

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 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 a student’s knowledge is to present them with a set of quiz
³⁰ questions, calculate the proportion they answer correctly, and provide them with feedback in the
³¹ form of a simple numeric or letter grade. While such a grade can provide *some* indication of whether
³² the student has mastered the to-be-learned material, any univariate measure of performance on a
³³ complex task sacrifices certain relevant information, risks conflating underlying factors, and so on.
³⁴ For example, consider the relative utility of the imaginary map described above that characterizes
³⁵ a student’s knowledge in detail, versus a single annotation saying that the student answered 85%
³⁶ of 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 instead be used to 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, 43]?

45 Or weaving a lecture’s atomic elements (e.g., its component words) into a structured network that
46 describes how those individual elements are related [26]? 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, 27, 40].

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, 27, 40) 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 clues as to what the document is about, just as memorizing a pas-
54 sage might provide some ability to answer simple questions about it. However, text embedding
55 models (e.g., 3–5, 7, 10, 25, 33) 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 [28, 29]. 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, wherein nearby vectors reflect conceptually related documents. A model
62 that succeeds at capturing an analogue of “understanding” is able to assign nearby feature vectors
63 to two conceptually related documents, *even when the specific words contained in those documents have*
64 *very little overlap.*

65 Given these insights, what form might a representation of the sum total of a person’s knowledge
66 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 conceptual knowledge is acquired during
77 learning. The central idea behind our framework is to use text embedding models to define the
78 coordinate systems of two maps: a *knowledge map* that describes the extent to which each concept is
79 currently known, and a *learning map* that describes changes in knowledge over time. Each location
80 on these maps represents a single concept, and the maps’ geometries are defined such that related
81 concepts are located nearby in space. We use this framework to analyze and interpret behavioral
82 data collected from an experiment that had participants answer sets multiple-choice questions
83 about a series of 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 random
91 features in otherwise meaningless categorized stimuli. However the objective of real-world train-
92 ing, or learning from life experiences more generally, is often to develop new knowledge that may
93 be applied in *useful* ways in the future. In this sense, the gap between modern learning theories and
94 modern pedagogical approaches that inform 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, 27].
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

101 that such maps might lead to real-world tools for improving how we educate. Taken together, our
102 work shows that existing course materials and evaluative tools like short multiple-choice quizzes
103 may be leveraged to gain highly detailed insights into what students know and how they learn.

104 Results

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

119 We asked participants in our study to complete brief multiple-choice quizzes before, between,
120 and after watching two lecture videos from the Khan Academy [22] platform (Fig. 1). The first
121 lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics:
122 gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*,
123 provided an overview of our current understanding of how stars form. We selected these particular
124 lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad
125 audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training on
126 our participants’ abilities to learn from the lectures. To this end, we selected two introductory videos



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.

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

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

To study in detail how participants' conceptual knowledge changed over the course of the

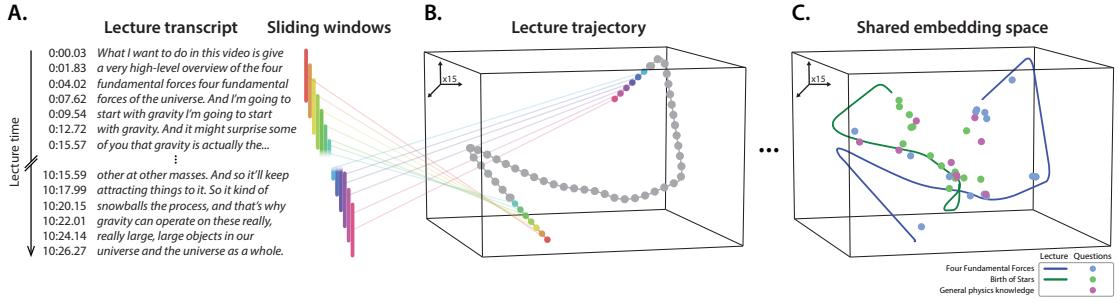


Figure 2: Modeling course content. **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The full set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding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

