training (Fig. 8).

593 ~~We view our work as making~~ Our work makes several contributions to the study of how people  
594 acquire conceptual knowledge. First, from a methodological standpoint, our modeling framework  
595 provides a systematic means of mapping out and characterizing knowledge in maps that have  
596 infinite (arbitrarily many) numbers of coordinates, and of “filling out” those maps using relatively  
597 small numbers of ~~multiple choice~~ ~~multiple-choice~~ quiz questions. Our experimental finding that we  
598 can use these maps to predict ~~responses to success on~~ held-out questions has several psychological  
599 implications as well. For example, concepts that are assigned to nearby coordinates by the text  
600 embedding model also appear to be “known to a similar extent” (as reflected by participants’  
601 responses to held-out questions; Fig. 6). This suggests that participants also *conceptualize* similarly  
602 the content reflected by nearby embedding coordinates. How participants’ knowledge ~~falls off~~  
603 ~~“falls off”~~ with spatial distance is captured by the knowledge maps we infer from their quiz  
604 responses (e.g., Figs. 7, 8). In other words, our study shows that knowledge about a given concept  
605 implies knowledge about related concepts, and ~~we also show how estimated knowledge falls off~~  
606 ~~with distance~~ ~~how far this implication extends~~ in text embedding space.

607 In our study, we characterize the “coordinates” of participants’ knowledge using a relatively  
608 simple “~~bag-of words~~ ~~bag-of-words~~” text embedding model [LDA; 8]. More sophisticated text em-  
609 bedding models, such as transformer-based models [18, 56, 69, 72] ~~can learn~~ ~~can leverage additional~~  
610 ~~textual information such as~~ complex grammatical and semantic relationships between words,  
611 higher-order syntactic structures, stylistic features, and more. We considered using transformer-  
612 based models in our study, but we found that the text embeddings derived from these models  
613 were surprisingly uninformative with respect to differentiating or otherwise characterizing the  
614 conceptual content of the lectures and questions we used (*see Supplementary results*). We suspect

that this reflects a broader challenge in constructing models that are both high-resolution within a given domain (e.g., the domain of physics lectures and questions) *and* sufficiently broad so as to enable them to cover a wide range of domains. ~~For example, we found that the embeddings derived even from much larger and more modern models like BERT [18], GPT [72], LLaMa [69], and others that are trained on enormous text corpora, end up yielding poor resolution within the content space spanned by individual course videos (Supp. Fig. 11).~~ Whereas the LDA embeddings of the lectures and questions are “near” each other (i.e., the convex hull enclosing the two lectures’ trajectories is highly overlapping with the convex hull enclosing the questions’ embeddings), the BERT embeddings of the lectures and questions are instead largely distinct (top row of Supp. Fig. 11). 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. 11). We also examined per-question “content matches” between individual questions and individual moments of each lecture (Figs. 4, 11). The time series 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 time series 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). Essentially, “larger” language models learn more complex features of language through training on enormous and diverse text corpora. But as a result, their embedding spaces also “span” an enormous

and diverse range of conceptual content, 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. 11). Taken together, these comparisons illustrate how LDA (trained on the specific content in question) sacrificing a degree of specificity in their capacities to distinguish subtle conceptual differences within a more narrow range of content. In comparing our LDA model (trained specifically on the lectures used in our study) to a larger transformer-based model (BERT), we found that our LDA model 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 individual questions, while BERT essentially relegates the contents of both lectures and all of the quiz questions (which are all broadly about “physics”) into to a tiny region of its embedding space, thereby blurring out meaningful distinctions between different specific concepts covered by the lectures and questions. (Supp. Fig. 11). We note that these are not criticisms of BERT (or, nor of other large language models trained on large and diverse corpora). Rather, our point is that simple fine-tuned simpler models trained on a relatively small but specialized corpus corpora can outperform much more complicated complex models trained on much larger corpora, when we are specifically interested in capturing subtle conceptual differences at the level of a single, narrowly focused course lecture or question. Of course quiz question. On the other hand, if our goal had been to find choose a model that generalized to many different content areas simultaneously, we would expect our approach LDA model to perform comparatively poorly relative to BERT or other much larger general-purpose models. We suggest that bridging the tradeoff between this tradeoff between achieving high resolution within each content area versus a single content area and the ability to generalize to many different diverse content areas will be an important challenge for future work in this domain.

671 At the opposite end of the spectrum from large language models, one could also imagine  
672 using an even *simpler* “model” than LDA that relates the contents of course lectures and quiz  
673 questions through explicit word-overlap metrics (rather than similarities in the latent topics they  
674 exhibit). In a supplementary analysis (Supp. Fig. 10), we compared the LDA-based question-lecture  
675 matches shown in Figure 4 with analogous matches based on the Jaccard similarity between each  
676 question’s text and each sliding window from the corresponding lecture’s transcript. As for  
677 the embeddings derived from BERT, we found that this word-matching approach also blurred  
678 meaningful distinctions between concepts presented in different parts of each lecture and tested  
679 by different quiz questions. But rather than characterizing their contents at too *broad* a semantic  
680 scale, the lack of specificity in this approach arises from considering too *narrow* a semantic scale:  
681 the sorts of concepts typically conveyed in course lectures and tested by quiz questions are not  
682 defined (and meaningful similarities and distinctions between them do not tend to emerge) at the  
683 level of individual words.

