

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

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 Designing and building procedures and tools for mapping out knowledge touches on deep
³⁰ questions about what it means to learn. For example, how do we acquire conceptual knowledge?
³¹ Memorizing course lectures or textbook chapters by rote can lead to the superficial *appearance*
³² of understanding the underlying content, but achieving true conceptual understanding seems
³³ to require something deeper and richer. Does conceptual understanding entail connecting newly
³⁴ acquired information to the scaffolding of one’s existing knowledge or experience [6, 10, 13, 14, 57]?
³⁵ Or weaving a lecture’s atomic elements (e.g., its component words) into a structured network
³⁶ that describes how those individual elements are related? Conceptual understanding could also
³⁷ involve building a mental model that transcends the meanings of those individual atomic elements
³⁸ by reflecting the deeper meaning underlying the gestalt whole [34, 38, 54].

³⁹ The difference between “understanding” and “memorizing,” as framed by researchers in ed-
⁴⁰ ucation, cognitive psychology, and cognitive neuroscience [e.g., 20, 26, 30, 38, 54] has profound
⁴¹ analogs in the fields of natural language processing and natural language understanding. For
⁴² example, considering the raw contents of a document (e.g., its constituent symbols, letters, and
⁴³ words) might provide some information about what the document is about, just as memorizing a
⁴⁴ passage might provide some ability to answer simple questions about it. However, text embedding

45 models [e.g., 7–9, 11, 15, 37, 45] also attempt to capture the deeper meaning *underlying* those atomic
46 elements. These models consider not only the co-occurrences of those elements within and across
47 documents, but also patterns in how those elements appear across different scales (e.g., sentences,
48 paragraphs, chapters, etc.), the temporal and grammatical properties of the elements, and other
49 high-level characteristics of how they are used [39, 40]. According to these models, the deep
50 conceptual meaning of a document may be captured by a feature vector in a high-dimensional
51 representation space, where nearby vectors reflect conceptually related documents. A model that
52 succeeds at capturing an analog of “understanding” is able to assign nearby feature vectors to
53 two conceptually related documents, *even when the words contained in those documents have very little*
54 *overlap*.

55 Given these insights, what form might the representation of the sum total of a person’s knowl-
56 edge take? First, we might require a means of systematically describing or representing the nearly
57 infinite set of possible things a person could know. Second, we might want to account for potential
58 associations between different concepts. For example, the concepts of “fish” and “water” might be
59 associated in the sense that fish live in water. Third, knowledge may have a critical dependency
60 structure, such that knowing about a particular concept might require first knowing about a set of
61 other concepts. For example, understanding the concept of a fish swimming in water first requires
62 understanding what fish and water *are*. Fourth, as we learn, our “current state of knowledge”
63 should change accordingly. Learning new concepts should both update our characterizations of
64 “what is known” and also unlock any now-satisfied dependencies of those newly learned concepts
65 so that they are “tagged” as available for future learning.

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

73 recorded course lectures.

74 Our primary research goal is to advance our understanding of what it means to acquire deep,
75 real-world conceptual knowledge. Traditional laboratory approaches to studying learning and
76 memory (e.g., list learning studies) often draw little distinction between memorization and under-
77 standing. Instead, these studies typically focus on whether information is effectively encoded or
78 retrieved, rather than whether the information is *understood*. Approaches to studying conceptual
79 learning, such as category learning experiments, can begin to investigate the distinction between
80 memorization and understanding, often by training participants to distinguish arbitrary or ran-
81 dom features in otherwise meaningless categorized stimuli. However the objective of real-world
82 training, or learning from life experiences more generally, is often to develop new knowledge that
83 may be applied in *useful* ways in the future. In this sense, the gap between modern learning theo-
84 ries and modern pedagogical approaches and classroom learning strategies is enormous: most of
85 our theories about *how* people learn are inspired by experimental paradigms and models that have
86 only peripheral relevance to the kinds of learning that students and teachers actually seek [26, 38].
87 To help bridge this gap, our study uses course materials from real online courses to inform, fit,
88 and test models of real-world conceptual learning. We also provide a demonstration of how our
89 models can be used to construct “maps” of what students know, and how their knowledge changes
90 with training. In addition to helping to visualize knowledge (and changes in knowledge), we hope
91 that such maps might lead to real-world tools for improving how we educate.

