

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 applied our framework to the videos' transcripts, and
¹² to text of the quiz questions, to quantify the content of each moment of video and each quiz
¹³ question. We used 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 knew.
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 knew
²³ the to-be-learned information already, or how much they knew about related concepts. For some
²⁴ students, they could potentially optimize their teaching efforts to maximize efficiency by focusing
²⁵ primarily on not-yet-known content. For other students (or other content areas), it might be more
²⁶ effective to optimize for direct connections between already known content and new material.
²⁷ Observing how the student’s knowledge changed over time, in response to their teaching, could
²⁸ also help to guide the teacher towards the most effective strategy for that individual student.

²⁹ A common approach to assessing students’ learning is to compile a set of quiz questions, tally
³⁰ up the proportion of correctly answered questions, and provide the student with feedback in the
³¹ form of a letter grade. Although a numerical or letter grades give *some* indication about whether
³² the student has mastered the to-be-learned material, any single measure of performance on a
³³ complex task is at risk of conflating underlying factors, losing relevant information, and so on. For
³⁴ example, consider the relative utility of the imaginary map described above that characterizes in
³⁵ detail a student’s knowledge, versus a single annotation saying that the student answered 85% of
³⁶ their quiz questions correctly, or that they received a ‘B’. Here, we show that the same quiz data
³⁷ required to compute proportion correct scores or letter grades can be used to instead obtain much
³⁸ more detailed insights into what the student knows at the time they took the quiz.

³⁹ Designing and building procedures and tools for mapping out knowledge touches on deep
⁴⁰ questions about what it means to learn. For example, how do we acquire conceptual knowledge?
⁴¹ Memorizing course lectures or textbook chapters by rote can lead to the superficial *appearance*
⁴² of understanding the underlying content, but achieving true conceptual understanding seems to
⁴³ require something deeper and richer. Does conceptual understanding entail connecting newly
⁴⁴ acquired information to the scaffolding of one’s existing knowledge or experience [2, 6, 8, 9, 42]?

45 Or weaving a lecture’s atomic elements (e.g., its component words) into a structured network
46 that describes how those individual elements are related? Conceptual understanding could also
47 involve building a mental model that transcends the meanings of those individual atomic elements
48 by reflecting the deeper meaning underlying the gestalt whole [23, 26, 39].

49 The difference between “understanding” and “memorizing,” as framed by researchers in ed-
50 ucation, cognitive psychology, and cognitive neuroscience [e.g., 14, 16, 19, 26, 39] has profound
51 analogs in the fields of natural language processing and natural language understanding. For
52 example, considering the raw contents of a document (e.g., its constituent symbols, letters, and
53 words) might provide some information about what the document is about, just as memorizing a
54 passage might provide some ability to answer simple questions about it. However, text embedding
55 models [e.g., 3–5, 7, 10, 25, 32] also attempt to capture the deeper meaning *underlying* those atomic
56 elements. These models consider not only the co-occurrences of those elements within and across
57 documents, but also patterns in how those elements appear across different scales (e.g., sentences,
58 paragraphs, chapters, etc.), the temporal and grammatical properties of the elements, and other
59 high-level characteristics of how they are used [27, 28]. According to these models, the deep
60 conceptual meaning of a document may be captured by a feature vector in a high-dimensional
61 representation space, where nearby vectors reflect conceptually related documents. A model that
62 succeeds at capturing an analog of “understanding” is able to assign nearby feature vectors to
63 two conceptually related documents, *even when the words contained in those documents have very little*
64 *overlap*.

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

73 should change accordingly. Learning new concepts should both update our characterizations of
74 “what is known” and also unlock any now-satisfied dependencies of those newly learned concepts
75 so that they are “tagged” as available for future learning.

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

84 Our primary research goal is to advance our understanding of what it means to acquire deep,
85 real-world conceptual knowledge. Traditional laboratory approaches to studying learning and
86 memory (e.g., list learning studies) often draw little distinction between memorization and under-
87 standing. Instead, these studies typically focus on whether information is effectively encoded or
88 retrieved, rather than whether the information is *understood*. Approaches to studying conceptual
89 learning, such as category learning experiments, can begin to investigate the distinction between
90 memorization and understanding, often by training participants to distinguish arbitrary or ran-
91 dom features in otherwise meaningless categorized stimuli. However the objective of real-world
92 training, or learning from life experiences more generally, is often to develop new knowledge that
93 may be applied in *useful* ways in the future. In this sense, the gap between modern learning theo-
94 ries and modern pedagogical approaches and classroom learning strategies is enormous: most of
95 our theories about *how* people learn are inspired by experimental paradigms and models that have
96 only peripheral relevance to the kinds of learning that students and teachers actually seek [16, 26].
97 To help bridge this gap, our study uses course materials from real online courses to inform, fit,
98 and test models of real-world conceptual learning. We also provide a demonstration of how our
99 models can be used to construct “maps” of what students know, and how their knowledge changes
100 with training. In addition to helping to visualize knowledge (and changes in knowledge), we hope

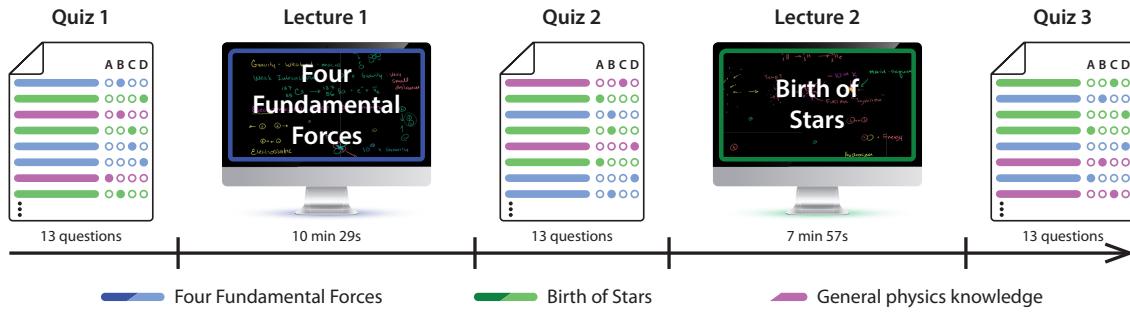


Figure 1: Experimental paradigm. Participants alternate between answering 13-question multiple-choice quizzes and watching two Khan Academy lectures. Each quiz contains a mix of 5 questions about lecture 1, 5 questions about lecture 2, and 3 questions about general physics knowledge. The specific questions reflected on each quiz, and the orders of each quiz's questions, were randomized across participants.