684 In other words, while the embedding spaces of more complex large language models afford  
685 low resolution at the scale of individual course lectures and questions because they “zoom out”  
686 too far, simpler word-matching measures afford low resolution because they “zoom in” too far. In  
687 this way, we view our approach as occupying a sort of “sweet spot” between simpler and more  
688 complex alternatives, in that it enables us to characterize the contents of course materials at the  
689 appropriate semantic scale where relevant concepts “come into focus.” Our approach enables us to  
690 accurately and consistently identify each question’s content in a way that matches it with specific  
691 content from the lectures and distinguishes it from other questions about similar content. In turn,  
692 this enables us to construct accurate predictions about participants’ knowledge of the conceptual  
693 content tested by individual quiz questions (Fig. 6).

694 Another application for large language models that does *not* require explicitly modeling the  
695 content of individual lectures or questions is to leverage ~~the~~ *these* models’ abilities to generate  
696 text. For example, generative text models like ChatGPT [56] and LLaMa [69] are already being  
697 used to build a new generation of interactive tutoring systems [e.g., 46]. Unlike the approach we  
698 have taken here, these generative text model-based systems do not explicitly model what learners

know, or how their knowledge changes over time with training. One could imagine building a hybrid system that combines the best of both worlds: a large language model that can *generate* text, combined with a smaller model that can *infer* what learners know and how their knowledge changes over time. Such a hybrid system could potentially be used to build the next generation of interactive tutoring systems that are able to adapt to learners' needs in real time, and **that are able to** provide more nuanced feedback about what learners know and what they do not know.

~~At the opposite end of the spectrum from large language models, one could also imagine simplifying some aspects of our LDA-based approach by computing simple word overlap metrics. For example, the Jaccard similarity between text  $A$  and  $B$  is computed as the number of unique words in the intersection of words from  $A$  and  $B$  divided by the number of unique words in the union of words from  $A$  and  $B$ . In a supplementary analysis (Supp. Fig. 10), we compared the LDA-based question-lecture matches we reported in Figure 4 with the Jaccard similarities between each question and each sliding window of text from the corresponding lecture. As shown in Supplementary Figure 10, this simple word matching approach does not appear to capture the same level of specificity as the LDA-based approach. Whereas the LDA-based approach often yields a clear peak in the time series of correlations between each question and the corresponding lecture, the Jaccard similarity-based approach does not. Furthermore, these LDA-based matches appear to capture conceptual overlaps between the questions and lectures (Supp. Tab. 3), whereas simple word matching does not. For example, one of the example questions examined in Supplementary Figure 10 asks "Which of the following occurs as a cloud of atoms gets more dense?" The LDA-based matches identify lecture timepoints where the relevant *topics* are discussed (e.g., when words like "cloud," "atom," "dense," etc., are mentioned *together*). The Jaccard similarity-based matches, on the other hand, are strong when *any* of these words are mentioned, even if they do not occur together.~~

~~We view our approach as occupying a sort of "sweet spot," between much larger language models and simple word matching-based approaches, that enables us to capture the relevant conceptual content of course materials at an appropriate semantic scale. Our approach enables us to accurately and consistently identify each question's content in a way that also matches up with~~

727 what is presented in the lectures. In turn, this enables us to construct accurate predictions about  
728 participants' knowledge of the conceptual content tested by held-out questions (Fig. 6).-

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

735 Within the past several years, ~~the~~ a global pandemic forced many educators to suddenly  
736 adapt to teaching remotely [36, 53, 65, 73]. This change in world circumstances is happening  
737 alongside (and perhaps accelerating) geometric growth in the availability of high-quality online  
738 courses from platforms such as Khan Academy [37], Coursera [74], EdX [39], and others [61].  
739 Continued expansion of the global internet backbone and improvements in computing hardware  
740 have also facilitated improvements in video streaming, enabling videos to be easily shared and  
741 viewed by increasingly large segments of the world's population. This exciting time for online  
742 course instruction provides an opportunity to re-evaluate how we, as a global community, educate  
743 ourselves and each other. For example, we can ask: what defines an effective course or training  
744 program? Which aspects of teaching might be optimized and/or augmented by automated tools?  
745 How and why do learning needs and goals vary across people? How might we lower barriers to  
746 receiving a high-quality education?

747 Alongside these questions, there is a growing desire to extend existing theories beyond the  
748 domain of lab testing rooms and into real classrooms [35]. In part, this has led to a recent  
749 resurgence of "naturalistic" or "observational" experimental paradigms that attempt to better  
750 reflect more ethologically valid phenomena that are more directly relevant to real-world situations  
751 and behaviors [54]. In turn, this has brought new challenges in data analysis and interpretation. A  
752 key step towards solving these challenges will be to build explicit models of real-world scenarios  
753 and how people behave in them (e.g., models of how people learn conceptual content from real-  
754 world courses, as in our current study). A second key step will be to understand which sorts

755 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 4,  
756 19, 51, 55, 58] might help to inform these models. A third major step will be to develop and  
757 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic  
758 paradigms.