92 Results

93 At its core, our main modeling approach is based around a simple assumption that we sought to
94 test empirically: all else being equal, knowledge about a given concept is predictive of knowledge
95 about similar or related concepts. From a geometric perspective, this assumption implies that
96 knowledge is fundamentally “smooth.” In other words, as one moves through a space representing
97 an individual’s knowledge (where similar concepts occupy nearby coordinates), their “level of
98 knowledge” should change relatively gradually throughout that space. To begin to test this

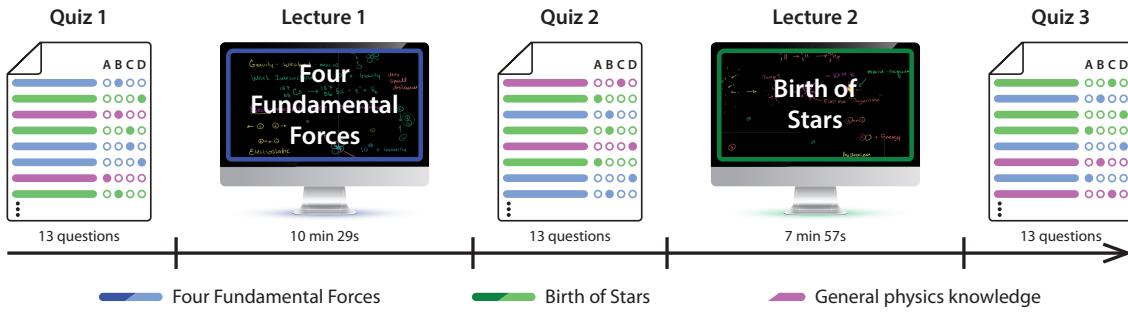


Figure 1: Experimental paradigm. Participants alternate between answering 13-question multiple-choice quizzes and watching two Khan Academy lectures. Each quiz contains a mix of 5 questions about lecture 1, 5 questions about lecture 2, and 3 questions about general physics knowledge. The specific questions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

smoothness assumption, we sought to track participants' knowledge and how it changed over time in response to training. Two overarching goals guide our approach. First, we want to gain detailed insights into what learners know, at different points in their training. For example, rather than simply reporting on the proportions of questions participants answer correctly (i.e., their overall performance), we seek estimates of their knowledge about a variety of specific concepts. Second, we want our approach to be potentially scalable to large numbers of concepts, courses, and students. This requires the conceptual content of interest to be discovered *automatically*, rather than relying on manually produced ratings or labels.

We asked participants in our study to complete brief multiple-choice quizzes before, between, and after watching two lecture videos from the Khan Academy [33] platform (Fig. 1). The first lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics: gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*, provided an overview of our current understanding of how stars form. We selected these particular lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training on our participants' abilities to learn from the lectures. To this end, we selected two introductory videos that were intended to be viewed at the start of students' training in their respective content areas. Second, we wanted both lectures to have some related content, so that we could test our approach's

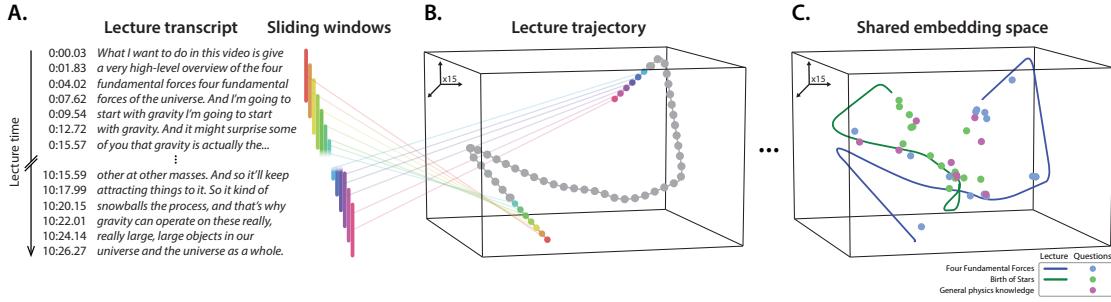


Figure 2: Modeling course content. **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding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

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).

