

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

⁴ Paxton C. Fitzpatrick¹, Andrew C. Heusser^{1, 2}, and Jeremy R. Manning^{1,*}

¹Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

²Akili Interactive Labs

Boston, MA 02110, USA

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

⁵ **Abstract**

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

¹⁷ **Keywords:** education, learning, knowledge, concepts, natural language processing

¹⁸ **Introduction**

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

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

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

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

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

73 ~

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

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

93 Our primary research goal is to advance our understanding of what it means to acquire deep,
94 real-world conceptual knowledge. Traditional laboratory approaches to studying learning and
95 memory (e.g., list-learning studies) often draw little distinction between memorization and under-
96 standing. Instead, these studies typically focus on whether information is effectively encoded or
97 retrieved, rather than whether the information is *understood*. Approaches to studying conceptual
98 learning, such as category learning experiments, can begin to investigate the distinction between
99 memorization and understanding, often by training participants to distinguish arbitrary or random
100 features in otherwise meaningless categorized stimuli [1, 17, 18, 21, 26, 52]. However the objective

101 of real-world training, or learning from life experiences more generally, is often to develop new
102 knowledge that may be applied in *useful* ways in the future. In this sense, the gap between modern
103 learning theories and modern pedagogical approaches that inform classroom learning strategies is
104 enormous: most of our theories about *how* people learn are inspired by experimental paradigms
105 and models that have only peripheral relevance to the kinds of learning that students and teachers
106 actually seek [23, 36]. To help bridge this gap, our study uses course materials from real on-
107 line courses to inform, fit, and test models of real-world conceptual learning. We also provide a
108 demonstration of how our models can be used to construct “maps” of what students know, and
109 how their knowledge changes with training. In addition to helping to visually capture knowledge
110 (and changes in knowledge), we hope that such maps might lead to real-world tools for improving
111 how we educate. Taken together, our work shows that existing course materials and evaluative
112 tools like short multiple-choice quizzes may be leveraged to gain highly detailed insights into what
113 students know and how they learn.

114 Results

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

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

129 We asked participants in our study to complete brief multiple-choice quizzes before, between,
 130 and after watching two lecture videos from the Khan Academy [31] platform (Fig. 1). The first
 131 lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics:
 132 gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*,
 133 provided an overview of our current understanding of how stars form. We selected these particular
 134 lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad
 135 audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training
 136 on participants' abilities to learn from the lectures. To this end, we selected two introductory
 137 videos that were intended to be viewed at the start of students' training in their respective content
 138 areas. Second, we wanted the two lectures to have some related content, so that we could test
 139 our approach's ability to distinguish similar conceptual content. To this end, we chose two videos
 140 from the same Khan Academy course domain, "Cosmology and Astronomy." Third, we sought to
 141 minimize dependencies and specific overlap between the videos. For example, we did not want
 142 participants' abilities to understand one video to (directly) influence their abilities to understand the
 143 other. To satisfy this last criterion, we chose videos from two different lecture series (Lectures 1 and
 144 2 were from the "Scale of the Universe" and "Stars, Black Holes, and Galaxies" series, respectively).

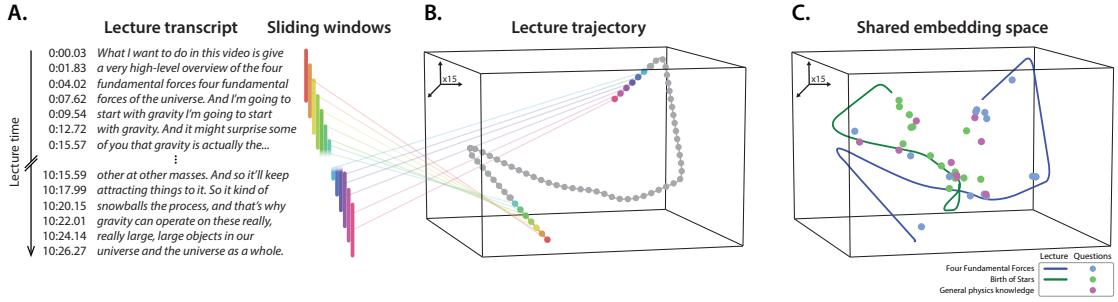


Figure 2: Modeling course content. **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The full set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding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

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

153 To study in detail how participants’ conceptual knowledge changed over the course of the
 154 experiment, we first sought to model the conceptual content presented to them at each moment
 155 throughout each of the two lectures. We adapted an approach we developed in prior work [24]
 156 to identify the latent themes in the lectures using a topic model [6]. Briefly, topic models take
 157 as input a collection of text documents, and learn a set of “topics” (i.e., latent themes) from their
 158 contents. Once fit, a topic model can be used to transform arbitrary (potentially new) documents
 159 into sets of “topic proportions,” describing the weighted blend of learned topics reflected in their

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

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



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

indicating that the topic model also exhibits some degree of specificity. The full set of pairwise comparisons between average topic weights for the lectures and question sets is reported in Supplementary Figure 2.

Another, more sensitive, way of summarizing the conceptual content of the lectures and questions is to look at *variability* in how topics are weighted over time and across different questions (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “information” [19] the lecture (or question set) reflects about that topic. For example, suppose a given topic is weighted on heavily throughout a lecture. That topic might be characteristic of some aspect or property of the lecture *overall* (conceptual or otherwise), but unless the topic’s weights changed in meaningful ways over time, the topic would be a poor indicator of any *specific* conceptual content in the lecture. We therefore also compared the variances in topic weights (across time or questions) between the lectures and questions. The variability in topic expression (over time and across questions) was similar for the Lecture 1 video and questions ($r(13) = 0.824, p < 0.001$, 95% CI = [0.696, 0.973]) and the Lecture 2 video and questions ($r(13) = 0.801, p < 0.001$, 95%

202 CI = [0.539, 0.958]). Simultaneously, as reported in Figure 3B, the variability in topic expression
203 across *different* videos and lecture-specific questions (i.e., Lecture 1 video vs. Lecture 2 questions;
204 Lecture 2 video vs. Lecture 1 questions) were negatively correlated, and neither video’s topic
205 variability was reliably correlated with the topic variability across general physics knowledge
206 questions. Taken together, the analyses reported in Figure 3 and Supplementary Figure 2 indicate
207 that a topic model fit to the videos’ transcripts can also reveal correspondences (at a coarse scale)
208 between the lectures and questions.