145 experiment, we first sought to model the conceptual content presented to them at each moment
 146 throughout each of the two lectures. We adapted an approach we developed in prior work [17] to
 147 identify the latent themes in the lectures using a topic model [4]. Briefly, topic models take as input
 148 a collection of text documents, and learn a set of “topics” (i.e., latent themes) from their contents.
 149 Once fit, a topic model can be used to transform arbitrary (potentially new) documents into sets
 150 of “topic proportions,” describing the weighted blend of learned topics reflected in their texts. We
 151 parsed automatically generated transcripts of the two lectures into overlapping sliding windows,
 152 where each window contained the text of the lecture transcript from a particular time span. We
 153 treated the set of text snippets (across all of these windows) as documents to fit our model (Fig. 2A;
 154 see *Constructing text embeddings of multiple lectures and questions*). Transforming the text from every
 155 sliding window with our model yielded a number-of-windows by number-of-topics (15) topic-
 156 proportions matrix that described the unique mixture of broad themes from both lectures reflected
 157 in each window’s text. Each window’s “topic vector” (i.e., column of the topic-proportions matrix)
 158 is analogous to a coordinate in a 15-dimensional space whose axes are topics discovered by the
 159 model. Within this space, each lecture’s sequence of topic vectors (i.e., corresponding to its

160 transcript’s overlapping text snippets across sliding windows) forms a *trajectory* that captures how
161 its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution
162 of one topic vector for each second of video (i.e., 1 Hz).

163 We hypothesized that a topic model trained on transcripts of the two lectures should also capture
164 the conceptual knowledge probed by each quiz question. If indeed the topic model could capture
165 information about the deeper conceptual content of the lectures (i.e., beyond surface-level details
166 such as particular word choices), then we should be able to recover a correspondence between each
167 lecture and questions *about* each lecture. Importantly, such a correspondence could not solely arise
168 from superficial text matching between lecture transcripts and questions, since the lectures and
169 questions used different words. Simply comparing the average topic weights from each lecture and
170 question set (averaging across time and questions, respectively) reveals a striking correspondence
171 (Fig. S1). Specifically, the average topic weights from lecture 1 are strongly correlated with the
172 average topic weights from lecture 1 questions ($r(13) = 0.809, p < 0.001$, 95% confidence interval
173 (CI) = [0.633, 0.962]), and the average topic weights from lecture 2 are strongly correlated with the
174 average topic weights from lecture 2 questions ($r(13) = 0.728, p = 0.002$, 95% CI = [0.456, 0.920]).
175 At the same time, the average topic weights from the two lectures are *negatively* correlated with
176 their non-matching question sets (lecture 1 video vs. lecture 2 questions: $r(13) = -0.547, p = 0.035$,
177 95% CI = [-0.812, -0.231]; lecture 2 video vs. lecture 1 questions: $r(13) = -0.612, p = 0.015$, 95%
178 CI = [-0.874, -0.281]), indicating that the topic model also exhibits some degree of specificity. The
179 full set of pairwise comparisons between average topic weights for the lectures and question sets
180 is reported in Figure S1.

181 Another, more sensitive, way of summarizing the conceptual content of the lectures and ques-
182 tions is to look at *variability* in how topics are weighted over time and across different questions
183 (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “infor-
184 mation” [13] the lecture (or question set) reflects about that topic. For example, suppose a given
185 topic is weighted on heavily throughout a lecture. That topic might be characteristic of some
186 aspect or property of the lecture *overall* (conceptual or otherwise), but unless the topic’s weights
187 changed in meaningful ways over time, the topic would be a poor indicator of any *specific* concep-



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 set. Each row and column corresponds to a bar plot in Panel A.

tual 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 ($r(13) = 0.824, p < 0.001, 95\% \text{ CI} = [0.696, 0.973]$) and the lecture 2 video and questions ($r(13) = 0.801, p < 0.001, 95\% \text{ CI} = [0.539, 0.958]$). However, as reported in Figure 3B, the variability in topic expressions across *different* videos and lecture-specific questions (i.e., lecture 1 video vs. lecture 2 questions; lecture 2 video vs. lecture 1 questions) were negatively correlated, and neither video’s topic variability was reliably correlated with the topic variability across general physics knowledge questions. Taken together, the analyses reported in Figures 3 and S1 indicate that a topic model fit to the videos’ transcripts can also reveal correspondences (at a coarse scale) between the lectures and questions.