132 To study in detail how participants' conceptual knowledge changed over the course of the
133 experiment, we first sought to model the conceptual content presented to them at each moment
134 throughout each of the two lectures. We adapted an approach we developed in prior work [27] to
135 identify the latent themes in the lectures using a topic model [8]. Briefly, topic models take as input
136 a collection of text documents and learn a set of "topics" (i.e., latent themes) from their contents.
137 Once fit, a topic model can be used to transform arbitrary (potentially new) documents into sets
138 of "topic proportions," describing the weighted blend of learned topics reflected in their texts. We
139 parsed automatically generated transcripts of the two lectures into overlapping sliding windows,
140 where each window contained the text of the lecture transcript from a particular time range. We
141 treated the set of text snippets (across all of these windows) as documents to fit our model (Fig. 2A;
142 see Constructing text embeddings of multiple lectures and questions). Transforming the text from
143 every sliding window with our model yielded a number-of-windows by number-of-topics (15)
144 topic-proportions matrix that described the unique mixture of broad themes from both lectures
145 reflected in each window's content. Each window's "topic vector" (i.e., column of the topic-
146 proportions matrix) is a coordinate in a 15-dimensional space whose axes are topics discovered by
147 the model. Within this space, each lecture's sequence of topic vectors (i.e., corresponding to its
148 transcript's overlapping text snippets across sliding windows) forms a *trajectory* that captures how
149 its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution
150 of one topic vector for each second of video (i.e., 1 Hz).

151 We hypothesized that a topic model trained on transcripts of the two lectures, should also
152 capture the conceptual knowledge probed by each quiz question. If indeed the topic model could
153 capture information about the deeper conceptual content of the lectures (i.e., beyond surface-level
154 details such as particular word choices) then we should be able to recover a correspondence between
155 each lecture and questions *about* each lecture. Importantly, such a correspondence could not solely
156 arise from superficial text matching between lecture transcripts and questions, since the lectures and
157 questions used different words. Simply comparing the average topic weights from each lecture and
158 question sets (averaging across time and questions, respectively) reveals a striking correspondence
159 (Fig. S1). Specifically, the average topic weights from lecture 1 are strongly correlated with the