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

227 The ability to quantify how much each question is “asking about” the content from each moment
228 of the lectures could enable high-resolution insights into participants’ knowledge. Traditional
229 approaches to estimating how much a student “knows” about the content of a given lecture entail

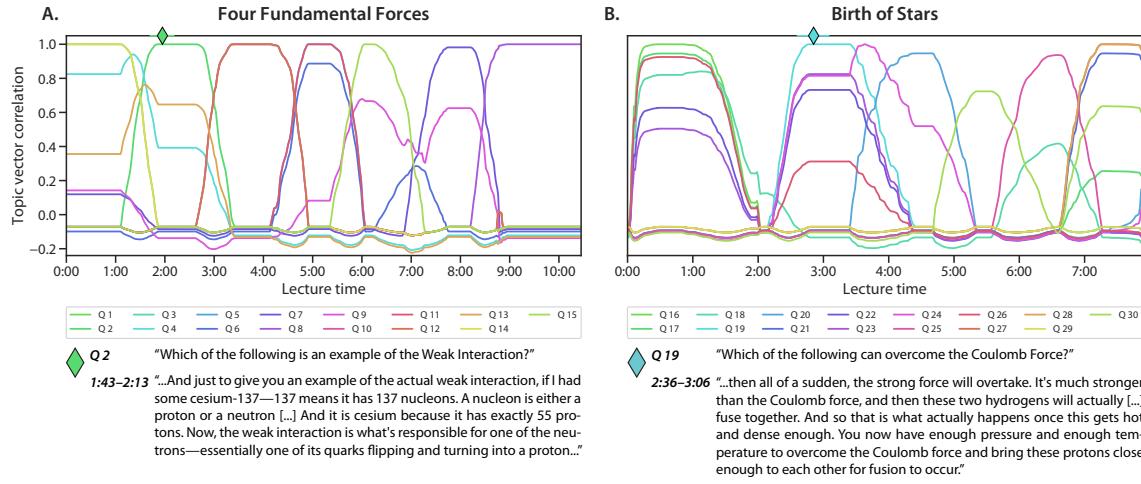


Figure 4: Which parts of each lecture are captured by each question? Each panel displays [timeseries-time series](#) plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions;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

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

We developed a simple formula (Eqn. 1) for using a participant’s responses to a small set

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

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

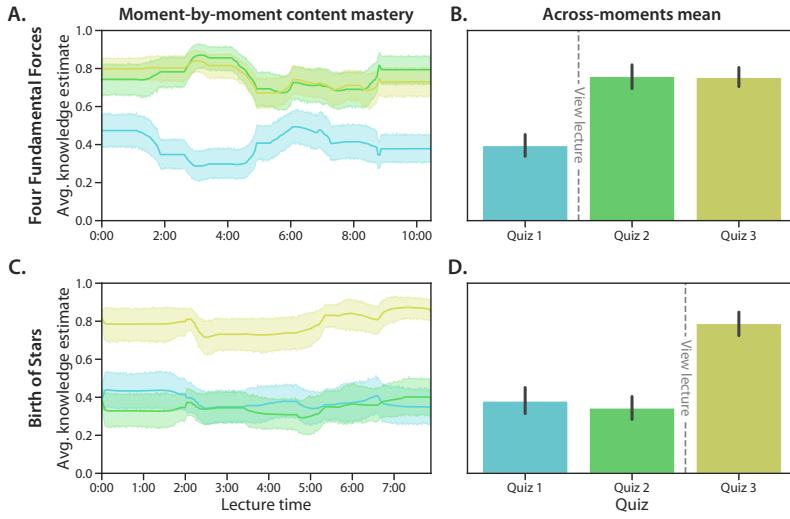


Figure 5: Estimating moment-by-moment knowledge acquisition. **A. Moment-by-moment knowledge about the Four Fundamental Forces** **Estimating knowledge about the content presented at each moment of each lecture.** **A. Knowledge about the time-varying content of Four Fundamental Forces.** Estimating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

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

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

289 We carried out these analyses in three different ways. First, we used all (but one) of the
290 questions from a given quiz (and participant) to predict knowledge at the embedding coordinate
291 of a held-out question (“All questions” in Fig. 6). This test was intended to serve as an overall
292 baseline for the predictive power of our approach. Second, we used questions about one lecture
293 to predict knowledge at the embedding coordinate of a held-out question about the *other* lecture,
294 from the same quiz and participant (“Across-lecture” in Fig. 6). This test was intended to test
295 the generalizability of our approach by asking whether our knowledge predictions held across the
296 content areas of the two lectures. Third, we used questions about one lecture to predict knowledge
297 at the embedding coordinate of a held-out question about the *same* lecture, from the same quiz and
298 participant (“Within-lecture” in Fig. 6). This test was intended to test the specificity of our approach

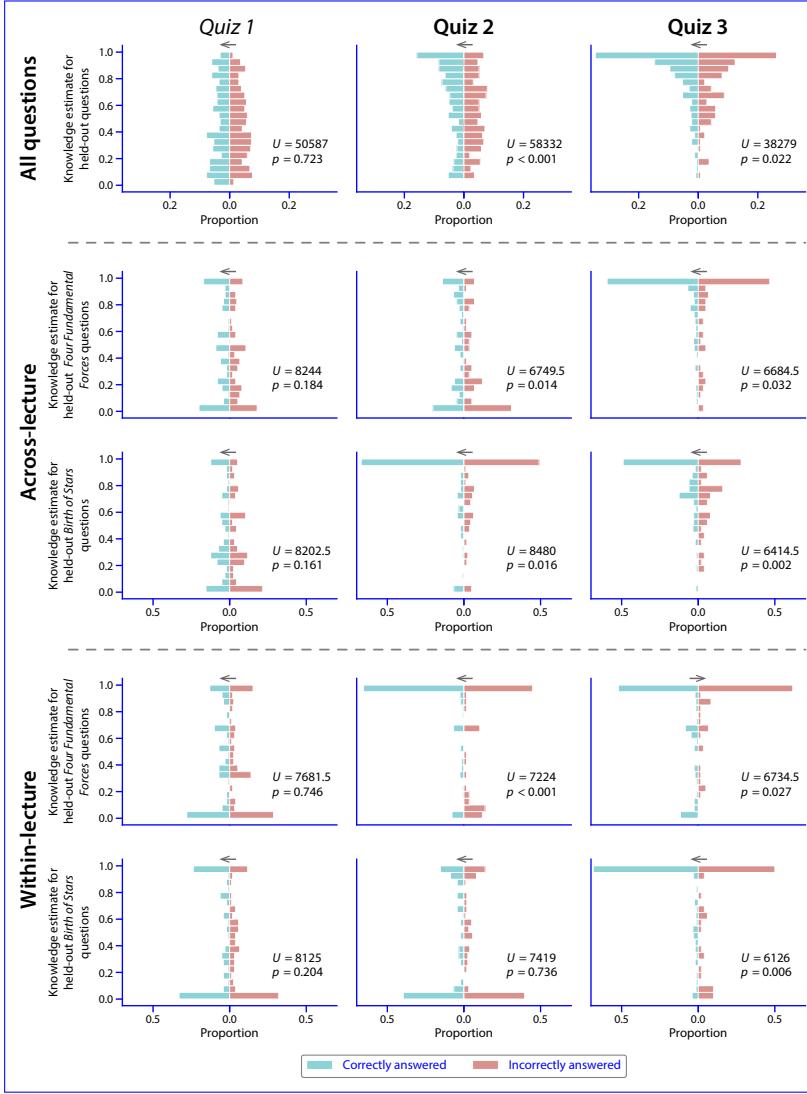


Figure 6: Estimating knowledge at the embedding coordinates of held-out questions. Predicting knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel/column), we plot the distributions of predicted knowledge at the embedding coordinates of each held-out correctly (blue) or incorrectly (red) answered question. The Mann-Whitney U-tests reported in each panel are between the distributions of estimated predicted knowledge at the coordinates of correctly versus and incorrectly answered held-out questions. In the top row ("All questions"), we used all quiz questions (from each quiz, for each participant) except one to predict knowledge at the held-out question's embedding coordinate. In the middle rows ("Across-lecture"), we used all questions about one lecture to predict knowledge at the embedding coordinate of a held-out question about the other lecture. In the bottom row ("Within-lecture"), we used all but one question about one lecture to predict knowledge at the embedding coordinate of a held-out question about the same lecture. We repeated each of these analyses using all possible held-out questions for each quiz and participant. The arrows at the tops of each panel indicate whether the average predicted knowledge was higher for held-out correctly answered (left) or incorrectly answered (right) questions.