Although a single lecture may be organized around a single broad theme at a coarse scale, at a finer scale each moment of a lecture typically covers a narrower range of content. We wondered whether a text embedding model trained on the lectures’ transcripts might capture some of this finer scale content. For example, if a particular question asks about the content from one small part

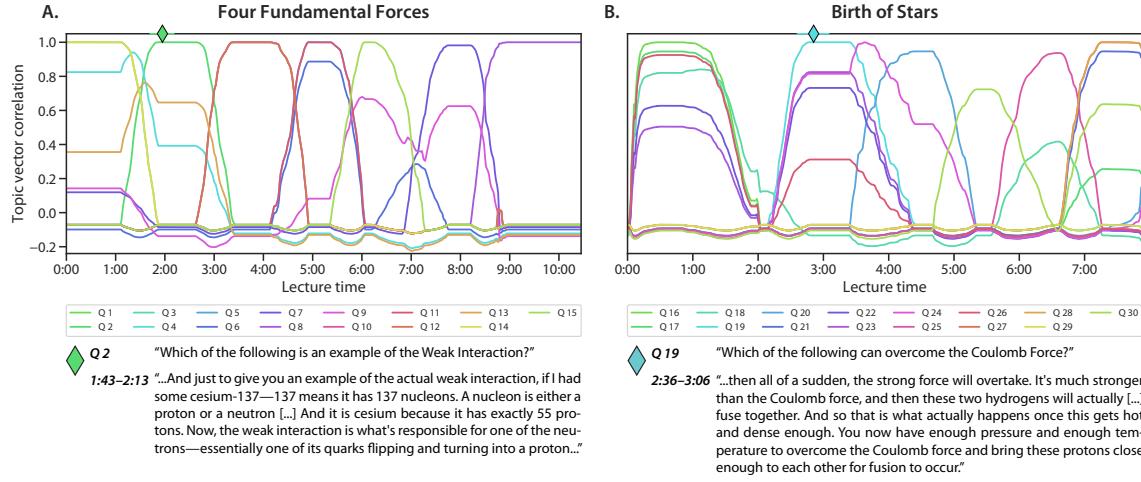


Figure 4: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel B.: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

of a lecture, we wondered whether the text embeddings could be used to automatically identify the “matching” moment(s) in the lecture. When we correlated each question’s topic vector with the topic vectors from each second of the lectures, we found some evidence that each question is temporally specific (Fig. 4). In particular, most questions’ topic vectors were maximally correlated with a well-defined (and relatively narrow) range of timepoints from their corresponding lectures, and the correlations fell off sharply outside of that range. We also qualitatively examined the best-matching intervals for each question by comparing the text of the question to the text of the most-correlated parts of the lectures. Despite that the questions were excluded from the text embedding model’s training set, in general we found (through manual inspection) a close correspondence between the conceptual content that each question probed and the content covered by the best-matching moments of the lectures. Two representative examples are shown at the bottom of Figure 4.

The ability to quantify how much each question is “asking about” the content from each moment

215 of the lectures could enable high-resolution insights into participants' knowledge. Traditional
216 approaches to estimating how much a student "knows" about the content of a given lecture entail
217 computing the proportion of correctly answered questions. But if two students receive identical
218 scores on an exam, might our modeling framework help us to gain more nuanced insights into
219 the *specific* content that each student has mastered (or failed to master)? For example, a student
220 who misses three questions that were all about the same concept (e.g., concept *A*) will have gotten
221 the same *proportion* of questions correct as another student who missed three questions about
222 three *different* concepts (e.g., *A*, *B*, and *C*). But if we wanted to fill in the "gaps" in the two
223 students' understandings, we might do well to focus on concept *A* for the first student, but to
224 also add in materials pertaining to concepts *B* and *C* for the second student. In other words, raw
225 "proportion-correct" measures may capture *how much* a student knows, but not *what* they know.
226 We wondered whether our modeling framework might enable us to (formally and automatically)
227 infer participants' knowledge at the scale of individual concepts (e.g., as captured by a single
228 moment of a lecture).