101 that such maps might lead to real-world tools for improving how we educate.

Results

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

116 We asked participants in our study to complete brief multiple-choice quizzes before, between,

118 and after watching two lecture videos from the Khan Academy [22] platform (Fig. 1). The first
119 lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics:
120 gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*,
121 provided an overview of our current understanding of how stars form. We selected these particular
122 lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad
123 audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training on
124 our participants' abilities to learn from the lectures. To this end, we selected two introductory videos
125 that were intended to be viewed at the start of students' training in their respective content areas.
126 Second, we wanted both lectures to have some related content, so that we could test our approach's
127 ability to distinguish similar conceptual content. To this end, we chose two videos from the same
128 (per instructor annotations) Khan Academy course domain, "Cosmology and Astronomy." Third,
129 we sought to minimize dependencies and specific overlap between the videos. For example, we
130 did not want participants' abilities to understand one video to (directly) influence their abilities to
131 understand the other. To satisfy this last criterion, we chose videos from two different lecture series
132 (lectures 1 and 2 were from the "Scale of the Universe" and "Stars, Black Holes, and Galaxies"
133 series, respectively).

134 We also wrote a set of multiple-choice quiz questions that we hoped would enable us to
135 evaluate participants' knowledge about each individual lecture, along with related knowledge
136 about physics not specifically presented in either video (see Tab. S1 for the full list of questions
137 in our stimulus pool). Participants answered questions randomly drawn from each content area
138 (lecture 1, lecture 2, and general physics knowledge) on each of the three quizzes. Quiz 1 was
139 intended to assess participants' "baseline" knowledge before training, quiz 2 assessed knowledge
140 after watching the *Four Fundamental Forces* video (i.e., lecture 1), and quiz 3 assessed knowledge
141 after watching the *Birth of Stars* video (i.e., lecture 2).

142 To study in detail how participants' conceptual knowledge changed over the course of the
143 experiment, we first sought to model the conceptual content presented to them at each moment
144 throughout each of the two lectures. We adapted an approach we developed in prior work [17] to
145 identify the latent themes in the lectures using a topic model [4]. Briefly, topic models take as input

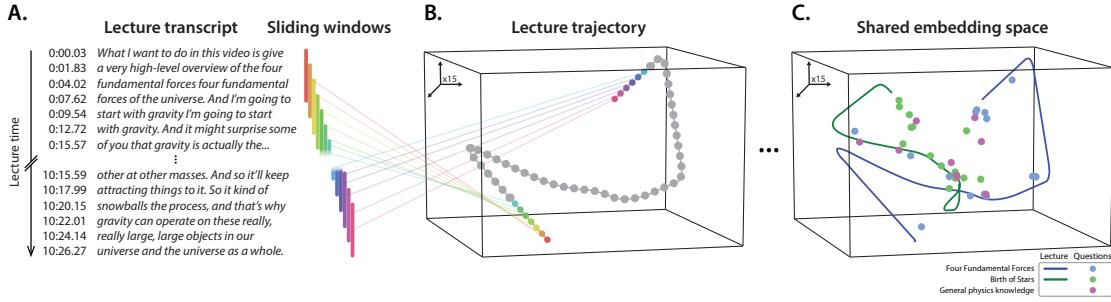


Figure 2: Modeling course content. **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding model. **B. Constructing lecture content trajectories.** After training our 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 (Tab. S1), to project them into a shared text embedding space. This results in one trajectory per lecture and one coordinate for each question. Here we have projected the 15-dimensional embeddings onto their first 3 principal components for visualization.

146 a collection of text documents and learn a set of “topics” (i.e., latent themes) from their contents.
 147 Once fit, a topic model can be used to transform arbitrary (potentially new) documents into sets
 148 of “topic proportions,” describing the weighted blend of learned topics reflected in their texts. We
 149 parsed automatically generated transcripts of the two lectures into overlapping sliding windows,
 150 where each window contained the text of the lecture transcript from a particular time range. We
 151 treated the set of text snippets (across all of these windows) as documents to fit our model (Fig. 2A;
 152 see Constructing text embeddings of multiple lectures and questions). Transforming the text from
 153 every sliding window with our model yielded a number-of-windows by number-of-topics (15)
 154 topic-proportions matrix that described the unique mixture of broad themes from both lectures
 155 reflected in each window’s content. Each window’s “topic vector” (i.e., column of the topic-
 156 proportions matrix) is a coordinate in a 15-dimensional space whose axes are topics discovered by
 157 the model. Within this space, each lecture’s sequence of topic vectors (i.e., corresponding to its
 158 transcript’s overlapping text snippets across sliding windows) forms a *trajectory* that captures how
 159 its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution
 160 of one topic vector for each second of video (i.e., 1 Hz).