299 by asking whether our knowledge predictions could distinguish between questions about different
300 content covered by the same lecture. We repeated each of these analyses using all possible held-out
301 questions for each quiz and participant.

302 For the initial quizzes participants took (prior to watching either lecture), participants' estimated
303 predicted knowledge tended to be low overall, and relatively unstructured (Fig. 6, left panel).
304 When we held out individual questions and estimated predicted their knowledge at the held-out
305 questions' embedding coordinates, we found no reliable differences in the estimates-predictions
306 when the held-out question had been correctly versus incorrectly answered ($t(633) = 0.577, p = 0.564$).
307 This. This "null" effect persisted when we used all of the Quiz 1 questions from a given participant
308 to predict a held-out question ("All questions"; $U = 50587, p = 0.723$), when we used questions
309 from one lecture to predict knowledge at the embedding coordinate of a held-out question
310 about the other lecture ("Across-lecture"; predicting knowledge for held-out Four Fundamental
311 Forces Questions using Birth of Stars questions: $U = 8244, p = 0.184$; predicting knowledge for
312 held-out Birth of Stars questions: $U = 8202.5, p = 0.161$), and when we used questions from
313 one lecture to predict knowledge at the embedding coordinate of a held-out question about
314 the same lecture ("Within-lecture"; Four Fundamental Forces: $U = 7681.5, p = 0.746$; Birth of Stars:
315 $U = 8125, p = 0.204$). We believe that this reflects a floor effect: when knowledge is low every-
316 where, there is little signal to differentiate between what is known versus unknown. After watching
317 the first lecture, estimated

318 After watching Four Fundamental Forces, predicted knowledge for held-out correctly answered
319 questions questions that were answered correctly (from the second quiz; Fig. 6, middle panel) exhibited a significant positive shift relative to held-out incorrectly answered questions questions
320 that were answered incorrectly. This held when we included all questions in the analysis ($U = 58332, p < 0.001$),
321 when we predicted knowledge across-lectures ($t(633) = 3.961, p < 0.001$). This second quiz provides
322 the maximally sensitive test for our knowledge predictions, since (if knowledge is estimated
323 accurately) participants' Quiz 2 responses should demonstrate specific knowledge about Lecture 1
324 content, but knowledge about Lecture 2 and general physics concepts should be roughly unchanged
325 from before they watched Lecture 1. After watching the second lecture, estimated knowledge

327 *Four Fundamental Forces*: $U = 6749.5, p = 0.014$; *Birth of Stars*: $U = 8480, p = 0.016$), and when we
328 predicted knowledge at the embedding coordinates of held-out *Four Fundamental Forces* questions
329 using other *Four Fundamental Forces* questions from the same quiz and participant ($U = 7224, p < 0.001$).
330 This difference did *not* hold for within-lecture knowledge predictions at knowledge at embedding
331 space coordinates of *Birth of Stars* questions ($U = 7419, p = 0.739$). Again, we suggest that this
332 might reflect a floor effect whereby, at that point in the participants' training, their knowledge
333 about the content of the *Birth of Stars* material is relatively low everywhere in that region of text
334 embedding space.

335 Finally, after watching *Birth of Stars*, predicted knowledge for held-out correctly answered
336 questions (from the third quiz; Fig. 6, right panel) was higher than for held-out incorrectly
337 answered questions. This held when we included all questions in the analysis ($U = 38279, p = 0.022$),
338 when we carried out across-lecture predictions (*Four Fundamental Forces*: $U = 6684.5, p = 0.032$;
339 *Birth of Stars*: $U = 6414.5, p = 0.002$), and when we carried out within-lecture knowledge
340 predictions for held-out *Birth of Stars* questions using other *Birth of Stars* questions from the same
341 quiz and participant ($U = 6126, p = 0.006$). However, we found the *opposite* effect when we carried
342 out within-lecture knowledge predictions for *all* questions exhibited a positive shift. However, the
343 estimated knowledge for held-out *Four Fundamental Forces* questions using other *Four Fundamental*
344 *Forces* questions from the same quiz and participant ($U = 6734, p = 0.027$). Specifically, on Quiz
345 3, our knowledge predictions for held-out correctly answered questions remained greater than
346 that for held-out incorrectly answered questions ($t(628) = 2.045, p = 0.041$) about *Four Fundamental*
347 *Forces* were reliably *lower* than those for their incorrectly answered counterparts. Speculatively, we
348 suggest that this may reflect participants forgetting some of the *Four Fundamental Forces* content. If
349 this forgetting happens in a relatively "random" way (with respect to spatial distance within the
350 text embedding space), then it could explain why some held-out questions about *Four Fundamental*
351 *Forces* were answered incorrectly, even if questions at nearby coordinates (i.e., about similar content)
352 were answered correctly. This might lead our approach to over-estimate knowledge for held-out
353 questions about "forgotten" knowledge that participants answered incorrectly. This third contrast
354 reflects a ceiling effect: when knowledge is relatively high everywhere, the signal differentiating

355 what is known versus unknown is relatively weak. Taken together, this set of analyses demonstrates
356 that our knowledge prediction framework is most informative when participants exhibit variability
357 in their knowledge of the content captured by the text embedding model the results in Figure 6
358 indicate that our approach can reliably predict acquired knowledge (especially about recently
359 learned content), and that the knowledge predictions are generalizable across the content areas
360 spanned by the two lectures, while also specific enough to distinguish between questions about
361 more subtly different content within the same lecture.

362 That the knowledge predictions derived from the text embedding space reliably distinguish
363 between held-out correctly versus incorrectly answered questions (Fig. 6) suggests that spatial
364 relationships within this space can help explain what participants know. But how far does this
365 explanatory power extend? For example, suppose we know that a participant correctly answered a
366 question at embedding coordinate x . As we move farther away from x in the embedding space, how
367 does the likelihood that the participant knows about the content at a given location “fall off” with
368 distance? Conversely, suppose the participant instead answered that same question *incorrectly*.
369 Again, as we move farther away from x in the embedding space, how does the likelihood that the
370 participant does *not* know about a coordinate’s content change with distance? We reasoned that,
371 assuming our embedding space is capturing something about how individuals actually organize
372 their knowledge, a participant’s ability to answer questions embedded very close to x should
373 tend to be similar to their ability to answer the question embedded *at* x . Whereas at another
374 extreme, once we reach some sufficiently large distance from x , our ability to infer whether or
375 not a participant will correctly answer a question based on their ability to answer the question
376 at x should be no better than guessing based on their *overall* proportion of correctly answered
377 questions. In other words, beyond the maximum distance at which the participant’s ability to
378 answer the question at x is informative of their ability to answer a second question at location y ,
379 then guessing the outcome at y based on x should be no more successful than guessing based on a
380 measure that does not consider embedding space distance.