229 We developed a simple formula (Eqn. 1) for using a participant's responses to a small set
230 of multiple-choice questions to estimate how much the participant "knows" about the concept
231 reflected by any arbitrary coordinate, x , in text embedding space (e.g., the content reflected by
232 any moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially,
233 the estimated knowledge at the coordinate is given by the weighted average proportion of quiz
234 questions the participant answered correctly, where the weights reflect how much each question
235 is "about" the content at x . When we apply this approach to estimate the participant's knowledge
236 about the content presented in each moment of each lecture, we can obtain a detailed timecourse
237 describing how much "knowledge" the participant has about any part of the lecture. As shown
238 in Figure 5, we can also apply this approach separately for the questions from each quiz the
239 participants took throughout the experiment. From just a few questions per quiz, we obtain a
240 high-resolution snapshot (at the time each quiz was taken) of what the participants knew about
241 any moment's content, from either of the two lectures they watched (comprising a total of 1,100
242 samples across the two lectures).



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.

243 Of course, even though the timecourses in Figure 5A and C provide detailed *estimates* about
244 participants' knowledge, those estimates are only *useful* to the extent that they accurately reflect what
245 participants actually know. As one sanity check, we anticipated that the knowledge estimates
246 should show a content-specific "boost" in participants' knowledge after watching each lecture.
247 In other words, if participants learn about each lecture's content when they watch each lecture,
248 the knowledge estimates should reflect that. After watching the *Four Fundamental Forces* lecture,
249 participants should show more knowledge for the content of that lecture than they had before,
250 and that knowledge should persist for the remainder of the experiment. Specifically, knowledge
251 about that lecture's content should be relatively low when estimated using Quiz 1 responses,
252 but should increase when estimated using Quiz 2 or 3 responses (Fig. 5B). Indeed, we found
253 that participants' estimated knowledge about the content of the *Four Fundamental Forces* was
254 substantially higher on Quiz 2 versus Quiz 1 ($t(49) = 8.764, p < 0.001$) and on Quiz 3 versus Quiz
255 1 ($t(49) = 10.519, p < 0.001$). We found no reliable differences in estimated knowledge about
256 that lecture's content on Quiz 2 versus 3 ($t(49) = 0.160, p = 0.874$). Similarly, we hypothesized
257 (and subsequently confirmed) that participants should show more estimated knowledge about the
258 content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 5D). Specifically, since
259 participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their
260 knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a "boost" on
261 Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge
262 about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ($t(49) = 1.013, p = 0.316$), but the
263 estimated knowledge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and
264 Quiz 3 versus 1 ($t(49) = 8.969, p < 0.001$).

265 If we are able to accurately estimate a participant's knowledge about the content tested by a
266 given question, our estimates of their knowledge should carry some predictive information about
267 whether the participant is likely to answer the question correctly or incorrectly. We developed
268 a statistical approach to test this claim. For each question in turn, for each participant, we used
269 Equation 1 to estimate (using all *other* questions from the same quiz, from the same participant)
270 the participant's knowledge at the held-out question's embedding coordinate. For each quiz, we



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.

grouped these estimates into two distributions: one for the estimated knowledge at the coordinates of each *correctly* answered question, and another for the estimated knowledge at the coordinates of each *incorrectly* answered question (Fig. 6). We then used independent samples t -tests to compare the means of these distributions of estimated knowledge.

For the initial quizzes participants took (prior to watching either lecture), participants' estimated knowledge tended to be low overall, and relatively unstructured (Fig. 6, left panel). When we held out individual questions and estimated their knowledge at the held-out questions' embedding 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 (estimated knowledge for correctly versus incorrectly answered Quiz 3 questions: $t(628) = 2.045, p = 0.041$).