161 We hypothesized that a topic model trained on transcripts of the two lectures, should also
162 capture the conceptual knowledge probed by each quiz question. If indeed the topic model could
163 capture information about the deeper conceptual content of the lectures (i.e., beyond surface-level
164 details such as particular word choices) then we should be able to recover a correspondence between
165 each lecture and questions *about* each lecture. Importantly, such a correspondence could not solely
166 arise from superficial text matching between lecture transcripts and questions, since the lectures and
167 questions used different words. Simply comparing the average topic weights from each lecture and
168 question sets (averaging across time and questions, respectively) reveals a striking correspondence
169 (Fig. S1). Specifically, the average topic weights from lecture 1 are strongly correlated with the
170 average topic weights from lecture 1 questions ($r(13) = XX, p = XX$, 95% confidence interval
171 (CI) = XX), and the average topic weights from lecture 2 are strongly correlated with the average
172 topic weights from lecture 2 questions ($r(13) = XX, p = XX, CI = XX$). At the same time, the
173 average topics from two lectures are *negatively* correlated ($r(13) = XX, p = XX, CI = XX$). The full
174 set of pairwise comparisons between topic vectors for the lectures and each question set is reported
175 in Figure S1.

176 Another, more sensitive, way of summarizing the conceptual content of the lectures and ques-
177 tions is to look at *variability* in how topics are weighted over time and across different questions
178 (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “infor-
179 mation” [13] the lecture (or questions) reflect about that topic. For example, suppose a given topic
180 is weighted on heavily throughout a lecture. That topic might be characteristic of some aspect or
181 property of the lecture *overall* (conceptual or otherwise), but unless the topic’s weights changed in
182 meaningful ways over time, the topic would be a poor indicator of any *specific* conceptual content
183 in the lecture. We therefore also compared the variance in topic weights (across time or questions)
184 between the lectures and questions. The variability in topic expression (over time and across ques-
185 tions) was similar for the lecture 1 video and questions ($r(13) = 0.824, p < 0.001, CI = [0.696, 0.973]$)
186 and the lecture 2 video and questions ($r(13) = 0.801, p < 0.001$, 95% CI = [0.539, 0.958]). However,
187 as reported in Figure 3B, the variability in topic expressions across *different* videos and lecture-
188 specific questions (i.e., lecture 1 video versus lecture 2 questions; lecture 2 video versus lecture 1



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 Table S2. **B. Relationships between topic weight variability.** Pairwise correlations between the distributions of topic weight variance for each lecture and question category. Each row and column corresponds to a bar plot in Panel A.

189 questions) were *negatively* correlated, and neither video’s topic variability was reliably correlated
190 with the topic variability across general physics knowledge questions. Taken together, the analyses
191 reported in Figures 3 and S1 indicate that a topic model fit to the videos’ transcripts can also reveal
192 correspondances (at a coarse scale) between the lectures and (held-out) questions.

193 Although a single lecture may be organized around a single broad theme at a coarse scale, at a
194 finer scale each moment of a lecture typically covers a narrower range of content. We wondered
195 whether a text embedding model trained on the lectures’ transcripts might capture some of this
196 finer scale content. For example, if a particular question asks about the content from one small
197 part of a lecture, we wondered whether our text embedding model could be used to automatically
198 identify the “matching” moment(s) in the lecture. When we correlated each question’s topic vector
199 with the topic vectors for each second of the lectures, we found some evidence that each question is
200 temporally specific (Fig. 4). In particular, most questions’ topic vectors were maximally correlated
201 with a well-defined (and relatively narrow) range of timepoints from their corresponding lectures,
202 and the correlations fell off sharply outside of that range. We also examined the best-matching

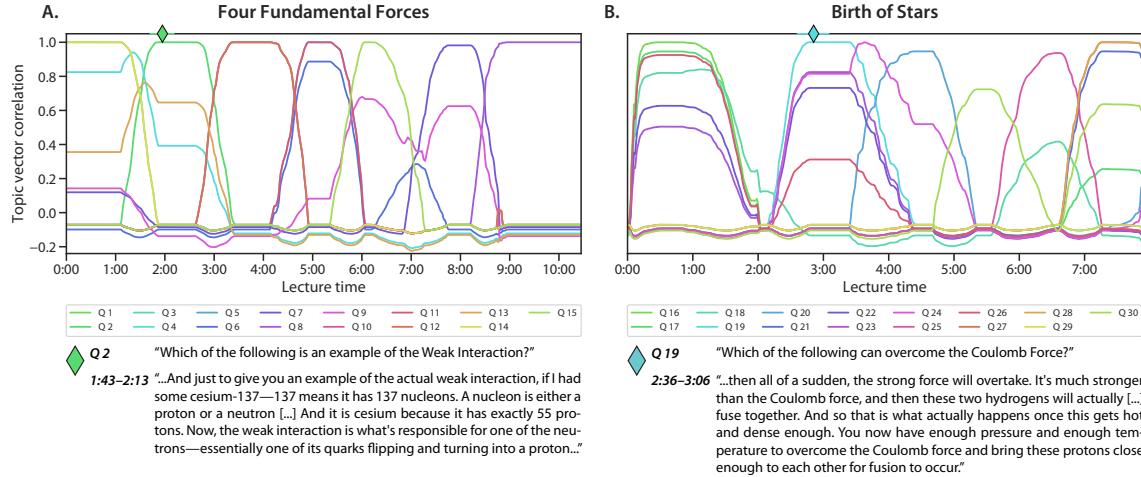


Figure 4: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel B.: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

203 intervals for each question qualitatively by comparing the text of the question to the text of the most-
204 correlated parts of the lectures. Despite that the questions were excluded from the text embedding
205 model’s training set, in general we found (through manual inspection) a close correspondence
206 between the conceptual content that each question covered and the content covered by the best-
207 matching moments of the lectures. Two representative examples are shown at the bottom of
208 Figure 4.

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

216 the same *proportion* of questions correct as another student who missed three questions about
217 three *different* concepts (e.g., *A*, *B*, and *C*). But if we wanted to fill in the “gaps” in the two
218 students’ understandings, we might do well to focus on concept *A* for the first student, but to
219 also add in materials pertaining to concepts *B* and *C* for the second student. In other words, raw
220 “proportion-correct” measures may capture *how much* a student knows, but not *what* they know.
221 We wondered whether our modeling framework might enable us to (formally and automatically)
222 infer participants’ knowledge at the scale of individual concepts (e.g., as captured by a single
223 question).