381 With these ideas in mind, we asked: conditioned on answering a question correctly, what
382 proportion of all questions (within some radius, r , of that question’s embedding coordinate)

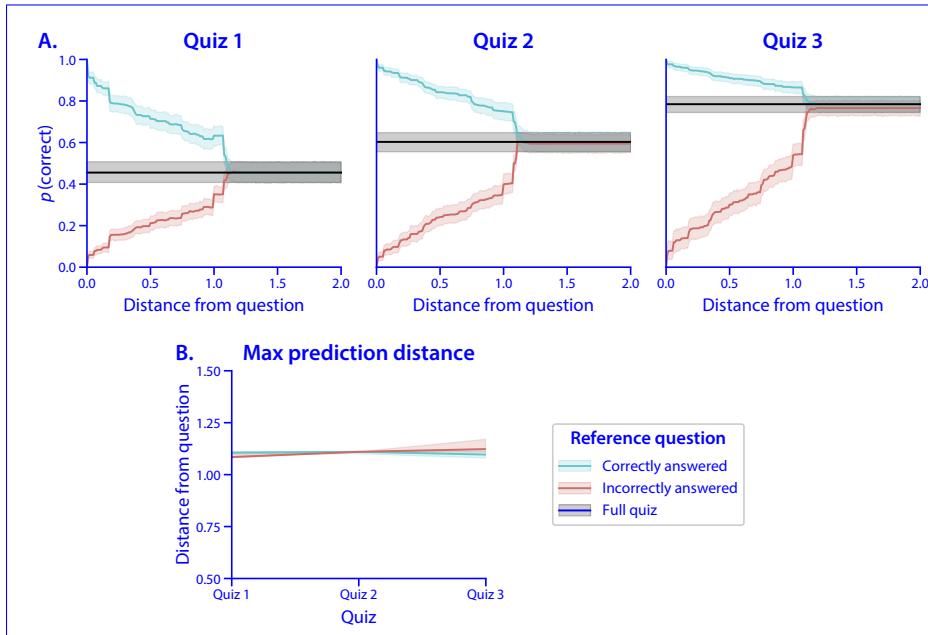


Figure 7: Knowledge falls off gradually in text embedding space. A. Performance versus distance. For each participant, for each correctly answered question (blue) or incorrectly answered question (red), we computed the proportion of correctly answered questions within a given distance of that question's 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.

383 were answered correctly? We plotted this proportion as a function of r . Similarly, we could
 384 ask, conditioned on answering a question incorrectly, how the proportion of correct responses
 385 changed with r . As shown in Figure 7, we found that quiz performance falls off smoothly with
 386 distance, and the “rate” of the falloff does not appear to change across the different quizzes, as
 387 measured by the distance at which performance becomes statistically indistinguishable from a
 388 simple proportion correct score (see *Estimating the “smoothness” of knowledge*). This suggests that,
 389 at least within the region of text embedding space covered by the questions our participants
 390 answered (and as characterized using our topic model), the rate at which knowledge changes

391 with distance is relatively constant, even as participants' overall level of knowledge varies across
392 quizzes or regions of the embedding space.

393 Knowledge estimates need not be limited to the content of the lectures. As illustrated in
394 Figure 8, our general approach to estimating knowledge from a small number of quiz questions
395 may be extended to *any* content, given its text embedding coordinate. To visualize how knowledge
396 “spreads” through text embedding space to content beyond the lectures participants watched, we
397 first fit a new topic model to the lectures’ sliding windows with ~~(up to)~~ $k = 100$ topics. Conceptually,
398 increasing the number of topics used by the model functions to increase the “resolution” of the
399 embedding space, providing a greater ability to estimate knowledge for content that is highly
400 similar to (but not precisely the same as) that contained in the two lectures. We note that we
401 used these 2D maps solely for visualization; all relevant comparisons, distance computations, and
402 statistical tests we report above were carried out in the original 15-dimensional space, using the
403 15-topic model. Aside from increasing the number of topics from 15 to 100, all other procedures
404 and model parameters were carried over from the preceding analyses. As in our other analyses,
405 we resampled each lecture’s topic trajectory to 1 Hz and projected each question into a shared text
406 embedding space.

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

416 Several features of the resulting knowledge maps are worth noting. The average knowledge
417 map estimated from Quiz 1 responses (Fig. 8A, leftmost map) shows that participants tended to
418 have relatively little knowledge about any parts of the text embedding space (i.e., the shading is

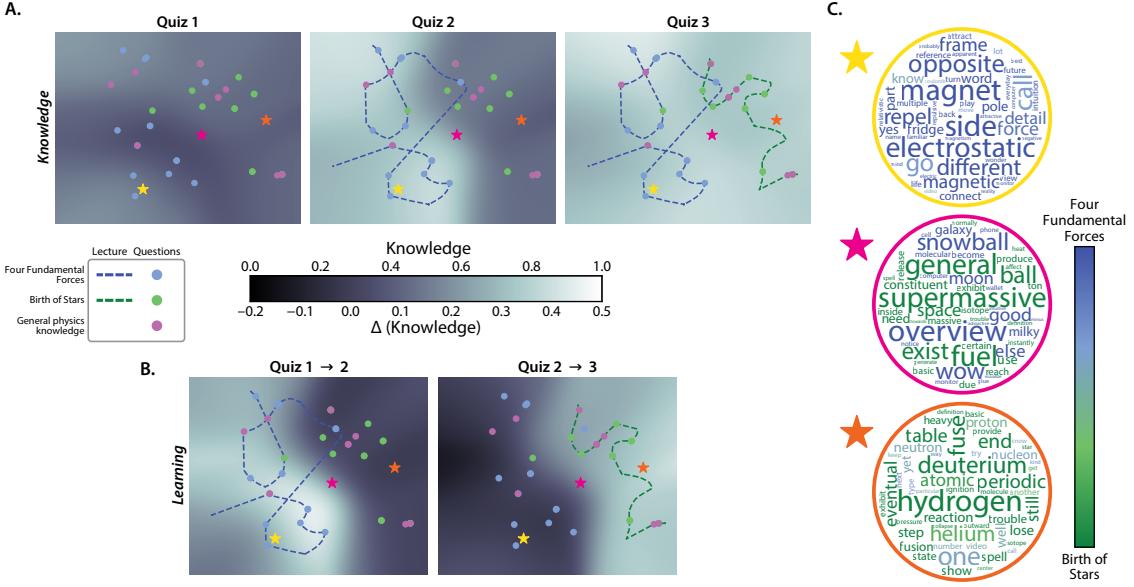


Figure 8: Mapping out the geometry of knowledge and learning. **A.** Average “knowledge maps” estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*). The topic trajectories of the two lectures are indicated by dotted lines (blue: Lecture 1; green: Lecture 2), and the coordinates of each question are indicated by dots (light blue: Lecture 1-related; light green: Lecture 2-related; purple: general physics knowledge). Each map reflects an average across all participants. For individual participants’ maps, see Supplementary Figures 7, 8, and 9. **B.** Average “learning maps” estimated between each successive pair of quizzes. The learning maps follow the same general format as the knowledge maps in Panel A, but here the shading at each coordinate indicates the *difference* between the corresponding coordinates in the indicated pair of knowledge maps—i.e., how much the estimated knowledge “changed” between the two quizzes. Each map reflects an average across all participants. For individual participants’ maps, see Supplementary Figures 10 and 11. **C.** Word clouds for sampled points in topic space. Each word cloud displays the weighted blend of words underlying the topic proportions represented at the corresponding colored star’s location on the maps. In each word cloud, the words’ relative sizes correspond to their relative weights at the starred location, and their colors indicate their relative weights in *the Four Fundamental Forces* (blue) versus *Birth of Stars* (green) lectures, on average, across all timepoints’ topic vectors.