287 Knowledge estimates need not be limited to the content of the lectures. As illustrated in
288 Figure 7, our general approach to estimating knowledge from a small number of quiz questions
289 may be applied to *any* content, given its text embedding coordinate. To visualize how knowledge
290 “spreads” through text embedding space to content beyond the lectures participants watched,
291 we first fit a new topic model to the lectures’ sliding windows with $k = 100$ topics. We hoped
292 that increasing the number of topics from 15 to 100 might help us to generalize the knowledge
293 predictions. (Aside from increasing the number of topics from 15 to 100, all other procedures and
294 model parameters were carried over from the preceding analyses.) As in our other analyses, we
295 resampled each lecture’s topic trajectory to 1 Hz and also projected each question into a shared
296 text embedding space.

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

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



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 are indicated by dotted lines, and the coordinates of each question are indicated by 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 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 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 on 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.

315 estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region
316 surrounding the embedding of the *Birth of Stars* lecture participants watched immediately prior to
317 taking Quiz 3.

318 Another way of visualizing these content-specific increases in knowledge after participants
319 viewed each lecture is displayed in Figure 7B. Taking the point-by-point difference between the
320 knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map*
321 that describes the *change* in knowledge estimates from one quiz to the next. These learning maps
322 highlight that the estimated knowledge increases we observed across maps were specific to the
323 regions around the embeddings of each lecture in turn.

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

338 Discussion

339 We developed a computational framework that uses short multiple-choice quizzes to gain nuanced
340 insights into what learners know and how their knowledge changes with training. First, we show

341 that our approach can automatically match the conceptual knowledge probed by individual quiz
342 questions to the corresponding moments in lecture videos when those concepts were presented
343 (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment “knowledge traces”
344 that reflect the degree of knowledge participants have about each video’s time-varying content,
345 and capture temporally specific increases in knowledge after viewing each lecture (Fig. 5). We
346 also show that these knowledge estimates can generalize to held-out questions (Fig. 6). Finally,
347 we use our framework to construct visual maps that provide snapshot estimates of how much
348 participants know about any concept within the scope of our text embedding model, and how
349 much their knowledge changes with training (Fig. 7).

350 Over the past several years, the global pandemic has forced many educators to teach re-
351 motely [21, 34, 42, 45]. This change in world circumstances is happening alongside (and perhaps
352 accelerating) geometric growth in the availability of high quality online courses on platforms such
353 as Khan Academy [22], Coursera [46], EdX [24], and others [39]. Continued expansion of the global
354 internet backbone and improvements in computing hardware have also facilitated improvements
355 in video streaming, enabling videos to be easily shared and viewed by large segments of the
356 world’s population. This exciting time for online course instruction provides an opportunity to
357 re-evaluate how we, as a global community, educate ourselves and each other. For example, we
358 can ask: what makes an effective course or training program? Which aspects of teaching might
359 be optimized and/or augmented by automated tools? How and why do learning needs and goals
360 vary across people? How might we lower barriers to achieving a high-quality education?

361 Alongside these questions, there is a growing desire to extend existing theories beyond the
362 domain of lab testing rooms and into real classrooms [20]. In part, this has led to a recent
363 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better
364 reflect more ethologically valid phenomena that are more directly relevant to real-world situations
365 and behaviors [35]. In turn, this has brought new challenges in data analysis and interpretation. A
366 key step towards solving these challenges will be to build explicit models of real-world scenarios
367 and how people behave in them (e.g., models of how people learn conceptual content from real-
368 world courses, as in our current study). A second key step will be to understand which sorts of

369 signals derived from behaviors and/or other measurements (e.g., neurophysiological data; 1, 12, 32,
370 36, 37) might help to inform these models. A third major step will be to develop and employ reliable
371 ways of evaluating the complex models and data that are a hallmark of naturalistic paradigms.

372 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also
373 relate to the notion of “theory of mind” of other individuals [15, 18, 31]. Considering others’ unique
374 perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and
375 communicate [38, 41, 44]. One could imagine future extensions of our work (e.g., analogous to
376 the knowledge and learning maps shown in Fig. 7), that attempt to characterize how well-aligned
377 different people’s knowledge bases or backgrounds are. In turn, this might be used to model how
378 knowledge (or other forms of communicable information) flows not just between teachers and
379 students, but between friends having a conversation, individuals on a first date, participants at
380 a business meeting, doctors and patients, experts and non-experts, political allies or adversaries,
381 and more. For example, the extent to which two people’s knowledge maps “match” or “align” in
382 a given region of text embedding space might serve as a predictor of how effectively they will be
383 able to communicate about the corresponding conceptual content.