759 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also  
760 relate to the notion of “theory of mind” of other individuals [27, 33, 50]. Considering others’ unique  
761 perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and  
762 communicate [59, 64, 68]. One could imagine future extensions of our work (e.g., analogous to  
763 the knowledge and learning maps shown in Fig. 8), that attempt to characterize how well-aligned  
764 different people’s knowledge bases or backgrounds are. In turn, this might be used to model how  
765 knowledge (or other forms of communicable information) flows not just between teachers and  
766 students, but between friends having a conversation, individuals on a first date, participants at  
767 a business meeting, doctors and patients, experts and non-experts, political allies or adversaries,  
768 and more. For example, the extent to which two people’s knowledge maps “match” or “align” in  
769 a given region of text embedding space might serve as a predictor of how effectively they will be  
770 able to communicate about the corresponding conceptual content.

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

777 **Materials and methods**

778 **Participants**

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

784 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09  
785 years). A total of 15 participants reported their gender as male and 35 participants reported their  
786 gender as female. A total of 49 participants reported their native language as "English" and 1  
787 reported having another native language. A total of 47 participants reported their ethnicity as  
788 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants  
789 reported their races as White (32 participants), Asian (14 participants), Black or African American  
790 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other  
791 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

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

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

832 ipants), Computing (2 participants), and other categories not listed in our survey (17 participants).  
833 Finally, we asked participants about in-person courses they had taken in different subject areas.  
834 They reported taking courses in Mathematics (38 participants), Science and engineering (37 partic-  
835 ipants), Arts and humanities (34 participants), Test preparation (27 participants), Economics and  
836 finance (26 participants), Computing (14 participants), College and careers (7 participants), ~~or~~and  
837 other courses not listed in our survey (6 participants).

## 838 Experiment

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

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

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

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

## 869 **Analysis**

### 870 **Statistics**

871 All of the statistical tests performed in our study were two-sided. The 95% confidence inter-  
872 vals we ~~reported report~~ for each correlation were estimated ~~by generating 10,000 from~~ bootstrap  
873 distributions of ~~correlation coefficients~~ 10,000 correlation coefficients obtained by sampling (with  
874 replacement) from the observed data.

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

876 We adapted an approach we developed in prior work [30] to embed each moment of the two  
877 lectures and each question in our pool in a common representational space. Briefly, our approach  
878 uses a topic model [Latent Dirichlet Allocation; 8] trained on a set of documents ~~, to discover a set~~ of  $k$  “topics” or “themes.” Formally, each topic is defined as a distribution of weights over words in  
879 the model’s vocabulary (i.e., the union of all unique words ~~, across all documents, excluding “stop~~ words.”). Conceptually, each topic is intended to give larger weights to words that are semantically  
880 related (as inferred from their tendency to co-occur in the same document). After fitting a topic  
881 model, each document in the training set, or any *new* document that contains at least some of  
882  
883

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

895 As illustrated in Figure 2A, we ~~start started~~ by building up a corpus of documents using over-  
896 lapping sliding windows that ~~span each video spanned each lecture~~’s transcript. Khan Academy  
897 provides professionally created, manual transcriptions of all lecture videos for closed captioning.  
898 However, such transcripts would not be readily available in all contexts to which our framework  
899 could potentially be applied. Khan Academy videos are hosted on the YouTube platform, which  
900 additionally provides automated captions. We opted to use these automated transcripts [which,  
901 in prior work, we have found to be of sufficiently near-human quality to yield reliable data in be-  
902 havioral studies; 75] when developing our framework in order to make it more directly extensible  
903 and adaptable by others in the future.

904 We fetched these automated transcripts using the youtube-transcript-api Python pack-  
905 age [17]. ~~The transcripts Each transcript~~ consisted of one timestamped line of text for every  
906 few seconds (mean: 2.34 s; standard deviation: 0.83 s) of spoken content in the ~~video~~lecture (i.e.,  
907 corresponding to each individual caption that would appear on-screen if viewing the lecture via  
908 YouTube, and when those lines would appear). We defined a sliding window length of (up to)  
909  $w = 30$  transcript lines ~~,~~ and assigned each window a timestamp corresponding to the midpoint  
910 between the timestamps for its first and last lines. This  $w$  parameter was chosen to match the  
911 same number of words per sliding window (rounded to the nearest whole word, and before pre-

912 processing) as the sliding windows we defined in our prior work [30; i.e., 185 words per sliding  
913 window].

914 These sliding windows ramped up and down in length at the beginning and end of each  
915 transcript, respectively. In other words, each transcript's first sliding window covered only its first  
916 line, the second sliding window covered the first two lines, and so on. This ensured that each line  
917 from the transcripts appeared in the same number ( $w$ ) of sliding windows. We next performed a  
918 series of standard text preprocessing steps: normalizing case, lemmatizing, removing punctuation  
919 and removing stop-words. We constructed our corpus of stop words by augmenting the Natural  
920 Language Toolkit [NLTK; 5] English stop word list with the following additional words, selected  
921 using one of the approaches suggested by Boyd-Graber et al. [9]: "actual," "actually," "also," "bit,"  
922 "could," "e," "even," "first," "follow," "following," "four," "let," "like," "mc," "really," "saw,"  
923 "see," "seen," "thing," and "two." This yielded sliding windows ~~with containing~~ an average of  
924 73.8 remaining words, and ~~lasting for spanning~~ an average of 62.22 seconds. We treated the text  
925 from each sliding window as a single "document" and combined these documents across the two  
926 ~~videoslectures~~ windows to create a single training corpus for the topic model.