419 relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked
420 increase in knowledge on the left side of the map (around roughly the same range of coordinates
421 traversed by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words,
422 participants' estimated increase in knowledge is localized to conceptual content that is nearby (i.e.,
423 related to) the content from the lecture they watched prior to taking Quiz 2. This localization is
424 non-trivial: these knowledge estimates are informed only by the embedded coordinates of the
425 *quiz questions*, not by the embeddings of either lecture (see Eqn. 4). Finally, the knowledge map
426 estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region
427 surrounding the embedding of the *Birth of Stars* lecture participants watched immediately prior to
428 taking Quiz 3.

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

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

⁴⁴⁷ top-weighted words at the example coordinate between the two lectures' embeddings show a
⁴⁴⁸ roughly even mix of words most strongly associated with each lecture.

⁴⁴⁹ Discussion

⁴⁵⁰ We developed a computational framework that uses short multiple-choice quizzes to gain nuanced
⁴⁵¹ insights into what learners know and how their knowledge changes with training. First, we show
⁴⁵² that our approach can automatically match the conceptual knowledge probed by individual quiz
⁴⁵³ questions to the corresponding moments in lecture videos when those concepts were presented
⁴⁵⁴ (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment "knowledge traces"
⁴⁵⁵ that reflect the degree of knowledge participants have about each video's time-varying content,
⁴⁵⁶ and capture temporally specific increases in knowledge after viewing each lecture (Fig. 5). We
⁴⁵⁷ also show that these knowledge estimates can generalize to held-out questions (Fig. 6). Finally,
⁴⁵⁸ we use our framework to construct visual maps that provide snapshot estimates of how much
⁴⁵⁹ participants know about any concept within the scope of our text embedding model, and how
⁴⁶⁰ much their knowledge of those concepts changes with training (Fig. 8).

⁴⁶¹ ~~Over the~~ We view our work as making several contributions to the study of how people
⁴⁶² acquire conceptual knowledge. First, from a methodological standpoint, our modeling framework
⁴⁶³ provides a systematic means of mapping out and characterizing knowledge in maps that have
⁴⁶⁴ infinite (arbitrarily many) numbers of coordinates, and of "filling out" those maps using relatively
⁴⁶⁵ small numbers of multiple choice quiz questions. Our experimental finding that we can use these
⁴⁶⁶ maps to predict responses to held-out questions has several psychological implications as well. For
⁴⁶⁷ example, concepts that are assigned to nearby coordinates by the text embedding model also appear
⁴⁶⁸ to be "known to a similar extent" (as reflected by participants' responses to held-out questions;
⁴⁶⁹ Fig. 6). This suggests that participants also *conceptualize* similarly the content reflected by nearby
⁴⁷⁰ embedding coordinates. The "spatial smoothness" of participants' knowledge (as estimated using
⁴⁷¹ quiz performance) is being captured by the knowledge maps we are inferring from their quiz
⁴⁷² responses (e.g., Figs. 7, 8). In other words, our study shows that knowledge about a given concept

473 implies knowledge about related concepts, and we also show how estimated knowledge falls off
474 with distance in text embedding space.

475 In our study, we characterize the “coordinates” of participants’ knowledge using a relatively
476 simple “bag of words” text embedding model [LDA; 6]. More sophisticated text embedding
477 models, such as transformer-based models [15, 48, 59, 62] can learn complex grammatical and
478 semantic relationships between words, higher-order syntactic structures, stylistic features, and
479 more. We considered using transformer-based models in our study, but we found that the
480 text embeddings derived from these models were surprisingly uninformative with respect to
481 differentiating or otherwise characterizing the conceptual content of the lectures and questions
482 we used. We suspect that this reflects a broader challenge in constructing models that are
483 high-resolution within a given domain (e.g., the domain of physics lectures and questions) *and*
484 sufficiently broad so as to enable them to cover a wide range of domains. For example, we found
485 that the embeddings derived even from much larger and more modern models like BERT [15]
486 , GPT [62], LLaMa [59], and others that are trained on enormous text corpora, end up yielding
487 poor resolution within the content space spanned by individual course videos (Supp. Fig. 6).
488 Whereas the LDA embeddings of the lectures and questions are “near” each other (i.e., the
489 convex hull enclosing the two lectures’ trajectories is highly overlapping with the convex hull
490 enclosing the questions’ embeddings), the BERT embeddings of the lectures and questions are
491 instead largely distinct (top row of Supp. Fig. 6). The LDA embeddings of the questions for
492 each lecture and the corresponding lecture’s trajectory are also similar. For example, as shown in
493 Fig. 2C, the LDA embeddings for *Four Fundamental Forces* questions (blue dots) appear closer to
494 the *Four Fundamental Forces* lecture trajectory (blue line), whereas the LDA embeddings for *Birth*
495 of *Stars* questions (green dots) appear closer to the *Birth of Stars* lecture trajectory (green line).
496 The BERT embeddings of the lectures and questions do not show this property (Supp. Fig. 6).
497 We also examined per-question “content matches” between individual questions and individual
498 moments of each lecture (Figs. 4, 6). The time series plot of individual questions’ correlations
499 are different from each other when computed using LDA (e.g., the traces can be clearly visually
500 separated), whereas the correlations computed from BERT embeddings of different questions all

501 look very similar. This tells us that LDA is capturing some differences in content between the
502 questions, whereas BERT is not. The time series plots of individual questions' correlations have
503 clear "peaks" when computed using LDA, but not when computed using BERT. This tells us that
504 LDA is capturing a "match" between the content of each question and a relatively well-defined
505 time window of the corresponding lectures. The BERT embeddings appear to blur together the
506 content of the questions versus specific moments of each lecture. Finally, we also compared the
507 pairwise correlations between embeddings of questions within versus across content areas (i.e.,
508 content covered by the individual lectures, lecture-specific questions, and by the "general physics
509 knowledge" questions). The LDA embeddings show a strong contrast between same-content
510 embeddings versus across-content embeddings. In other words, the embeddings of questions
511 about the *Four Fundamental Forces* material are highly correlated with the embeddings of the *Four*
512 *Fundamental Forces* lecture, but not with the embeddings of *Birth of Stars*, questions about *Birth of*
513 *Stars*, or general physics knowledge questions. We see a similar pattern with the LDA embeddings
514 of the *Birth of Stars* questions (Fig. 3, Supp. Fig. 2). In contrast, the BERT embeddings are all
515 highly correlated with each other (Supp. Fig. 6). Taken together, these comparisons illustrate
516 how LDA (trained on the specific content in question) provides both coverage of the requisite
517 material and specificity at the level of the content covered by individual questions. BERT, on the
518 other hand, essentially assigns both lectures and all of the questions (which are all broadly about
519 "physics") into a tiny region of its embedding space, thereby blurring out meaningful distinctions
520 between different specific concepts covered by the lectures and questions. We note that these are
521 not criticisms of BERT (or other large language models trained on large and diverse corpora).
522 Rather, our point is that simple fine-tuned models trained on a relatively small but specialized
523 corpus can outperform much more complicated models trained on much larger corpora, when we
524 are specifically interested in capturing subtle conceptual differences at the level of a single course
525 lecture or question. Of course if our goal had been to find a model that generalized to many
526 different content areas, we would expect our approach to perform comparatively poorly relative to
527 BERT or other much larger models. We suggest that bridging the tradeoff between high resolution
528 within each content area versus the ability to generalize to many different content areas will be an

529 important challenge for future work in this domain.