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

390 Materials and methods

391 Participants

392 We enrolled a total of 50 Dartmouth undergraduate students in our study. Participants received
393 course credit for enrolling. We asked each participant to complete a demographic survey that

394 included questions about their age, gender, native spoken language, ethnicity, race, hearing, color
395 vision, sleep, coffee consumption, level of alertness, and several aspects of their educational back-
396 ground and prior coursework.

397 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
398 years). A total of 15 participants reported their gender as male and 35 participants reported their
399 gender as female. A total of 49 participants reported their native language as "English" and 1
400 reported having another native language. A total of 47 participants reported their ethnicity as
401 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants
402 reported their races as White (32 participants), Asian (14 participants), Black or African American
403 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
404 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

417 Participants reported their undergraduate major(s) as "social sciences" (28 participants), "nat-
418 ural sciences" (16 participants), "professional" (e.g., pre-med or pre-law; 8 participants), "mathe-
419 matics and engineering" (7 participants), "humanities" (4 participants), or "undecided" (3 partici-
420 pants). Note that some participants selected multiple categories for their undergraduate major. We
421 also asked participants about the courses they had taken. In total, 45 participants reported having

422 taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan
423 Academy courses. Of those who reported having watched at least one Khan Academy course,
424 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8
425 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We
426 also asked participants about the specific courses they had watched, categorized under different
427 subject areas. In the “Mathematics” area, participants reported having watched videos on AP
428 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-
429 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry
430 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential
431 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),
432 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other
433 videos not listed in our survey (5 participants). In the “Science and engineering” area, participants
434 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-
435 ipants); Physics, AP Physics I, or AP Physics II (15 participants); Biology, AP Biology; or High
436 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed
437 in our survey (19 participants). We also asked participants whether they had specifically seen the
438 videos used in our experiment. Of the 45 participants who reported having having taken at least
439 one Khan Academy course in the past, 44 participants reported that they had not watched the *Four*
440 *Fundamental Forces* video, and 1 participant reported that they were not sure whether they had
441 watched it. All participants reported that they had not watched the *Birth of Stars* video. When
442 we asked participants about non-Khan Academy online courses, they reported having watched
443 or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test
444 preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 partic-
445 ipants), Computing (2 participants), and other categories not listed in our survey (18 participants).
446 Finally, we asked participants about in-person courses they had taken in different subject areas.
447 They reported taking courses in Mathematics (39 participants), Science and engineering (38 par-
448 ticipants), Arts and humanities (35 participants), Test preparation (27 participants), Economics
449 and finance (26 participants), Computing (15 participants), College and careers (7 participants), or

450 other courses not listed in our survey (6 participants).

451 **Experiment**

452 We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces*
453 (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force;
454 duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed;
455 duration: 7 minutes and 57 seconds). We then hand-created 39 multiple-choice questions: 15 about
456 the conceptual content of *Four Fundamental Forces* (i.e., lecture 1), 15 about the conceptual content
457 of *Birth of Stars* (i.e., lecture 2), and 9 questions that tested for general conceptual knowledge about
458 basic physics (covering material that was not presented in either video). The full set of questions
459 and answer choices may be found in Table S1.

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

470 **Analysis**

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

472 We adapted an approach we developed in prior work [17] to embed each moment of the two
473 lectures and each question in our pool in a common representational space. Briefly, our approach
474 uses a topic model (Latent Dirichlet Allocation; 4), trained on a set of documents, to discover a set

475 of k “topics” or “themes.” Formally, each topic is defined as a set of weights over each word in
476 the model’s vocabulary (i.e., the union of all unique words, across all documents, excluding “stop
477 words.”). Conceptually, each topic is intended to give larger weights to words that are semantically
478 related or tend to co-occur in the same documents. After fitting a topic model, each document
479 in the training set, or any *new* document that contains at least some of the words in the model’s
480 vocabulary, may be represented as a k -dimensional vector describing how much the document
481 (most probably) reflects each topic. (Unless, otherwise noted, we used $k = 15$ topics.)