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

934 We transformed each sliding window's text into a topic vector, and then used linear interpolation  
935 (independently for each topic dimension) to resample the resulting time series to one vector  
936 per second. We also used the fitted model to obtain topic vectors for each ~~quiz~~ question in our pool  
937 (see Supp. Tab. 1). Taken together, we obtained a *trajectory* for each ~~lecture~~ video, describing its path  
938 through topic space, and a single coordinate for each question (Fig. 2C). Embedding both ~~videos~~  
939 ~~lectures~~ and all of the questions using a common model enables us to compare the content from

940 different moments of ~~videos~~the lectures, compare the content across ~~videos~~lectures, and estimate  
941 potential associations between specific questions and specific moments of ~~video~~lecture content.

942 **Estimating dynamic knowledge traces**

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

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

945 where

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

946 and where mincorr and maxcorr are the minimum and maximum correlations between the topic  
947 vectors for any lecture timepoint and quiz question, taken over all timepoints in the given lecture  
948 ~~, and all five and all~~ questions *about* that lecture appearing on the given quiz. We also define  
949  $f(s, \Omega)$  as the  $s^{\text{th}}$  topic vector from the set of topic vectors  $\Omega$ . Here  $t$  indexes the set time series of  
950 lecture topic vectors ~~L~~, and  $i$  and  $j$  index the topic vectors of questions  $Q$  used to estimate the  
951 knowledge trace, ~~Q~~participant's knowledge. Note that "~~correct~~correct" denotes the set of indices  
952 of the questions the participant answered correctly on the given quiz.

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

961    **Generalized linear mixed models**

962    In the set of analyses reported in Figure 6, we assessed whether estimates of participants' knowledge  
963    at the embedding coordinates of individual quiz questions could be used to reliably predict their  
964    **ability abilities** to correctly answer those questions. In essence, we treated each question a given  
965    participant answered on a given quiz as a "lecture" consisting of a single timepoint, and used  
966    Equation 1 to estimate the participant's knowledge for its embedding coordinate based on their  
967    performance on all *other* questions they answered on that same quiz ("All questions"; Fig. 6,  
968    top row). Additionally, for each lecture-related question (i.e., excluding questions about general  
969    physics knowledge), we computed analogous knowledge estimates based on **all other two different**  
970    **subsets of** questions the participant answered on the same quiz**about**: (1) **all other questions about**  
971    the same lecture as the target question ("Within-lecture"; Fig. 6, middle rows), and (2) **all questions**  
972    **about** the other of the two lectures ("Across-lecture"; Fig. 6, bottom rows).

973    In **each** performing these analyses, our null hypothesis is that the knowledge estimates we  
974    compute based on the quiz questions' embedding coordinates do *not* provide useful information  
975    about participants' abilities to correctly answer those questions—in other words, that there is no  
976    meaningful difference (on average) between the knowledge estimates we compute for questions  
977    participants answered correctly versus incorrectly. Specifically, since we estimate knowledge for a  
978    given embedding coordinate as a weighted proportion-correct score (where each question's weight  
979    reflects its embedding-space distance from the target coordinate; see Eqn. 1), if these weights are  
980    uninformative (e.g., randomly distributed), then our estimates of participants' knowledge should  
981    be equivalent (on average) to the *unweighted* proportion of correctly answered questions used  
982    to compute them. In general, for a given participant and quiz, this expected null value (i.e.,  
983    **that participant's proportion-correct score on that quiz**) is the same for any coordinate in the  
984    embedding space (e.g., any lecture timepoint, quiz question, etc.). However, in the "All questions"  
985    and "Within-lecture" versions of the analyses shown in Figure 6, we estimate each participant's  
986    knowledge for each target question using *all other* questions (or *all other* questions about the same  
987    lecture) they answered on the same quiz. This introduces a systematic dependency between

988 a participant's success on a target question and their proportion-correct score on the remaining  
989 questions available to estimate their knowledge for it. For example, suppose a participant correctly  
990 answered  $n$  out of  $q$  questions on a given quiz. If we hold out a single *correctly* answered question as  
991 the target, the proportion of remaining questions answered correctly would be  $\frac{n-1}{q-1}$ , whereas if we  
992 hold out a single *incorrectly* answered question, the proportion of remaining questions answered  
993 correctly would be  $\frac{n}{q-1}$ . Thus, the proportion of correctly answered remaining questions (and  
994 therefore the null-hypothesized value of a knowledge estimate computed from them) is always  
995 *lower* for target questions a participant answered correctly than for those they answered incorrectly.

996

997 To correct for this baseline difference under our null hypothesis, we used a rebalancing  
998 procedure that ensured our knowledge estimates for questions each participant answered correctly  
999 and incorrectly were computed from the *same* proportion of correctly answered questions. For  
1000 each target question on a given participant's quiz, we first identified all remaining questions  
1001 with the opposite "correctness" label (i.e., if the target question was answered correctly, we  
1002 identified all remaining incorrectly answered questions, and vice versa). We then held out  
1003 each of these opposite-label questions, in turn, along with the target question, and estimated  
1004 the participant's knowledge for the target question using all *other* remaining questions. Since each  
1005 of these subsets of remaining questions was constructed by holding out one correctly answered  
1006 question and one incorrectly answered question from the participant's quiz responses, if the  
1007 participant correctly answered  $n$  out of  $q$  questions total, then their proportion-correct score on  
1008 each subset of questions used to estimate their knowledge would be  $\frac{n-1}{q-2}$ , regardless of whether they  
1009 answered the target question correctly or incorrectly. Finally, we averaged over these per-subset  
1010 knowledge estimates to obtain a rebalanced estimate of the participant's knowledge for the  
1011 target question that leveraged information from all remaining questions' embedding coordinates,  
1012 but whose expected value under our null hypothesis was the same as that of each individual  
1013 subset ( $\frac{n-1}{q-2}$ ). By equalizing the null-hypothesized values of knowledge estimates for correctly  
1014 and incorrectly answered questions, this procedure ensures that any meaningful relationships  
1015 we observe between participants' estimated knowledge for individual quiz questions and their