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

541 At the opposite end of the spectrum from large language models, one could also imagine
542 simplifying some aspects of our LDA-based approach by computing simple word overlap metrics.
543 For example, the Jaccard similarity between text *A* and *B* is computed as the number of unique
544 words in the intersection of words from *A* and *B* divided by the number of unique words in the
545 union of words from *A* and *B*. In a supplementary analysis (Supp. Fig. 5), we compared the
546 LDA-based question-lecture matches we reported in Figure 4 with the Jaccard similarities between
547 each question and each sliding window of text from the corresponding lecture. As shown in
548 Supplementary Figure 5, this simple word-matching approach does not appear to capture the same
549 level of specificity as the LDA-based approach. Whereas the LDA-based approach often yields a
550 clear peak in the time series of correlations between each question and the corresponding lecture,
551 the Jaccard similarity-based approach does not. Furthermore, these LDA-based matches appear
552 to capture conceptual overlaps between the questions and lectures (Supp. Tab. 3), whereas simple
553 word matching does not. For example, one of the example questions examined in Supplementary
554 Figure 5 asks "Which of the following occurs as a cloud of atoms gets more dense?" The LDA-based
555 matches identify lecture timepoints where the relevant *topics* are discussed (e.g., when words like
556 "cloud," "atom," "dense," etc., are mentioned *together*). The Jaccard similarity-based matches,

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

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

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

571 Within the past several years, the global pandemic ~~has~~ forced many educators to suddenly
572 adapt to teaching remotely [30, 45, 56, 63]. This change in world circumstances is happening
573 alongside (and perhaps accelerating) geometric growth in the availability of high-quality online
574 courses from platforms such as Khan Academy [31], Coursera [64], EdX [33], and others [53].
575 Continued expansion of the global internet backbone and improvements in computing hardware
576 have also facilitated improvements in video streaming, enabling videos to be easily shared and
577 viewed by increasingly large segments of the world’s population. This exciting time for online
578 course instruction provides an opportunity to re-evaluate how we, as a global community, educate
579 ourselves and each other. For example, we can ask: what defines an effective course or training
580 program? Which aspects of teaching might be optimized and/or augmented by automated tools?
581 How and why do learning needs and goals vary across people? How might we lower barriers ~~of~~
582 ~~access to~~ ~~to receiving~~ a high-quality education?

583 Alongside these questions, there is a growing desire to extend existing theories beyond the
584 domain of lab testing rooms and into real classrooms [29]. In part, this has led to a recent resur-

gence of “naturalistic” or “observational” experimental paradigms that attempt to better reflect more ethologically valid phenomena that are more directly relevant to real-world situations and behaviors [46]. In turn, this has brought new challenges in data analysis and interpretation. A key step towards solving these challenges will be to build explicit models of real-world scenarios and how people behave in them (e.g., models of how people learn conceptual content from real-world courses, as in our current study). A second key step will be to understand which sorts of signals derived from behaviors and/or other measurements (e.g., [neurophysiological data](#); 2, 16, 43, 47, 50) [e.g., [neurophysiological data](#); 2, 16, 43, 47, 50] might help to inform these models. A third major step will be to develop and employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic paradigms.

Beyond specifically predicting what people *know*, the fundamental ideas we develop here also relate to the notion of “theory of mind” of other individuals [22, 27, 42]. Considering others’ unique perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and communicate [51, 55, 58]. One could imagine future extensions of our work (e.g., analogous to the knowledge and learning maps shown in Fig. 8), that attempt to characterize how well-aligned different people’s knowledge bases or backgrounds are. In turn, this might be used to model how knowledge (or other forms of communicable information) flows not just between teachers and students, but between friends having a conversation, individuals on a first date, participants at a business meeting, doctors and patients, experts and non-experts, political allies or adversaries, and more. For example, the extent to which two people’s knowledge maps “match” or “align” in a given region of text embedding space might serve as a predictor of how effectively they will be able to communicate about the corresponding conceptual content.

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

613 **Materials and methods**

614 **Participants**

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

620 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
621 years). A total of 15 participants reported their gender as male and 35 participants reported their
622 gender as female. A total of 49 participants reported their native language as "English" and 1
623 reported having another native language. A total of 47 participants reported their ethnicity as
624 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants
625 reported their races as White (32 participants), Asian (14 participants), Black or African American
626 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
627 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

Participants reported their undergraduate major(s) as “social sciences” (28 participants), “natural sciences” (16 participants), “professional” (e.g., pre-med or pre-law; 8 participants), “mathematics and engineering” (7 participants), “humanities” (4 participants), or “undecided” (3 participants). Note that some participants selected multiple categories for their undergraduate major(s). We also asked participants about the courses they had taken. In total, 45 participants reported having taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan Academy courses. Of those who reported having watched at least one Khan Academy course, 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We also asked participants about the specific courses they had watched, categorized under different subject areas. In the “Mathematics” area, participants reported having watched videos on AP Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Calculus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants), Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other videos not listed in our survey (5 participants). In the “Science and engineering” area, participants reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 participants); Physics, AP Physics I, or AP Physics II (18 participants); Biology, AP Biology; or High school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed in our survey (5 participants). We also asked participants whether they had specifically seen the videos used in our experiment. Of the 45 participants who reported having having taken at least one Khan Academy course in the past, 44 participants reported that they had not watched the *Four Fundamental Forces* video, and 1 participant reported that they were not sure whether they had watched it. All participants reported that they had not watched the *Birth of Stars* video. When we asked participants about non-Khan Academy online courses, they reported having watched or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 participants).

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

674 Experiment

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

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

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

695 quizzes: the first before viewing Lecture 1, the second between Lectures 1 and 2, and the third after
696 viewing Lecture 2 (see Fig. 1). The questions appearing on each quiz, for each participant, were
697 randomly chosen from the full set of 39, with the constraints that (a) each quiz ~~contain~~contained
698 exactly 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general
699 physics knowledge, and (b) each question appear exactly once for each participant. The orders of
700 questions on each quiz, and the orders of answer options for each question, were also randomized.
701 ~~Our~~We obtained informed consent from all participants, and our experimental protocol was
702 approved by the Committee for the Protection of Human Subjects at Dartmouth College. We used
703 this experiment to develop and test our computational framework for estimating knowledge and
704 learning.

705 Analysis

706 Statistics

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

710 Constructing text embeddings of multiple lectures and questions

711 We adapted an approach we developed in prior work [24] to embed each moment of the two
712 lectures and each question in our pool in a common representational space. Briefly, our approach
713 uses a topic model(~~Latent Dirichlet Allocation; 6~~, [Latent Dirichlet Allocation; 6]) trained on a set
714 of documents, to discover a set of ~~(up to)~~^k “topics” or “themes.” Formally, each topic is defined
715 as a distribution of weights over ~~each word~~words in the model’s vocabulary (i.e., the union of
716 all unique words, across all documents, excluding “stop words.”). Conceptually, each topic is
717 intended to give larger weights to words that are semantically related ~~or tend (as inferred from~~
718 their tendency to co-occur in the same ~~documents~~document). After fitting a topic model, each
719 document in the training set, or any new document that contains at least some of the words in

720 the model’s vocabulary, may be represented as a k -dimensional vector describing how much the
721 document (most probably) reflects each topic. To select an appropriate k for our model, as a starting
722 point, we identified the minimum number of topics that yielded at least one “unused” topic (i.e., in
723 which all words in the vocabulary were assigned uniform weights) after training. This indicated
724 that the number of topics was sufficient to capture the set of latent themes present in the two lectures
725 (from which we constructed our document corpus, as described below). We found this value to be
726 $k = 15$ topics. We found that with a limited number of additional adjustments following [7], such
727 as removing corpus-specific stop-words, the model yielded (subjectively) sensible and coherent
728 topics. The distribution of weights over words in the vocabulary for each discovered topic is shown
729 in Supplementary Figure 1, and each topic’s top-weighted words may be found in Supplementary
730 Table 2.