224 We developed a simple formula (Eqn. 1) for using a participant’s responses to a small set
225 of multiple-choice questions to estimate how much the participant “knows” about the concept
226 reflected by any arbitrary coordinate, x , in text embedding space (e.g., the content reflected by any
227 moment in a lecture they had watched; see Estimating dynamic knowledge traces). Essentially,
228 the estimated knowledge at the coordinate is given by the weighted average proportion of quiz
229 questions the participant answered correctly, where the weights reflect how much each question
230 is “about” the content at x . When we apply this approach to estimate the participant’s knowledge
231 about the content presented in each moment of each lecture, we can obtain a detailed timecourse
232 describing how much “knowledge” the participant has about any part of the lecture. As shown
233 in Figure 5, we can also apply this approach separately for the questions from each quiz the
234 participants took throughout the experiment. From just 13 questions per quiz, we obtain a high-
235 resolution snapshot (at the time each quiz was taken) of what the participants knew about any
236 moment’s content, from either of the two lectures they watched (comprising a total of 1106 samples
237 across the two lectures).

238 Of course, even though the timecourses in Figure 5A and C provide detailed *estimates* about
239 participants’ knowlege, those estimates are only *useful* to the extent that they accurately reflect what
240 participants actually know. As one sanity check, we anticipated that the knowledge estimates
241 should show a content-specific “boost” in participants’ knowledge after watching each lecture.
242 In other words, if participants learn about each lecture’s content when they watch each lecture,
243 the knowledge estimates should reflect that. After watching the *Four Fundamental Forces* lecture,



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

244 participants should show more knowledge for the content of that lecture than they had before,
245 and that knowledge should persist for the remainder of the experiment. Specifically, knowledge
246 about that lecture's content should be relatively low when estimated using Quiz 1 responses,
247 but should increase when estimated using Quiz 2 or 3 responses (Fig. 5B). Indeed, we found
248 that participants' estimated knowledge about the content of the *Four Fundamental Forces* was
249 substantially higher on Quiz 2 versus Quiz 1 ($t(49) = 8.764, p < 0.001$) and on Quiz 3 versus Quiz
250 1 ($t(49) = 10.519, p < 0.001$). We found no reliable differences in estimated knowledge about
251 that lecture's content on Quiz 2 versus 3 ($t(49) = 0.160, p = 0.874$). Similarly, we hypothesized
252 (and subsequently confirmed) that participants should show more estimated knowledge about the
253 content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 5D). Specifically, since
254 participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their
255 knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a "boost" on
256 Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge
257 about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ($t(49) = 1.013, p = 0.316$), but the
258 estimated knowledge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and
259 Quiz 3 versus 1 ($t(49) = 8.969, p < 0.001$).

260 If we are able to accurately estimate a participant's knowledge about the content tested by a
261 given question, the estimated knowledge should have some predictive information about whether
262 the participant is likely to answer the question correctly or incorrectly. For each question in turn, for
263 each participant, we used Equation 1 to estimate (using all *other* questions from the same quiz, from
264 the same participant) the participant's knowledge at the held-out question's embedding coordinate.
265 For each quiz, we grouped these estimates into two distributions: one for the estimated knowledge
266 at the coordinates of each *correctly* answered question, and another for the estimated knowledge at
267 the coordinates of each *incorrectly* answered question (Fig. 6). We then used independent samples
268 t -tests to compare the means of these distributions of estimated knowledge.

269 For the initial quizzes participants took (prior to watching either lecture), participants' estimated
270 knowledge tended to be low overall, and relatively unstructured (Fig. 6, left panel). When we held
271 out individual questions and estimated their knowledge at the held-out questions' embedding

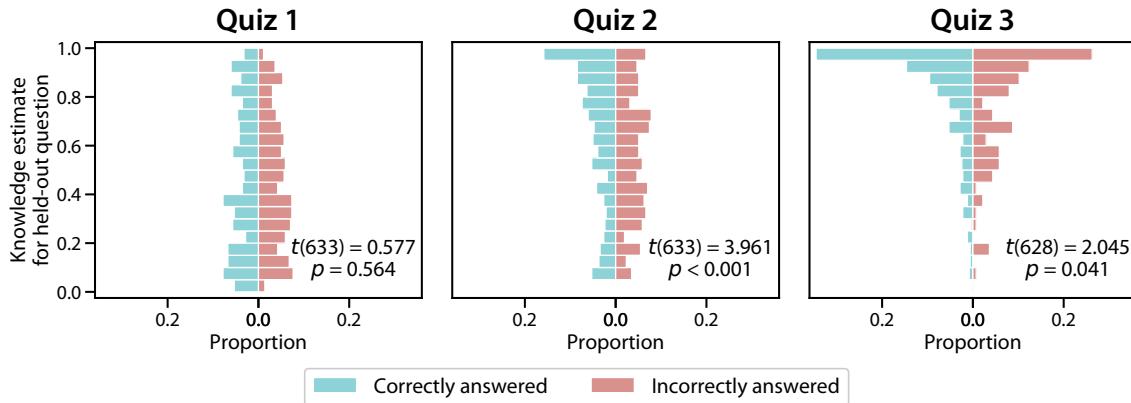


Figure 6: Estimating knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held out correctly (blue) or incorrectly (red) answered question. The t -tests reported in each panel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

coordinates, we found no reliable differences in the estimates when the held-out question had been correctly versus incorrectly answered ($t(633) = 0.577, p = 0.564$). After watching the first video, estimated knowledge for held-out correctly answered questions (from the second quiz; Fig. 6, middle panel) exhibited a positive shift relative to held-out incorrectly answered questions ($t(633) = 3.961, p < 0.001$). After watching the second video, estimated knowledge (from the third quiz; Fig. 6, right panel) for *all* questions exhibited a positive shift. However, the increase in estimated knowledge for held-out correctly answered questions was larger than for held-out incorrectly answered questions ($t(628) = 2.045, p = 0.041$).