482 As illustrated in Figure 2A, we start by building up a corpus of documents using overlapping
483 sliding windows that span each video’s transcript. Khan Academy provides professionally created,
484 manual transcriptions of all videos for closed captioning. However, such transcripts would not
485 be readily available in all contexts to which our framework could potentially be applied. Khan
486 Academy videos are hosted on the YouTube platform, which additionally provides automated
487 captions. We opted to use these automated transcripts (which, in prior work, we have found are
488 of sufficiently near-human quality yield reliable data in behavioral studies; 47) when developing
489 our framework in order to make it more directly extensible and adaptable by others in the future.

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

503 corpus for the topic model. The top words from each of the 15 discovered topics may be found in
504 Table S2.

505 After fitting a topic model to the two videos' transcripts, we could use the trained model to
506 transform arbitrary (potentially new) documents into k -dimensional topic vectors. A convenient
507 property of these topic vectors is that documents that reflect similar blends of topics (i.e., documents
508 that reflect similar themes, according to the model) will yield similar coordinates (in terms of
509 Euclidean distance, correlation, or other geometric measures). In general, the similarity between
510 different documents' topic vectors may be used to characterize the similarity in conceptual content
511 between the documents.

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

520 **Estimating dynamic knowledge traces**

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

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

523 where

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

524 and where mincorr and maxcorr are the minimum and maximum correlations between any lecture
525 timepoint and question, taken over all timepoints in the given lecture, and all five questions *about*

526 that lecture appearing on the given quiz. We also define $f(s, \Omega)$ as the s^{th} topic vector from the set
527 of topic vectors Ω . Here t indexes the set of lecture topic vectors, L , and i and j index the topic
528 vectors of questions used to estimate the knowledge trace, Q . Note that “correct” denotes the set
529 of indices of the questions the participant answered correctly on the given quiz.

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

538 **Creating knowledge and learning map visualizations**

539 An important feature of our approach is that, given a trained text embedding model and partic-
540 ipants’ quiz performance on each question, we can estimate their knowledge about *any* content
541 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-
542 tions or even appearing in the lectures. To visualize these estimates (Figs. 7, S2, S3, S4, S5, and S6),
543 we used Uniform Manifold Approximation and Projection (UMAP; 30) to construct a 2D projection
544 of the text embedding space. Sampling the original 100-dimensional space at high resolution to
545 obtain an adequate set of topic vectors spanning the embedding space would be computationally
546 intractable. However, sampling a 2D grid is trivial.

547 At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing
548 the cross-entropy between the pairwise (clustered) distances between the observations in their
549 original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional
550 embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise
551 distances in the original high-dimensional space were defined as 1 minus the correlation between
552 the pair of coordinates, and pairwise distances in the low-dimensional embedding space were

553 defied as the Euclidean distance between the pair of coordinates.

554 In our application, all of the coordinates we embedded were topic vectors, whose elements
555 are always non-negative. Although UMAP is an invertible transformation at the embedding
556 locations of the original data, other locations in the embedding space will not necessarily follow
557 the same implicit “rules” as the original high-dimensional data. For example, inverting an arbitrary
558 coordinate in the embedding space might result in negative-valued vectors, which are incompatible
559 with the topic modeling framework. To protect against this issue, we log-transformed the topic
560 vectors prior to embedding them in the 2D space. When we inverted the embedded vectors (e.g.,
561 to estimate topic vectors or word clouds, as in Fig. 7C), we passed the inverted (log-transformed)
562 values through the exponential function to obtain a vector of non-negative values.

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

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

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

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

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

575 Intuitively, Equation 4 computes the weighted proportion of correctly answered questions, where
576 the weights are given by how nearby (in the 2D space) each question is to the x . We also defined

577 learning maps as the coordinate-by-coordinate differences between any pair of knowledge maps.
578 Intuitively, learning maps reflect the *change* in knowledge across two maps.