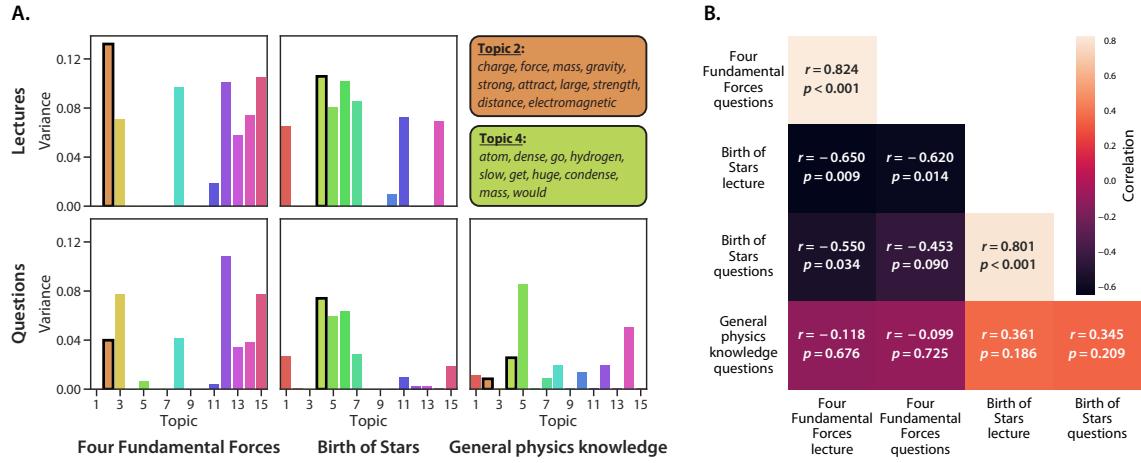


Figure 3: Lecture and question topic overlap. **A. Topic weight variability.** The bar plots display the variance of each topic's weight across lecture timepoints (top row) and questions (bottom row); colors denote topics. The top-weighted words from the most “expressive” (i.e., variable across observations) 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.

average topic weights from lecture 1 questions ($r(13) = XX, p = XX, 95\% \text{ confidence interval } (\text{CI}) = XX$), and the average topic weights from lecture 2 are strongly correlated with the average topic weights from lecture 2 questions ($r(13) = XX, p = XX, \text{CI} = XX$). At the same time, the average topics from two lectures are *negatively* correlated ($r(13) = XX, p = XX, \text{CI} = XX$). The full set of pairwise comparisons between topic vectors for the lectures and each question set is reported in Figure S1.

Another, more sensitive, way of summarizing the conceptual content of the lectures and questions is to look at *variability* in how topics are weighted over time and across different questions (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “information” [19] the lecture (or questions) reflect about that topic. For example, suppose a given topic is weighted on heavily throughout a lecture. That topic might be characteristic of some aspect or property of the lecture *overall* (conceptual or otherwise), but unless the topic’s weights changed in meaningful ways over time, the topic would be a poor indicator of any *specific* conceptual content in the lecture. We therefore also compared the variance in topic weights (across time or questions)

174 between the lectures and questions. The variability in topic expression (over time and across ques-
175 tions) was similar for the lecture 1 video and questions ($r(13) = 0.824$, $p < 0.001$, $CI = [0.696, 0.973]$)
176 and the lecture 2 video and questions ($r(13) = 0.801$, $p < 0.001$, 95% $CI = [0.539, 0.958]$). However,
177 as reported in Figure 3B, the variability in topic expressions across *different* videos and lecture-
178 specific questions (i.e., lecture 1 video versus lecture 2 questions; lecture 2 video versus lecture 1
179 questions) were *negatively* correlated, and neither video’s topic variability was reliably correlated
180 with the topic variability across general physics knowledge questions. Taken together, the analyses
181 reported in Figures 3 and S1 indicate that a topic model fit to the videos’ transcripts can also reveal
182 correspondances (at a coarse scale) between the lectures and (held-out) questions.

183 Although a single lecture may be organized around a single broad theme at a coarse scale, at a
184 finer scale each moment of a lecture typically covers a narrower range of content. We wondered
185 whether a text embedding model trained on the lectures’ transcripts might capture some of this
186 finer scale content. For example, if a particular question asks about the content from one small
187 part of a lecture, we wondered whether our text embedding model could be used to automatically
188 identify the “matching” moment(s) in the lecture. When we correlated each question’s topic vector
189 with the topic vectors for each second of the lectures, we found some evidence that each question is
190 temporally specific (Fig. 4). In particular, most questions’ topic vectors were maximally correlated
191 with a well-defined (and relatively narrow) range of timepoints from their corresponding lectures,
192 and the correlations fell off sharply outside of that range. We also examined the best-matching
193 intervals for each question qualitatively by comparing the text of the question to the text of the most-
194 correlated parts of the lectures. Despite that the questions were excluded from the text embedding
195 model’s training set, in general we found (through manual inspection) a close correspondence
196 between the conceptual content that each question covered and the content covered by the best-
197 matching moments of the lectures. Two representative examples are shown at the bottom of
198 Figure 4.