Knowledge estimates need not be limited to the content of the lectures. As illustrated in Figure 7, our general approach to estimating knowledge from a small number of quiz questions may be applied to *any* content, given its text embedding coordinate. To visualize how knowledge “spreads” through text embedding space to content beyond the lectures participants watched, we first fit a new topic model to the lectures’ sliding windows with $k = 100$ topics. We hoped that increasing the number of topics from 15 to 100 might help us to generalize the knowledge predictions. (Aside from increasing the number of topics from 15 to 100, all other procedures and model parameters were carried over from the preceding analyses.) As in our other analyses, we

288 resampled each lecture’s topic trajectory to 1 Hz and also projected each question into a shared
289 text embedding space.

290 We projected the resulting 100-dimensional topic vectors (for each second of video and for
291 each question) into a shared 2-dimensional space (see Creating knowledge and learning map
292 visualizations). Next, we sampled points evenly from a 100×100 grid of coordinates that evenly
293 tiled a rectangle enclosing the 2D projections of the videos and questions. We used Equation 4 to
294 estimate participants’ knowledge at each of these 10,000 sampled locations, and we averaged these
295 estimates across participants to obtain an estimated average *knowledge map* (Fig. 7). Intuitively,
296 the knowledge map constructed from a given quiz’s responses provides a visualization of how
297 “much” participants know about any content expressible by the fitted text embedding model.

298 Several features of the resulting knowledge maps are worth noting. The average knowledge
299 map estimated from Quiz 1 responses (Fig. 7, leftmost map) shows that participants tended to
300 have relatively little knowledge about any parts of the text embedding space (i.e., the shading
301 is relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a
302 marked increase in knowledge on the left side of the map (around roughly the same range of
303 coordinates covered by the *Four Fundamental Forces* lecture, indicated by the dotted blue line).
304 In other words, participants’ estimated increase in knowledge is localized to conceptual content
305 that is nearby (i.e., related to) the content from the lecture they watched prior to taking Quiz
306 2. This localization is non-trivial: the knowledge estimates are informed only by the embedded
307 coordinates of the *quiz questions*, not by the embeddings of either lecture (Eqn. 4). Finally, the
308 knowledge map estimated from Quiz 3 responses shows a second increase in knowledge, localized
309 to the region surrounding the embedding of the *Birth of Stars* lecture participants watched prior to
310 taking Quiz 3.

311 Another way of visualizing these content-specific increases in knowledge (apparently driven
312 by watching each lecture) is displayed in Figure 7B. Taking the point-by-point difference between
313 the knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map*
314 that describes the *change* in knowledge estimates from one quiz to the next. These learning maps
315 highlight that the estimated knowledge increases we observed across maps were specific to the



Figure 7: 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 each lecture and the coordinates of each question are indicated by dotted lines and dots. Each map reflects an average across all participants. For individual participants’ maps, see Figures S2, S3, and S4. **B.** Average “learning maps” estimated between each successive pair of quizzes. The learning maps are in the same general format as the knowledge maps in Panel A, but each coordinate in the learning maps indicates the difference between the corresponding coordinates in the indicated pair of knowledge maps—i.e., how much the estimated knowledge “changed” across the two quizzes. Each map reflects an average across all participants. For individual participants’ maps, see Figures S5 and S6. **C.** Word clouds for sampled points in topic space. Each word cloud displays the relative weights of each word (via their relative sizes) reflected by the blend of topics represented at the locations of the stars in the maps. The words’ colors indicate how much each word is weighted, on average, across all timepoints’ topic vectors in the *Four Fundamental Forces* (blue) and *Birth of Stars* (green) videos, respectively.

316 regions around the embeddings of each lecture in turn.

317 Because the 2D projection we used to construct the knowledge and learning maps is invertible,
318 we may gain additional insights into the estimates by reconstructing the original high-dimensional
319 topic vectors for any point(s) in the maps we are interested in. For example, this could serve as
320 a useful tool for an instructor looking to better understand which content areas a student (or a
321 group of students) knows well (or poorly). As a demonstration, we show the top-weighted words
322 from the blends of topics reconstructed from three example locations on the maps (Fig. 7C): one
323 point near the *Four Fundamental Forces* embedding (yellow); a second point near the *Birth of Stars*
324 embedding (orange), and a third point somewhere in between the two lectures' embeddings (pink).
325 As shown in the word clouds in the Panel, the top-weighted words at the example coordinate near
326 the *Four Fundamental Forces* embedding also tended to be weighted heavily by the topics expressed
327 in that lecture. Similarly, the top-weighted words at the example coordinate near the *Birth of Stars*
328 embedding tended to be weighted most heavily by the topics expressed in *that* lecture. And the
329 top-weighted words at the example coordinate between the two lectures' embeddings show a
330 roughly even mix of words most strongly associated with each lecture.

331 Discussion

332 We developed a computational framework that uses short multiple choice quizzes to provide
333 nuanced insights into what learners know and how their knowledge changes with training. First,
334 we show that our approach can automatically match up the conceptual content of individual
335 questions with the corresponding moments in lecture videos where those concepts were presented
336 (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment “knowledge traces”
337 that reflect how much knowledge participants have before and after watching each lecture video
338 (Fig. 5). We also show that these knowledge estimates can generalize to held-out questions (Fig. 6).
339 Finally, we use our framework to construct visual maps that provide snapshot estimates of how
340 much participants know about any concept within the scope of our text embedding model, and
341 how much their knowledge changes with training (Fig. 7).