579 Author contributions

580 Conceptualization: PCF, ACH, and JRM. Methodology: PCF, ACH, and JRM. Software: PCF.
581 Validation: PCF. Formal analysis: PCF. Resources: PCF, ACH, and JRM. Data curation: PCF.
582 Writing (original draft): JRM. Writing (review and editing): PCF, ACH, and JRM. Visualization:
583 PCF and JRM. Supervision: JRM. Project administration: PCF. Funding acquisition: JRM.

584 Data and code availability

585 All of the data analyzed in this manuscript, along with all of the code for running our experiment
586 and carrying out the analyses may be found at <https://github.com/ContextLab/efficient-learning-khan>.
587

588 Acknowledgements

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

596 **References**

- 597 [1] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and
598 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
599 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 600 [2] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
601 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
602 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 603 [3] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International
604 Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
605 Machinery.
- 606 [4] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine
607 Learning Research*, 3:993–1022.
- 608 [5] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
609 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
610 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
611 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
612 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 613 [6] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
614 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 615 [7] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
616 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
617 sentence encoder. *arXiv*, 1803.11175.
- 618 [8] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
619 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.

- 620 [9] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
621 Evidence for a new conceptualization of semantic representation in the left and right cerebral
622 hemispheres. *Cortex*, 40(3):467–478.
- 623 [10] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
624 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,
625 41(6):391–407.
- 626 [11] Depoix, J. (2019). YouTube transcript/subtitle API. <https://github.com/jdepoix/youtube-transcript-api>.
- 628 [12] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
629 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
630 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 631 [13] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical
632 Transactions of the Royal Society A*, 222(602):309–368.
- 633 [14] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
634 *School Science and Mathematics*, 100(6):310–318.
- 635 [15] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of
636 Cognition and Development*, 13(1):19–37.
- 637 [16] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
638 learning, pages 212–221. Sage Publications.
- 639 [17] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal be-
640 havioral and neural signatures of transforming naturalistic experiences into episodic memories.
641 *Nature Human Behavior*, 5:905–919.
- 642 [18] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
643 empathy and theory of mind. In *Social Behavior from Rodents to Humans*, pages 193–206. Springer.

- 644 [19] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
645 Columbia University Press.
- 646 [20] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
647 326(7382):213–216.
- 648 [21] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
649 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International
650 Journal of Environmental Research and Public Health*, 18(5):2672.
- 651 [22] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 652 [23] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 653 [24] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
654 *The Chronicle of Higher Education*, 21:1–5.
- 655 [25] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic
656 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
657 104:211–240.
- 658 [26] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic
659 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.
- 660 [27] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of
661 Educational Studies*, 53(2):129–147.
- 662 [28] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
663 *Handbook of Human Memory*. Oxford University Press.
- 664 [29] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
665 function? *Psychological Review*, 128(4):711–725.
- 666 [30] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
667 projection for dimension reduction. *arXiv*, 1802(03426).

- 668 [31] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
669 mind. In *The Wiley-Blackwell handbook of childhood cognitive development*. Wiley-Blackwell.
- 670 [32] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
671 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
672 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 673 [33] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
674 tations in vector space. *arXiv*, 1301.3781.
- 675 [34] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
676 from a national survey of language educators. *System*, 97:102431.
- 677 [35] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
678 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 679 [36] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
680 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*
681 *Neuroscience*, 17(4):367–376.
- 682 [37] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
683 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
684 7:43916.
- 685 [38] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*
686 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 687 [39] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
688 higher education: unmasking power and raising questions about the movement’s democratic
689 potential. *Educational Theory*, 63(1):87–110.
- 690 [40] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
691 Student conceptions and conceptual learning in science. Routledge.

- 692 [41] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
693 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*
694 *tion in Nursing*, 22:32–42.
- 695 [42] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching
696 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 697 [43] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
698 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*
699 *Mathematics Education*, 35(5):305–329.
- 700 [44] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal*
701 *Medicine*, 21:524–530.
- 702 [45] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
703 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 704 [46] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
705 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 706 [47] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is
707 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*
708 *Research Methods*, 50:2597–2605.