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

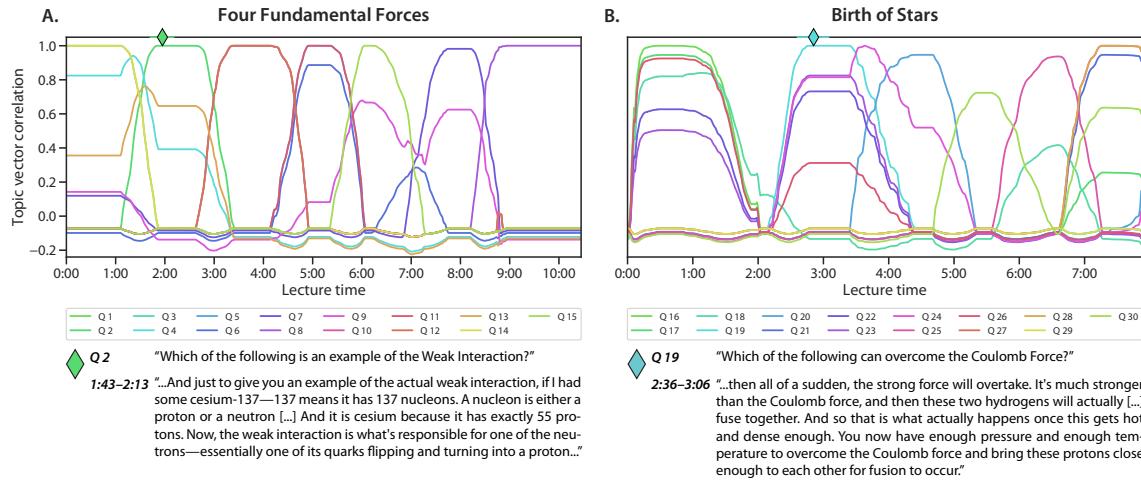


Figure 4: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel B.: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

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

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

Of course, even though the timecourses in Figure 5A and C provide detailed *estimates* about participants’ knowledge, those estimates are only *useful* to the extent that they accurately reflect what



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.

230 participants actually know. As one sanity check, we anticipated that the knowledge estimates
231 should show a content-specific “boost” in participants’ knowledge after watching each lecture.
232 In other words, if participants learn about each lecture’s content when they watch each lecture,
233 the knowledge estimates should reflect that. After watching the *Four Fundamental Forces* lecture,
234 participants should show more knowledge for the content of that lecture than they had before,
235 and that knowledge should persist for the remainder of the experiment. Specifically, knowledge
236 about that lecture’s content should be relatively low when estimated using Quiz 1 responses,
237 but should increase when estimated using Quiz 2 or 3 responses (Fig. 5B). Indeed, we found
238 that participants’ estimated knowledge about the content of the *Four Fundamental Forces* was
239 substantially higher on Quiz 2 versus Quiz 1 ($t(49) = 8.764, p < 0.001$) and on Quiz 3 versus Quiz
240 1 ($t(49) = 10.519, p < 0.001$). We found no reliable differences in estimated knowledge about
241 that lecture’s content on Quiz 2 versus 3 ($t(49) = 0.160, p = 0.874$). Similarly, we hypothesized
242 (and subsequently confirmed) that participants should show more estimated knowledge about the
243 content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 5D). Specifically, since
244 participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their
245 knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a “boost” on
246 Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge
247 about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ($t(49) = 1.013, p = 0.316$), but the
248 estimated knowledge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and
249 Quiz 3 versus 1 ($t(49) = 8.969, p < 0.001$).

250 If we are able to accurately estimate a participant’s knowledge about the content tested by a
251 given question, the estimated knowledge should have some predictive information about whether
252 the participant is likely to answer the question correctly or incorrectly. For each question in turn, for
253 each participant, we used Equation 1 to estimate (using all *other* questions from the same quiz, from
254 the same participant) the participant’s knowledge at the held-out question’s embedding coordinate.
255 For each quiz, we grouped these estimates into two distributions: one for the estimated knowledge
256 at the coordinates of each *correctly* answered question, and another for the estimated knowledge at
257 the coordinates of each *incorrectly* answered question (Fig. 6). We then used independent samples



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.