342 Over the past several years, the global pandemic has forced many educators to teach re-
343 motely [21, 33, 41, 44]. This change in world circumstances is happening alongside (and perhaps
344 accelerating) geometric growth in the availability of high quality online courses on platforms such
345 as Khan Academy [22], Coursera [45], EdX [24], and others [38]. Continued expansion of the global
346 internet backbone and improvements in computing hardware have also facilitated improvements
347 in video streaming, enabling videos to be easily downloaded and shared by large segments of the
348 world’s population. This exciting time for online course instruction provides an opportunity to
349 re-evaluate how we, as a global community, educate ourselves and each other. For example, we
350 can ask: what makes an effective course or training program? Which aspects of teaching might be
351 optimized or automated? How can we provide How and why do learning needs and goals vary
352 across people? How might we lower barriers to achieving a high quality education?

353 Alongside these questions, there is a growing desire to extend existing theories beyond the
354 domain of lab testing rooms and into real classrooms [20]. In part, this has led to a recent
355 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better
356 reflect more ethologically valid phenomena that are more directly relevant to real-world situations
357 and behaviors [34]. In turn, this has brought new challenges in data analysis and interpretation. A
358 key step towards solving these challenges will be to build explicit models of real-world scenarios
359 and how people behave in them (e.g., models of how people learn conceptual content from real-
360 world courses, as in our current study). A second key step will be to understand which sorts
361 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 1,
362 12, 31, 35, 36] might help to inform these models. A third major step will be to develop and
363 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic
364 paradigms.

365 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also
366 relate to the notion of “theory of mind” of other individuals [15, 18, 30]. Considering others’
367 unique perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively
368 interact and communicate [37, 40, 43]. One could imagine future extensions of our work (e.g.,
369 analogous to the knowledge and learning maps shown in Fig. 7), that attempt to characterize how

370 well-aligned different people's knowledge bases or backgrounds are. In turn, this might be used to
371 model how knowledge flow not just between teachers and students, but between friends having
372 a conversation, individuals out on a first date, participants at a business meeting, doctors and
373 patients, experts and non-experts, political allies and adversaries, and more. For example, the
374 extent to which two people's knowledge maps "match" in a given region of text embedding space
375 might serve as a predictor of how effectively the people will communicate about the corresponding
376 conceptual content.

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

383 Materials and methods

384 Participants

385 We enrolled a total of 50 Dartmouth undergraduate students in our study. Participants received
386 course credit for enrolling. We asked each participant to fill out a demographic survey that included
387 questions about their age, gender, native spoken language, ethnicity, race, hearing, color vision,
388 sleep, coffee consumption, level of alertness, and several aspects of their educational background
389 and prior coursework.

390 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
391 years). A total of 15 participants reported their gender as male and 35 participants reported their
392 gender as female. A total of 49 participants reported their native language as "English" and 1
393 reported having another native language. A total of 47 participants reported their ethnicity as
394 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants

395 reported their races as White (32 participants), Asian (14 participants), Black or African American
396 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
397 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

410 Participants reported their undergraduate major(s) as “social sciences” (28 participants), “nat-
411 ural sciences” (16 participants), “professional” (e.g., pre-med or pre-law; 8 participants), “mathe-
412 matics and engineering” (7 participants), “humanities” (4 participants), or “undecided” (3 partici-
413 pants). Note that some participants selected multiple categories for their undergraduate major. We
414 also asked participants about the courses they had taken. In total, 45 participants reported having
415 taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan
416 Academy courses. Of those who reported having watched at least one Khan Academy course,
417 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8
418 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We
419 also asked participants about the specific courses they had watched, categorized under different
420 subject areas. In the “Mathematics” area, participants reported having watched videos on AP
421 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-
422 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry

423 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential
424 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),
425 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other
426 videos not listed in our survey (5 participants). In the “Science and engineering” area, participants
427 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-
428 ipants); Physics, AP Physics I, or AP Physics II (15 participants); Biology, AP Biology; or High
429 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed
430 in our survey (19 participants). We also asked participants whether they had specifically seen the
431 videos used in our experiment. Of the 45 participants who reported having having taken at least
432 one Khan Academy course in the past, 44 participants reported that they had not watched the *Four*
433 *Fundamental Forces* video, and 1 participant reported that they were not sure whether they had
434 watched it. All participants reported that they had not watched the *Birth of Stars* video. When
435 we asked participants about non-Khan Academy online courses, they reported having watched
436 or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test
437 preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 partic-
438 ipants), Computing (2 participants), and other categories not listed in our survey (18 participants).
439 Finally, we asked participants about in-person courses they had taken in different subject areas.
440 They reported taking courses in Mathematics (39 participants), Science and engineering (38 par-
441 ticipants), Arts and humanities (35 participants), Test preparation (27 participants), Economics
442 and finance (26 participants), Computing (15 participants), College and careers (7 participants), or
443 other courses not listed in our survey (6 participants).

444 Experiment

445 We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces*
446 (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force;
447 duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed;
448 duration: 7 minutes and 57 seconds). We then hand-created 39 multiple-choice questions: 15 about
449 the conceptual content of *Four Fundamental Forces* (i.e., lecture 1), 15 about the conceptual content

450 of *Birth of Stars* (i.e., lecture 2), and 9 questions that tested for general conceptual knowledge about
451 basic physics (covering material that was not presented in either video). The full set of questions
452 and answer choices may be found in Table S1.