1016 abilities to correctly answer them reflect the predictive power of the embedding-space distances  
1017 we use to weight questions' contributions to the knowledge estimates, rather than an artifact of our  
1018 testing procedure. Note that if a participant answered all or no questions on a given quiz correctly,  
1019 their responses contained no opposite-label questions with which to perform this rebalancing,  
1020 and we therefore excluded their data from our analyses for that quiz. We used this rebalancing  
1021 procedure when constructing knowledge estimates for the "All questions" and "Within-lecture"  
1022 versions of the analyses shown in Figure 6, but not for the "Across-lecture" analyses as, in this  
1023 case, the target questions and the questions used to estimate participants' knowledge for them  
1024 were drawn from different subsets of quiz questions (those about one lecture, and those about the  
1025 other), and were therefore independent.

1026 In each version of this analysis (i.e., row in Fig. 6), and separately for each of the three quizzes  
1027 (i.e., column in Fig. 6), we then fit a generalized linear mixed model (GLMM) with a logistic link  
1028 function to the set of knowledge estimates for all questions (or all questions about a particular  
1029 lecture) that participants answered on the given quiz. We implemented these models in R using  
1030 the `lme4` package [3] and fit them following guidance from Bates et al. [2] and Matuschek et al.  
1031 [47]. Specifically, we initially fit each model with the maximal random effects structure afforded  
1032 by our design, which we identified as:

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

1033 where "accuracy" is a binary value indicating whether each target question was answered cor-  
1034 rectly or incorrectly, "knowledge" is estimated knowledge at each target question's embedding  
1035 coordinate, "participant" is a unique identifier assigned to each participant, and "question" is a  
1036 unique identifier assigned to each quiz question. For models we fit using knowledge estimates for  
1037 target questions about multiple content areas (i.e., in the "All questions" version of the analysis), we  
1038 also included an additional random effect term,  $(\text{knowledge} | \text{lecture})$ , where "lecture" is a cate-  
1039 gorical value denoting whether the target question was about *Four Fundamental Forces*, *Birth of Stars*,  
1040 or general physics knowledge. Note that with our coding scheme, identifiers for each question

1041 are implicitly nested within levels of lecture and so do not require explicit nesting in our model  
1042 formula. We then iteratively removed random effects from the maximal model until it successfully  
1043 converged with a full-rank (i.e., non-singular) full-rank random effects variance-covariance ma-  
1044 trix. We obtained the odds ratios reported in Figure 6 by exponentiating the estimated coefficient  
1045 for “knowledge” from each fitted model. Conceptually, these odds ratios represent how many  
1046 times greater the odds are that a given participant will answer a given question correctly if their  
1047 estimated knowledge for its embedding coordinate is 1, compared to if it is 0. We estimated 95%  
1048 confidence intervals for each odds ratio by generating 10,000 random subsamples (of full size, with  
1049 replacement) from the data used to fit each model, and refitting the models to each subsample to  
1050 obtain bootstrap distributions of 10,000 odds ratios.

1051 To assess the predictive value of our knowledge estimates, we compared each GLMM’s abil-  
1052 ity to discriminate between correctly and incorrectly answered explain participants’ success on  
1053 individual quiz questions to that of an analogous model that which assumed (as we assume under  
1054 our null hypothesis) that knowledge estimates for correctly and incorrectly answered questions  
1055 did not consider estimated knowledge systematically differ, on average. Specifically, we used the  
1056 same sets of observations with to which we fit each “full” model to fit a second “null” model, with  
1057 the formula:

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

1058 where “accuracy”, “participant”, and “question” are as defined above. As with our full models,  
1059 the null models we fit for the “All questions” version of the analysis for each quiz contained an  
1060 additional term,  $(1 | \text{lecture})$ , where “lecture” is as defined above with the same random effects  
1061 structure, but with the coefficient for the fixed effect of “knowledge” constrained to zero (i.e.,  
1062 we removed this term from the null model). We then compared each full model to its reduced  
1063 (null) equivalent using a likelihood-ratio test (LRT). Because the typical standard asymptotic  $\chi^2_d$   
1064 approximation of the null distribution for the LRT statistic ( $\lambda_{LR}$ ) is can be anti-conservative for  
1065 models that differ in their random slope terms finite sample sizes [26, 62, 67], we computed  $p$ -values  
1066 for these tests using a parametric bootstrap procedure [14, 28]. For each of 10,000 bootstraps, we

1067 used the fitted null model to simulate a sample of observations of equal size to our original sample.  
1068 We then re-fit both the null and full models to this simulated sample and compared them via an  
1069 LRT. This yielded a distribution of  $\lambda_{LR}$  statistics we may expect to observe ~~under given data that~~  
1070 ~~conforms to~~ our null hypothesis. ~~Following Ewens [22], we~~ We computed a corrected  $p$ -value for  
1071 our observed  $\lambda_{LR}$  as  $\frac{r}{n} \frac{r+1}{n+1}$ , where  $r$  is the number of simulated model comparisons that yielded a  $\lambda_{LR}$   
1072 greater than ~~or equal to~~ our observed value and  $n$  is the number of simulations we ran (~~10~~10,000).