731 As illustrated in Figure 2A, we start by building up a corpus of documents using overlap-
732 ping sliding windows that span each video’s transcript. Khan Academy provides professionally
733 created, manual transcriptions of all videos for closed captioning. However, such transcripts
734 would not be readily available in all contexts to which our framework could potentially be ap-
735 plied. Khan Academy videos are hosted on the YouTube platform, which additionally provides
736 automated captions. We opted to use these automated transcripts(~~which, in prior work, we~~
737 ~~have found to be of sufficiently near-human quality to yield reliable data in behavioral studies; 65~~
738 [which, in prior work, we have found to be of sufficiently near-human quality to yield reliable data in behavioral studi
739 when developing our framework in order to make it more directly extensible and adaptable by
740 others in the future.

741 We fetched these automated transcripts using the `youtube-transcript-api` Python pack-
742 age [14]. The transcripts consisted of one timestamped line of text for every few seconds (mean:
743 2.34 s; standard deviation: 0.83 s) of spoken content in the video (i.e., corresponding to each indi-
744 vidual caption that would appear on-screen if viewing the lecture via YouTube, and when those
745 lines would appear). We defined a sliding window length of (up to) $w = 30$ transcript lines, and
746 assigned each window a timestamp corresponding to the midpoint between the timestamps for its
747 first and last lines. This w parameter was chosen to match the same number of words per sliding

748 window (rounded to the nearest whole word, and before preprocessing) as the sliding windows
749 we defined in our prior work [24] (i.e., 185 words per sliding window).

750 These sliding windows ramped up and down in length at the beginning and end of each
751 transcript, respectively. In other words, each transcript’s first sliding window covered only its first
752 line, the second sliding window covered the first two lines, and so on. This ensured that each
753 line from the transcripts appeared in the same number (w) of sliding windows. ~~After performing~~
754 ~~various~~ We next performed a series of standard text preprocessing (e.g., steps: normalizing case,
755 lemmatizing, removing punctuation and removing stop-words), we. We constructed our corpus
756 of stop words by augmenting the Natural Language Toolkit [NLTK; 3] English stop word list with
757 the following additional words, selected using one of the approaches suggested by [7]: “actual,”
758 “actually,” “also,” “bit,” “could,” “e,” “even,” “first,” “follow,” “following,” “four,” “let,” “like,”
759 “mc,” “really,” “saw,” “see,” “seen,” “thing,” and “two.” This yielded sliding windows with an
760 average of 73.8 remaining words, and lasting for an average of 62.22 seconds. We treated the text
761 from each sliding window as a single “document,” and combined these documents across the two
762 videos’ windows to create a single training corpus for the topic model.

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

770 We transformed each sliding window’s text into a topic vector, and then used linear interpolation
771 (independently for each topic dimension) to resample the resulting ~~timeseries~~ time series
772 to one vector per second. We also used the fitted model to obtain topic vectors for each question
773 in our pool (see Supp. Tab. 1). Taken together, we obtained a trajectory for each video, describing
774 its path through topic space, and a single coordinate for each question (Fig. 2C). Embedding both
775 videos and all of the questions using a common model enables us to compare the content from dif-

776 ferent moments of videos, compare the content across videos, and estimate potential associations
777 between specific questions and specific moments of video.

778 **Estimating dynamic knowledge traces**

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

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

781 where

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

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

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

796 **Estimating the "smoothness" of knowledge**

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

802 For each of 10,000 iterations, we drew a random subsample (with replacement) of 50 participants
803 from our dataset full dataset. Within each iteration, we first computed the 95% confidence interval
804 (CI) of the across-subsample-participants mean proportion correct on each of the three quizzes,
805 separately. To compute this interval for each quiz, we repeatedly (1,000 times) subsampled
806 participants (with replacement, from the outer subsample for the current iteration) and computed
807 the mean proportion correct each of these inner subsamples. We then identified the 2.5th and
808 97.5th percentiles of the resulting distributions of 1,000 means. These three intervals (one for each
809 quiz) served as our thresholds for confidence that the proportion correct within a given distance
810 from a reference question was reliably different (at the $p < 0.05$ significance level) from the average
811 proportion correct across all questions on the given quiz.

812 Next, for each participant in the current subsample, and for each of the three quizzes they
813 completed (separately), we iteratively treated each of the 15 questions appearing on the given
814 quiz as the "reference" question. We constructed a series of concentric 15-dimensional "spheres"
815 centered on the reference question's embedding space coordinate, where each successive sphere's
816 radius increased by 0.01 (correlation distance) between 0 and 2, inclusive (i.e., tiling the range
817 of possible correlation distances with 201 spheres in total). We then computed the proportion
818 of questions enclosed within each sphere that the participant answered correctly, and averaged
819 these per-radius proportion correct scores across reference questions that were answered correctly,
820 and those that were answered incorrectly. This resulted in two number-of-spheres sequences of
821 proportion-correct scores for each subsample participant and quiz: one derived from correctly
822 answered reference questions, and one derived from incorrectly answered reference questions.

823 We computed the across-subsample-participants mean proportion correct for each radius
824 value (i.e., sphere) and "correctness" of reference question. This yielded two sequences of

825 proportion-correct scores for each quiz, analogous to the blue and red lines displayed in Figure 7A,
826 but for the present subsample. For each quiz, we then found the minimum distance from
827 the reference question (i.e., sphere radius) at which each of these two sequences of per-radius
828 proportion correct scores intersected the 95% confidence interval for the overall proportion correct
829 (i.e., analogous to the black error bands in Fig. 7A).

830 This resulted in two “intersection” distances for each quiz (for correctly answered and incorrectly
831 answered reference questions). Repeating this full process for each of the 10,000 bootstrap iterations
832 output two distributions of intersection distances for each of the three quizzes. The means and
833 95% confidence intervals for these distributions are plotted in Figure 7B.

834 Creating knowledge and learning map visualizations

835 An important feature of our approach is that, given a trained text embedding model and participants’ quiz performance on each question, we can estimate their knowledge about *any* content expressible by the embedding model—not solely the content explicitly probed by the quiz questions, or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 7, 8, 9, 10, and 11), we used Uniform Manifold Approximation and Projection(UMAP; 40, 41) [UMAP; 40, 41] to construct a 2D projection of the text embedding space. ~~Sampling the original Whereas our main analyses used a 15-topic embedding space, we used a 100-topic embedding space for these visualizations. This change in the number of topics overcame an undesirable behavior in the UMAP embedding procedure, whereby embedding coordinates for the 15-topic model tended to be “clumped” into separated clusters, rather than forming a smooth trajectory through the 2D space. When we increased the number of topics to 100, the embedding coordinates in the 2D space formed a smooth trajectory through the space, with substantially less clumping (Fig. 8). Creating a “map” by sampling this~~ 100-dimensional space at high resolution to obtain an adequate set of topic vectors spanning the embedding space would be computationally intractable. However, sampling a 2D grid is trivial.