453 Over the course of the experiment, participants completed three 13-question multiple-choice
454 quizzes: the first before viewing lecture 1, the second between lectures 1 and 2, and the third
455 after viewing lecture 2 (Fig. 1). The questions appearing on each quiz, for each participant, were
456 randomly chosen from the full set of 39, with the constraints that (a) each quiz contain 5 questions
457 about lecture 1, 5 questions about lecture 2, and 3 questions about general physics knowledge, and
458 (b) each question appear exactly once for each participant. The order of questions on each quiz, and
459 the order of answer options for each question, were also randomized. Our experimental protocol
460 was approved by the Committee for the Protection of Human Subjects at Dartmouth College. We
461 used the experiment to develop and test our computational framework for estimating knowledge
462 and learning.

463 Analysis

464 Constructing text embeddings of multiple lectures and questions

465 We adapted an approach we developed in prior work [17] to embed each moment of the two
466 lectures and each question in our pool in a common representational space. Briefly, our approach
467 uses a topic model (Latent Dirichlet Allocation; 4), trained on a set of documents, to discover a
468 set of k “topics” or “themes.” Formally, each topic is defined as a set of weights over each word
469 in the model’s vocabulary (i.e., the union of all unique words, across all documents, excluding
470 “stop words.”). Conceptually, each topic is intended to give larger weights to words that are
471 conceptually related or that tend to co-occur in the same documents. After fitting a topic model,
472 each document in the training set, or any *new* document that contains at least some of the words in
473 the model’s vocabulary, may be represented as a k -dimensional vector describing how much the
474 document (most probably) reflects each topic. (Unless, otherwise noted, we used $k = 15$ topics.)

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

476 sliding windows that span each video’s transcript. Khan Academy provides professionally created,
477 manual transcriptions of all videos for closed captioning. However, such transcripts would not
478 be readily available in all contexts to which our framework could potentially be applied. Khan
479 Academy videos are hosted on the YouTube platform, which additionally provides automated
480 captions. We opted to use these automated transcripts (which, in prior work, we have found are
481 sufficiently near-human quality yield reliable data in behavioral studies; 46) when developing our
482 framework in order to make it more easily extensible and adaptable by others in the future.

483 We fetched these automated transcripts using the `youtube-transcript-api` Python pack-
484 age [11]. The transcripts consisted of one timestamped line of text for every few seconds (mean:
485 2.34 s; standard deviation: 0.83 s) of spoken content in the video (i.e., corresponding to each indi-
486 vidual caption that would appear on-screen if viewing the lecture via YouTube, and when those
487 lines would appear). We defined a sliding window length of (up to) $w = 30$ transcript lines, and
488 assigned each window a timestamp corresponding to the midpoint between its first and last lines’
489 timestamps. These sliding windows ramped up and down in length at the very beginning and
490 end of the transcript, respectively. In other words, the first sliding window covered only the first
491 line from the transcript; the second sliding window covered the first two lines; and so on. This
492 insured that each line of the transcript appeared in the same number (w) of sliding windows. After
493 performing various standard text preprocessing (e.g., normalizing case, lemmatizing, removing
494 punctuation and stop-words), we treated the text from each sliding window as a single “doc-
495 ument,” and we combined these documents across the two videos’ windows to create a single
496 training corpus for the topic model. The top words from each of the 15 discovered topics may be
497 found in Table S2.

498 After fitting a topic model to each videos’ transcripts, we could use the trained model to
499 transform arbitrary (potentially new) documents into k -dimensional topic vectors. A convenient
500 property of these topic vectors is that documents that reflect similar blends of topics (i.e., documents
501 that reflect similar themes, according to the model) will yield similar coordinates (in terms of
502 Euclidean distance, correlation, or other geometric measures). In general, the similarity between
503 different documents’ topic vectors may be used to characterize the similarity in conceptual content

504 between the documents.

505 We transformed each sliding window’s text into a topic vector, and then used linear interpolation
506 (independently for each topic dimension) to resample the resulting timeseries to one vector
507 per second. We also used the fitted model to obtain topic vectors for each question in our pool
508 (Tab. S1). Taken together, we obtained a *trajectory* for each video, describing its path through topic
509 space, and a single coordinate for each question (Fig. 2C). Embedding both videos and all of the
510 questions using a common model enables us to compare the content from different moments of
511 videos, compare the content across videos, and estimate potential associations between specific
512 questions and specific moments of video.

513 **Estimating dynamic knowledge traces**

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

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

516 where

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

517 and where mincorr and maxcorr are the minimum and maximum correlations between any lecture
518 timepoint and question, taken over all timepoints and questions across both lectures and all five
519 question used to estimate the knowledge trace. We also define $f(s, \Omega)$ as the s^{th} topic vector from
520 the set of topic vectors Ω . Here t indexes the set of lecture topic vectors, L , and i and j index the
521 topic vectors of questions used to estimate the knowledge trace, Q . Note that “correct” denotes
522 the set of indices of the questions the participant answered correctly on the given quiz.

523 Intuitively, $\text{ncorr}(x, y)$ is the correlation between two topic vectors (e.g., the topic vector from one
524 timepoint in a lecture, x , and the topic vector for one question, y), normalized by the minimum and
525 maximum correlations (across all timepoints t and questions Q) to range between 0 and 1, inclusive.
526 Equation 1 then computes the weighted average proportion of correctly answered questions about

527 the content presented at timepoint t , where the weights are given by the normalized correlations
528 between timepoint t 's topic vector and the topic vectors for each question. The normalization
529 step (i.e., using ncorr instead of the raw correlations) insures that every question contributes some
530 non-zero amount to the knowledge estimate.

531 **Creating knowledge and learning map visualizations**

532 An important feature of our approach is that, given a trained text embedding model and partic-
533 ipants' quiz performance on each question, we can estimate their knowledge about *any* content
534 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-
535 tions or even appearing in the lectures. To visualize these estimates (Figs. 7, S2, S3, S4, S5, and S6),
536 we used Uniform Manifold Approximation and Projection (UMAP; 29) to construct a 2D projection
537 of the text embedding space. Sampling the original 100-dimensional space at high resolution to
538 obtain an adequate set of topic vectors spanning the embedding space would be computationally
539 intractable. However, sampling a 2D grid is trivial.