1073 **Estimating the “smoothness” of knowledge**

1074 In the analysis reported in Figure 7A, we show how participants’ ability to correctly answer  
1075 quiz questions changes as a function of distance from a given correctly or incorrectly answered  
1076 reference question. We used a bootstrap-based approach to estimate the maximum distances over  
1077 which these proportions of correctly answered questions could be reliably distinguished from  
1078 participants’ overall average proportion of correctly answered questions.

1079 For each of 10,000 iterations, we drew a random subsample (with replacement) of 50 partic-  
1080 ipants from our dataset. Within each iteration, we first computed the 95% confidence interval  
1081 (CI) of the across-subsample-participants mean proportion correct on each of the three quizzes,  
1082 separately. To compute this interval for each quiz, we repeatedly (1,000 times) subsampled par-  
1083 ticipants (with replacement, from the outer subsample for the current iteration) and computed  
1084 the mean proportion correct of each of these inner subsamples. We then identified the 2.5<sup>th</sup> and  
1085 97.5<sup>th</sup> percentiles of the resulting distributions of 1,000 means. These three intervals (one for each  
1086 quiz) served as our thresholds for confidence that the proportion correct within a given distance  
1087 from a reference question was reliably different (at the  $p < 0.05$  significance level) from the average  
1088 proportion correct across all questions on the given quiz.

1089 Next, for each participant in the current subsample, and for each of the three quizzes they  
1090 completed (separately), we iteratively treated each of the 15 questions appearing on the given  
1091 quiz as the “reference” question. We constructed a series of concentric 15-dimensional “spheres”  
1092 centered on the reference question’s ~~embedding space~~ ~~embedding-space~~ coordinate, where each  
1093 successive sphere’s radius increased by 0.01 (correlation distance) between 0 and 2, inclusive (i.e.,

1094 tiling the range of possible correlation distances with 201 spheres in total). We then computed the  
1095 proportion of questions enclosed within each sphere that the participant answered correctly, and  
1096 averaged these per-radius proportion correct proportion-correct scores across reference questions  
1097 that were answered correctly, and those that were answered incorrectly. This resulted in two  
1098 number-of-spheres sequences of proportion-correct scores for each subsample participant and  
1099 quiz: one derived from correctly answered reference questions, and one derived from incorrectly  
1100 answered reference questions.

1101 We computed the across-subsample-participants mean proportion correct for each radius value  
1102 (i.e., sphere) and “correctness” of reference question. This yielded two sequences of proportion-  
1103 correct scores for each quiz, analogous to the blue and red lines displayed in Figure 7A, but for  
1104 the present subsample. For each quiz, we then found the minimum distance from the reference  
1105 question (i.e., sphere radius) at which each of these two sequences of per-radius proportion correct  
1106 proportion-correct scores intersected the 95% confidence interval for the overall proportion correct  
1107 (i.e., analogous to the black error bands in Fig. 7A).

1108 This resulted in two “intersection” distances for each quiz (for correctly answered and incor-  
1109 rectly answered reference questions). Repeating this full process for each of the 10,000 bootstrap  
1110 iterations output two distributions of intersection distances for each of the three quizzes. The  
1111 means and 95% confidence intervals for these distributions are plotted in Figure 7B.

## 1112 Creating knowledge and learning map visualizations

1113 An important feature of our approach is that, given a trained text embedding model and partici-  
1114 pants’ quiz performance on each quiz question, we can estimate their knowledge about *any* content  
1115 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-  
1116 tions, or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 5, 6, 7, 8,  
1117 and 9), we used Uniform Manifold Approximation and Projection [UMAP; 48, 49] to construct a  
1118 2D projection of the text embedding space. Whereas our main analyses used a 15-topic embedding  
1119 space, we used a 100-topic embedding space for these visualizations. This change in the number  
1120 of topics overcame an undesirable behavior in the UMAP embedding procedure, whereby embed-

1121 ding coordinates for the 15-topic model tended to be “clumped” into separated clusters, rather  
1122 than forming a smooth trajectory through the 2D space. When we increased the number of topics  
1123 to 100, the embedding coordinates in the 2D space formed a smooth trajectory through the space,  
1124 with substantially less clumping (Fig. 8). Creating a “map” by sampling this 100-dimensional  
1125 space at high resolution to obtain an adequate set of topic vectors spanning the embedding space  
1126 would be computationally intractable. However, sampling a 2D grid is trivial.

1127 At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing  
1128 the cross-entropy between the pairwise (clustered) distances between the observations in their  
1129 original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional  
1130 embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise  
1131 distances in the original high-dimensional space were defined as 1 minus the correlation between  
1132 each pair of coordinates, and pairwise distances in the low-dimensional embedding space were  
1133 defined as the Euclidean distance between each pair of coordinates.