258 t -tests to compare the means of these distributions of estimated knowledge.

259 For the initial quizzes participants took (prior to watching either lecture), participants' estimated
 260 knowledge tended to be low overall, and relatively unstructured (Fig. 6, left panel). When we held
 261 out individual questions and estimated their knowledge at the held-out questions' embedding
 262 coordinates, we found no reliable differences in the estimates when the held-out question had
 263 been correctly versus incorrectly answered ($t(633) = 0.577, p = 0.564$). After watching the first
 264 video, estimated knowledge for held-out correctly answered questions (from the second quiz;
 265 Fig. 6, middle panel) exhibited a positive shift relative to held-out incorrectly answered questions
 266 ($t(633) = 3.961, p < 0.001$). After watching the second video, estimated knowledge (from the
 267 third quiz; Fig. 6, right panel) for *all* questions exhibited a positive shift. However, the increase
 268 in estimated knowledge for held-out correctly answered questions was larger than for held-out
 269 incorrectly answered questions ($t(628) = 2.045, p = 0.041$).

270 Knowledge estimates need not be limited to the content of the lectures. As illustrated in
 271 Figure 7, our general approach to estimating knowledge from a small number of quiz questions
 272 may be applied to *any* content, given its text embedding coordinate. To visualize how knowledge
 273 "spreads" through text embedding space to content beyond the lectures participants watched,

274 we first fit a new topic model to the lectures' sliding windows with $k = 100$ topics. We hoped
275 that increasing the number of topics from 15 to 100 might help us to generalize the knowledge
276 predictions. (Aside from increasing the number of topics from 15 to 100, all other procedures and
277 model parameters were carried over from the preceding analyses.) As in our other analyses, we
278 resampled each lecture's topic trajectory to 1 Hz and also projected each question into a shared
279 text embedding space.

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

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

301 Another way of visualizing these content-specific increases in knowledge (apparently driven



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.

302 by watching each lecture) is displayed in Figure 7B. Taking the point-by-point difference between
303 the knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map*
304 that describes the *change* in knowledge estimates from one quiz to the next. These learning maps
305 highlight that the estimated knowledge increases we observed across maps were specific to the
306 regions around the embeddings of each lecture in turn.

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

321 Discussion

322 Teaching, like effective writing and speaking, is fundamentally about empathy [1, 44, 59]. Great
323 teachers consider students' interests [12, 60], backgrounds [16, 47, 53], and working memory capac-
324 ities [2], and flexibly optimize their teaching strategies within those constraints [4, 23, 28]. In the
325 classroom, empathizing with students also means maintaining open lines of communication [65] by
326 fostering an environment in which all students feel comfortable speaking up if they have an excit-
327 ing new idea, or if they are having trouble understanding something [21, 61]. In-person instruction

328 also often entails dynamic student-teacher and student-student interactions. These in-person in-
329 teractions can provide the instructor with valuable information about students' understanding of
330 the course material, beyond what they can glean solely from exams or assignments [18, 25, 62].
331 In turn, this can allow the instructor to adapt their teaching approaches on-the-fly according to
332 students' questions and behaviors. But what does great teaching look like in asynchronous online
333 courses, when the instructor typically prepares course lectures and materials without knowing
334 who will ultimately be learning from them? Can the empathetic side of teaching be automated
335 and scaled?

336 The notion of empathy also related to "theory of mind" of other individuals [22, 29, 42].
337 Considering others' unique perspectives, prior experiences, knowledge, goals, etc., can help us
338 to more effectively interact and communicate [51, 55, 58]. The knowledge and learning maps
339 we estimate in our study (Fig. 7) hint at one potential form that an automated "empathetic"
340 teacher might take. We imagine automated content delivery systems that adapt lessons on the
341 fly according to continually updated estimates of what students know and how quickly they are
342 learning different conceptual content [e.g., building on ideas such as 3, 24, 36, 64, and others].