540 At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing
541 the cross entropy between the pairwise (clustered) distances between the observations in their
542 original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional
543 embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise
544 distances in the original high-dimensional space were defined as 1 minus the correlation between
545 the pair of coordinates, and pairwise distances in the low-dimensional embedding space were
546 defined as the Euclidean distance between the pair of coordinates.

547 Although UMAP is not fully invertible, due to its clustering step, it is possible to approximately
548 invert the embedded low-dimensional points. In our application, we sought to embed topic
549 vectors, whose elements are always non-negative. Because the inversion step is inexact, sometimes
550 it results in negative-valued vectors (even if the original observations were all non-negative),
551 which are incompatible with the topic modeling framework. To protect against this issue, we
552 log-transformed the topic vectors prior to embedding them in the 2D space. When we inverted
553 the embedded vectors (e.g., to estimate topic vectors or word clouds, as in Fig. 7C), we passed

554 to inverted (log-transformed) values through the exponential function to obtain a vector of non-
555 negative values.

556 After embedding both lectures' topic trajectories and the topic vectors of every question, we
557 defined a rectangle enclosing the 2D projections of the lectures' and quizzes' embeddings. We then
558 sampled points from a regular 100×100 grid of coordinates that evenly tiled this enclosing rectangle.
559 We sought to estimate participants' knowledge (and learning, i.e., changes in knowledge) at each
560 of the resulting 10K coordinates.

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

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

565 The λ term in the RBF equation controls the "smoothness" of the function, where larger values
566 of λ result in smoother maps. In our implementation we used $\lambda = 50$. Next, we estimated the
567 "knowledge" at each coordinate, x , using:

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

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

572 Author contributions

573 Conceptualization: PCF, ACH, and JRM. Methodology: PCF, ACH, and JRM. Software: PCF.
574 Validation: PCF. Formal analysis: PCF. Resources: PCF, ACH, and JRM. Data curation: PCF.
575 Writing (original draft): JRM. Writing (review and editing): PCF, ACH, and JRM. Visualization:

576 PCF and JRM. Supervision: JRM. Project administration: PCF. Funding acquisition: JRM.

577 **Data and code availability**

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

581 **Acknowledgements**

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

589 **References**

- 590 [1] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and
591 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
592 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 593 [2] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
594 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
595 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 596 [3] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*

- 597 Conference on Machine Learning, pages 113–120, New York, NY. Association for Computing
598 Machinery.
- 599 [4] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
600 *Learning Research*, 3:993–1022.
- 601 [5] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
602 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
603 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
604 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
605 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 606 [6] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
607 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 608 [7] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
609 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
610 sentence encoder. *arXiv*, 1803.11175.
- 611 [8] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
612 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 613 [9] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
614 Evidence for a new conceptualization of semantic representation in the left and right cerebral
615 hemispheres. *Cortex*, 40(3):467–478.
- 616 [10] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
617 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,
618 41(6):391–407.
- 619 [11] Depoix, J. (2019). YouTube transcript/subtitle API. <https://github.com/jdepoix/youtube-transcript-api>.

- 621 [12] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
622 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
623 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 624 [13] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical
625 Transactions of the Royal Society A*, 222(602):309–368.
- 626 [14] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
627 *School Science and Mathematics*, 100(6):310–318.
- 628 [15] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of
629 Cognition and Development*, 13(1):19–37.
- 630 [16] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
631 learning, pages 212–221. Sage Publications.
- 632 [17] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal be-
633 havioral and neural signatures of transforming naturalistic experiences into episodic memories.
634 *Nature Human Behavior*, 5:905–919.
- 635 [18] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
636 empathy and theory of mind. In *Social Behavior from Rodents to Humans*, pages 193–206. Springer.
- 637 [19] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
638 Columbia University Press.
- 639 [20] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
640 326(7382):213–216.
- 641 [21] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
642 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International
643 Journal of Environmental Research and Public Health*, 18(5):2672.
- 644 [22] Khan, S. (2004). *The Khan Academy*. Salman Khan.

- 645 [23] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 646 [24] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
647 *The Chronicle of Higher Education*, 21:1–5.
- 648 [25] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic
649 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
650 104:211–240.
- 651 [26] Macellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of
652 Educational Studies*, 53(2):129–147.
- 653 [27] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
654 *Handbook of Human Memory*. Oxford University Press.
- 655 [28] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum "memory wave"
656 function? *Psychological Review*, 128(4):711–725.
- 657 [29] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
658 projection for dimension reduction. *arXiv*, 1802(03426).
- 659 [30] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
660 mind. In *The Wiley-Blackwell handbook of childhood cognitive development*. Wiley-Blackwell.
- 661 [31] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
662 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
663 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 664 [32] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
665 tations in vector space. *arXiv*, 1301.3781.
- 666 [33] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
667 from a national survey of language educators. *System*, 97:102431.

- 668 [34] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
669 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 670 [35] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
671 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*
672 *Neuroscience*, 17(4):367–376.
- 673 [36] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
674 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
675 7:43916.
- 676 [37] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*
677 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 678 [38] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
679 higher education: unmasking power and raising questions about the movement’s democratic
680 potential. *Educational Theory*, 63(1):87–110.
- 681 [39] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
682 Student conceptions and conceptual learning in science. Routledge.
- 683 [40] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
684 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*
685 *tion in Nursing*, 22:32–42.
- 686 [41] Shim, T. E. and Lee, S. Y. (2020). College students’ experience of emergency remote teaching
687 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 688 [42] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
689 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*
690 *Mathematics Education*, 35(5):305–329.
- 691 [43] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal*
692 *Medicine*, 21:524–530.

- 693 [44] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
694 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 695 [45] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
696 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 697 [46] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is
698 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior
699 Research Methods*, 50:2597–2605.