1134 In our application, all of the coordinates we embedded were topic vectors, whose elements  
1135 are always non-negative and sum to one. Although UMAP is an invertible transformation at  
1136 the embedding locations of the original data, other locations in the embedding space will not  
1137 necessarily follow the same implicit “rules” as the original high-dimensional data. For example,  
1138 inverting an arbitrary coordinate in the embedding space might result in negative-valued vectors,  
1139 which are incompatible with the topic modeling framework. To protect against this issue, we  
1140 log-transformed the topic vectors prior to embedding them in the 2D space. When we inverted  
1141 the embedded vectors (e.g., to estimate topic vectors ~~or~~ for word clouds, as in Fig. 8C), we passed  
1142 the inverted (log-transformed) values through the exponential function to obtain a vector of non-  
1143 negative values, and normalized them to sum to one.

1144 After embedding both lectures’ topic trajectories and the topic vectors of every question, we  
1145 defined a rectangle enclosing the 2D projections of the lectures’ and quizzes’ embeddings. We then  
1146 sampled points from a regular  $100 \times 100$  grid of coordinates that evenly tiled this enclosing rectangle.  
1147 We sought to estimate participants’ knowledge (and learning, i.e., changes in knowledge) at each  
1148 of the resulting 10,000 coordinates.

1149 To generate our estimates, we placed a set of 39 radial basis functions (RBFs) throughout the  
1150 embedding space, centered on the 2D projections for each question (i.e., we included one RBF for  
1151 each question). At coordinate  $x$ , the value of an RBF centered on a question's coordinate  $\mu$  is  
1152 given by:

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

1153 The  $\lambda$  term in the RBF equation controls the “smoothness” of the function, where larger values  
1154 of  $\lambda$  result in smoother maps. In our implementation we used  $\lambda = 50$ . Next, we estimated the  
1155 “knowledge” at each coordinate,  $x$ , using:

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

1156 **Intuitively**, Equation 4 computes the weighted proportion of correctly answered questions, where  
1157 the weights are given by how nearby (in the 2D space) each question is to the  $x$ . We also defined  
1158 *learning maps* as the coordinate-by-coordinate differences between any pair of knowledge maps.  
1159 Intuitively, learning maps reflect the *change* in knowledge across two maps.

## 1160 Author contributions

1161 Conceptualization: P.C.F., A.C.H., and J.R.M. Methodology: P.C.F., A.C.H., and J.R.M. Software:  
1162 P.C.F. Validation: P.C.F. Formal analysis: P.C.F. Resources: P.C.F., A.C.H., and J.R.M. Data curation:  
1163 P.C.F. Writing (original draft): J.R.M. Writing (review and editing): P.C.F., A.C.H., and J.R.M. Visu-  
1164 alization: P.C.F. and J.R.M. Supervision: J.R.M. Project administration: P.C.F. Funding acquisition:  
1165 J.R.M.

## 1166 Data availability

1167 All of the data analyzed in this manuscript may be found at <https://github.com/ContextLab/effic->  
1168 ient-learning-khan.

1169 **Code availability**

1170 All of the code for running our experiment and carrying out the analyses may be found at  
1171 <https://github.com/ContextLab/efficient-learning-khan>.

1172 **Acknowledgements**

1173 We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of  
1174 this study, and assistance with data collection efforts from Will Baxley, Max Bluestone, Daniel  
1175 Carstensen, Kunal Jha, Caroline Lee, Lucy Owen, Xinming Xu, and Kirsten Ziman. Our work was  
1176 supported in part by NSF CAREER Award Number 2145172 to J.R.M. The content is solely the  
1177 responsibility of the authors and does not necessarily represent the official views of our supporting  
1178 organizations. The funders had no role in study design, data collection and analysis, decision to  
1179 publish, or preparation of the manuscript.

1180 **References**

- 1181 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,  
1182 56:149–178.
- 1183 [2] Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015a). Parsimonious mixed models. *arXiv*,  
1184 1506.04967.
- 1185 [3] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015b). Fitting linear mixed-effects models  
1186 using lme4. *Journal of Statistical Software*, 67(1):1–48.
- 1187 [4] Bevilacqua, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and  
1188 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom  
1189 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.

- 1190 [5] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text*  
1191 *with the natural language toolkit*. Reilly Media, Inc.
- 1192 [6] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-  
1193 dren: distinguishing response flexibility from conceptual flexibility; the protracted development  
1194 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 1195 [7] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*  
1196 *Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing  
1197 Machinery.
- 1198 [8] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*  
1199 *Learning Research*, 3:993–1022.
- 1200 [9] Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models:  
1201 problems, diagnostics, and improvements. In Airolidi, E. M., Blei, D. M., Erosheva, E. A., and  
1202 Fienberg, S. E., editors, *Handbook of Mixed Membership Models and Their Applications*. CRC Press.
- 1203 [10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,  
1204 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,  
1205 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,  
1206 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,  
1207 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 1208 [11] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the  
1209 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 1210 [12] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-  
1211 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal  
1212 sentence encoder. *arXiv*, 1803.11175.
- 1213 [13] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual  
1214 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.

- 1215 [14] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge  
1216 Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- 1217 [15] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).  
1218 Evidence for a new conceptualization of semantic representation in the left and right cerebral  
1219 hemispheres. *Cortex*, 40(3):467–478.
- 1220 [16] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).  
1221 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,  
1222 41(6):391–407.
- 1223 [17] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 1225 [18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep  
1226 bidirectional transformers for language understanding. *arXiv*, 1810.04805.
- 1227 [19] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,  
1228 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony  
1229 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 1230 [20] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 1231 [21] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of*  
1232 *Experimental Psychology: General*, 115:155–174.
- 1233 [22] Ewens, W. J. (2003). On Estimating *P* Values by Monte Carlo Methods. *American Journal of*  
1234 *Human Genetics*, 72(2):496–498.
- 1235 [23] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*  
1236 *Transactions of the Royal Society A*, 222(602):309–368.
- 1237 [24] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.  
1238 *School Science and Mathematics*, 100(6):310–318.