840 At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing
841 the cross-entropy between the pairwise (clustered) distances between the observations in their

852 original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional
853 embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise
854 distances in the original high-dimensional space were defined as 1 minus the correlation between
855 each pair of coordinates, and pairwise distances in the low-dimensional embedding space were
856 defined as the Euclidean distance between each pair of coordinates.

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

867 After embedding both lectures’ topic trajectories and the topic vectors of every question, we
868 defined a rectangle enclosing the 2D projections of the lectures’ and quizzes’ embeddings. We then
869 sampled points from a regular 100×100 grid of coordinates that evenly tiled this enclosing rectangle.
870 We sought to estimate participants’ knowledge (and learning, i.e., changes in knowledge) at each
871 of the resulting 10,000 coordinates.

872 To generate our estimates, we placed a set of 39 radial basis functions (RBFs) throughout the
873 embedding space, centered on the 2D projections for each question (i.e., we included one RBF for
874 each question). At coordinate x , the value of an RBF centered on a question’s coordinate μ , is given
875 by:

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

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

878 “knowledge” at each coordinate, x , using:

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

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

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

883 Author contributions

884 Conceptualization: ~~PCF, ACH, and JRM~~. Methodology: ~~PCF, ACH, and JRM~~. Software: ~~PCF~~. Validation:
885 ~~PCF, P.C.F., A.C.H., and J.R.M.~~ Methodology: P.C.F., A.C.H., and J.R.M. Software: P.C.F. Validation:
886 P.C.F. Formal analysis: ~~PCF~~. Resources: ~~PCF, ACH, and JRM~~. P.C.F. Resources: P.C.F., A.C.H., and
887 J.R.M. Data curation: ~~PCF, P.C.F.~~ Writing (original draft): ~~JRM, J.R.M.~~ Writing (review and editing):
888 ~~PCF, ACH, and JRM~~. Visualization: ~~PCF and JRM~~. Supervision: ~~JRM, P.C.F., A.C.H., and J.R.M.~~
889 Visualization: P.C.F. and J.R.M. Supervision: J.R.M. Project administration: ~~PCF, P.C.F.~~ Funding
890 acquisition: ~~JRM, J.R.M.~~ J.R.M.

891 Data and code availability

892 All of the data analyzed in this manuscript, ~~along with all~~ may be found at <https://github.com/ContextLab/efficient-learning-khan>.

894 Code availability

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

897 **Acknowledgements**

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

905 **References**

- 906 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,
907 56:149–178.
- 908 [2] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and
909 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
910 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 911 [3] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text
912 with the natural language toolkit*. Reilly Media, Inc.
- 913 [4] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
914 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
915 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 916 [5] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International
917 Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
918 Machinery.
- 919 [6] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine
920 Learning Research*, 3:993–1022.

- 921 [7] Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models:
922 problems, diagnostics, and improvements. In Airoldi, E. M., Blei, D. M., Erosheva, E. A., and
923 Fienberg, S. E., editors, *Handbook of Mixed Membership Models and Their Applications*. CRC Press.
- 924 [8] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
925 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
926 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
927 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
928 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 929 [9] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
930 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 931 [10] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
932 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
933 sentence encoder. *arXiv*, 1803.11175.
- 934 [11] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
935 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 936 [12] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
937 Evidence for a new conceptualization of semantic representation in the left and right cerebral
938 hemispheres. *Cortex*, 40(3):467–478.
- 939 [13] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
940 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,
941 41(6):391–407.
- 942 [14] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 943 [15] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep
944 bidirectional transformers for language understanding. *arXiv*, 1810.04805.

- 946 [16] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
947 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
948 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 949 [17] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 950 [18] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of*
951 *Experimental Psychology: General*, 115:155–174.
- 952 [19] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*
953 *Transactions of the Royal Society A*, 222(602):309–368.
- 954 [20] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
955 *School Science and Mathematics*, 100(6):310–318.
- 956 [21] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather
957 prediction” task? individual variability in strategies for probabilistic category learning. *Learning*
958 *and Memory*, 9:408–418.
- 959 [22] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*
960 *Cognition and Development*, 13(1):19–37.
- 961 [23] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
962 learning, pages 212–221. Sage Publications.
- 963 [24] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-
964 ioral and neural signatures of transforming experiences into memories. *Nature Human Behavior*,
965 5:905–919.
- 966 [25] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge au-
967 tomatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*,
968 9:doi.org/10.3389/fpsyg.2018.00133.

- 969 [26] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-
970 peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–
971 4008.
- 972 [27] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
973 empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 974 [28] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
975 Columbia University Press.
- 976 [29] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
977 326(7382):213–216.
- 978 [30] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
979 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International
980 Journal of Environmental Research and Public Health*, 18(5):2672.
- 981 [31] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 982 [32] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 983 [33] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
984 *The Chronicle of Higher Education*, 21:1–5.
- 985 [34] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic
986 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
987 104:211–240.
- 988 [35] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic
989 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.
- 990 [36] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of
991 Educational Studies*, 53(2):129–147.

- 992 [37] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
993 *Handbook of Human Memory*. Oxford University Press.
- 994 [38] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
995 function? *Psychological Review*, 128(4):711–725.
- 996 [39] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension
997 for adding LLM-driven chatbots to interactive notebooks. [https://github.com/ContextLab/
998 chatify](https://github.com/ContextLab/chatify).
- 999 [40] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and
1000 projection for dimension reduction. *arXiv*, 1802(03426).
- 1001 [41] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold
1002 Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 1003 [42] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
1004 mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 1005 [43] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
1006 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
1007 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 1008 [44] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
1009 tations in vector space. *arXiv*, 1301.3781.
- 1010 [45] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
1011 from a national survey of language educators. *System*, 97:102431.
- 1012 [46] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
1013 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1014 [47] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
1015 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective
1016 Neuroscience*, 17(4):367–376.

- 1017 [48] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 1018 [49] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models.
1019 *arXiv*, 2208.02957.
- 1020 [50] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
1021 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
1022 7:43916.
- 1023 [51] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of
1024 Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 1025 [52] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.
1026 *Biological Cybernetics*, 45(1):35–41.
- 1027 [53] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
1028 higher education: unmasking power and raising questions about the movement’s democratic
1029 potential. *Educational Theory*, 63(1):87–110.
- 1030 [54] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
1031 Student conceptions and conceptual learning in science. Routledge.
- 1032 [55] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
1033 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-
1034 tion in Nursing*, 22:32–42.
- 1035 [56] Shim, T. E. and Lee, S. Y. (2020). College students’ experience of emergency remote teaching
1036 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 1037 [57] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
1038 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in
1039 Mathematics Education*, 35(5):305–329.
- 1040 [58] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal
1041 Medicine*, 21:524–530.

- 1042 [59] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,
1043 Goyal, N., Hambro, E., Azhar, F., Rodriguz, A., Joulin, A., Grave, E., and Lample, G. (2023).
1044 LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 1045 [60] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Pio-
1046 ntkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust
1047 detection of AI-generated texts. *arXiv*, 2306.04723.
- 1048 [61] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold?
1049 distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of*
1050 *the Cognitive Science Society*, 43(43).
- 1051 [62] Viswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and
1052 Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing*
1053 *Systems*.
- 1054 [63] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
1055 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 1056 [64] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
1057 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 1058 [65] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is
1059 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*
1060 *Research Methods*, 50:2597–2605.