- 1239 [25] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather  
1240 prediction” task? individual variability in strategies for probabilistic category learning. *Learning*  
1241 and *Memory*, 9:408–418.
- 1242 [26] Goldman, N. and Whelan, S. (2000). Statistical Tests of Gamma-Distributed Rate Heterogeneity  
1243 in Models of Sequence Evolution in Phylogenetics. *Molecular Biology and Evolution*, 17(6):975–978.
- 1244 [27] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*  
1245 *Cognition and Development*, 13(1):19–37.
- 1246 [28] Halekoh, U. and Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric  
1247 Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *Journal of*  
1248 *Statistical Software*, 59(9):1–32.
- 1249 [29] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual  
1250 learning, pages 212–221. Sage Publications.
- 1251 [30] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-  
1252 ioral and neural signatures of transforming experiences into memories. *Nature Human Behaviour*,  
1253 5:905–919.
- 1254 [31] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge au-  
1255 tomatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*,  
1256 9:doi.org/10.3389/fpsyg.2018.00133.
- 1257 [32] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-  
1258 peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–  
1259 4008.
- 1260 [33] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating  
1261 empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 1262 [34] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.  
1263 Columbia University Press.

- 1264 [35] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,  
1265 326(7382):213–216.
- 1266 [36] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).  
1267 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International  
1268 Journal of Environmental Research and Public Health*, 18(5):2672.
- 1269 [37] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 1270 [38] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 1271 [39] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.  
1272 *The Chronicle of Higher Education*, 21:1–5.
- 1273 [40] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic  
1274 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,  
1275 104:211–240.
- 1276 [41] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic  
1277 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.
- 1278 [42] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of  
1279 Educational Studies*, 53(2):129–147.
- 1280 [43] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”  
1281 function? *Psychological Review*, 128(4):711–725.
- 1282 [44] Manning, J. R. (2023). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,  
1283 *Handbook of Human Memory*. Oxford University Press.
- 1284 [45] Manning, J. R. and Kahana, M. J. (2012). Interpreting semantic clustering effects in free recall.  
1285 *Memory*, 20(5):511–517.
- 1286 [46] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension  
1287 for adding LLM-driven chatbots to interactive notebooks. [https://github.com/ContextLab/  
1288 chatify](https://github.com/ContextLab/chatify).

- 1289 [47] Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type i error  
1290 and power in linear mixed models. *Journal of Memory and Language*, 94:305–315.
- 1291 [48] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and  
1292 projection for dimension reduction. *arXiv*, 1802(03426).
- 1293 [49] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold  
1294 Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 1295 [50] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of  
1296 mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 1297 [51] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,  
1298 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to  
1299 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 1300 [52] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-  
1301 tations in vector space. *arXiv*, 1301.3781.
- 1302 [53] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications  
1303 from a national survey of language educators. *System*, 97:102431.
- 1304 [54] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of  
1305 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1306 [55] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).  
1307 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective  
1308 Neuroscience*, 17(4):367–376.
- 1309 [56] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 1310 [57] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models.  
1311 *arXiv*, 2208.02957.

- 1312 [58] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG  
1313 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,  
1314 7:43916.
- 1315 [59] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*  
1316 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 1317 [60] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.  
1318 *Biological Cybernetics*, 45(1):35–41.
- 1319 [61] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in  
1320 higher education: unmasking power and raising questions about the movement’s democratic  
1321 potential. *Educational Theory*, 63(1):87–110.
- 1322 [62] Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for a zero random  
1323 effect variance or polynomial regression in additive and linear mixed models. *Computational*  
1324 *Statistics & Data Analysis*, 52(7):3283–3299.
- 1325 [63] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter  
1326 Student conceptions and conceptual learning in science. Routledge.
- 1327 [64] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-  
1328 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*  
1329 *tion in Nursing*, 22:32–42.
- 1330 [65] Shim, T. E. and Lee, S. Y. (2020). College students’ experience of emergency remote teaching  
1331 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 1332 [66] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for  
1333 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*  
1334 *Mathematics Education*, 35(5):305–329.
- 1335 [67] Snijders, T. A. B. and Bosker, R. (2011). More powerful tests for variance parameters. In

- 1336      *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, chapter 6, pages  
1337      94–108. Sage Publications, 2nd edition.
- 1338      [68] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal  
1339      Medicine*, 21:524–530.
- 1340      [69] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,  
1341      Goyal, N., Hambro, E., Azhar, F., Rodriguz, A., Joulin, A., Grave, E., and Lample, G. (2023).  
1342      LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 1343      [70] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Pio-  
1344      ntkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust  
1345      detection of AI-generated texts. *arXiv*, 2306.04723.
- 1346      [71] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold?  
1347      distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of  
1348      the Cognitive Science Society*, 43(43).
- 1349      [72] Viswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and  
1350      Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing  
1351      Systems*.
- 1352      [73] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned  
1353      from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 1354      [74] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from  
1355      free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 1356      [75] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is  
1357      automatic speech-to-text transcription ready for use in psychological experiments? *Behavior  
1358      Research Methods*, 50:2597–2605.