343 Over the past several years, the global pandemic has forced many educators to teach re-
344 motely [32, 46, 56, 63]. This change in world circumstances is happening alongside (and perhaps
345 accelerating) geometric growth in the availability of high quality online courses on platforms such
346 as Khan Academy [33], Coursera [66], EdX [35], and others [52]. Continued expansion of the global
347 internet backbone and improvements in computing hardware have also facilitated improvements
348 in video streaming, enabling videos to be easily downloaded and shared by large segments of the
349 world's population. This exciting time for online course instruction provides an opportunity to
350 re-evaluate how we, as a global community, educate ourselves and each other. For example, we
351 can ask: what makes an effective course or training program? Which aspects of teaching might be
352 optimized or automated? How can we provide How and why do learning needs and goals vary
353 across people? How might we lower barriers to achieving a high quality education?

354 Alongside these questions, there is a growing desire to extend existing theories beyond the
355 domain of lab testing rooms and into real classrooms [31]. In part, this has led to a recent

356 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better
357 reflect more ethologically valid phenomena that are more directly relevant to real-world situations
358 and behaviors [48]. In turn, this has brought new challenges in data analysis and interpretation. A
359 key step towards solving these challenges will be to build explicit models of real-world scenarios
360 and how people behave in them (e.g., models of how people learn conceptual content from real-
361 world courses, as in our current study). A second key step will be to understand which sorts
362 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 5,
363 17, 43, 49, 50] might help to inform these models. A third major step will be to develop and
364 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic
365 paradigms.

366 Ultimately, our work suggests a new line of questions regarding the future of education:
367 which aspects of teaching can be optimized and/or automated? The social benefits of face-to-face
368 instruction, such as social interactions, friendships, and emotional support, cannot (and perhaps
369 should not) be fully replaced by an automated computer-based system. Nor can modern computer
370 systems experience emotional empathy in the human sense of the word. On the other hand,
371 perhaps it is possible to separate out the social aspects of classroom instruction from the purely
372 learning-related aspects. Our study shows that text embedding models can uncover detailed
373 insights into students’ knowledge and how it changes over time during learning. We hope that
374 these advances might help pave the way for new ways of teaching or delivering educational content
375 that are tailored to individual students’ learning needs and goals.

376 Materials and methods

377 Participants

378 We enrolled a total of 50 Dartmouth undergraduate students in our study. Participants received
379 course credit for enrolling. We asked each participant to fill out a demographic survey that included
380 questions about their age, gender, native spoken language, ethnicity, race, hearing, color vision,

381 sleep, coffee consumption, level of alertness, and several aspects of their educational background
382 and prior coursework.

383 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
384 years). A total of 15 participants reported their gender as male and 35 participants reported their
385 gender as female. A total of 49 participants reported their native language as "English" and 1
386 reported having another native language. A total of 47 participants reported their ethnicity as
387 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants
388 reported their races as White (32 participants), Asian (14 participants), Black or African American
389 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
390 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

403 Participants reported their undergraduate major(s) as "social sciences" (28 participants), "nat-
404 ural sciences" (16 participants), "professional" (e.g., pre-med or pre-law; 8 participants), "mathe-
405 matics and engineering" (7 participants), "humanities" (4 participants), or "undecided" (3 partici-
406 pants). Note that some participants selected multiple categories for their undergraduate major. We
407 also asked participants about the courses they had taken. In total, 45 participants reported having
408 taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan

409 Academy courses. Of those who reported having watched at least one Khan Academy course,
410 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8
411 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We
412 also asked participants about the specific courses they had watched, categorized under different
413 subject areas. In the “Mathematics” area, participants reported having watched videos on AP
414 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-
415 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry
416 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential
417 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),
418 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other
419 videos not listed in our survey (5 participants). In the “Science and engineering” area, participants
420 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-
421 ipants); Physics, AP Physics I, or AP Physics II (15 participants); Biology, AP Biology; or High
422 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed
423 in our survey (19 participants). We also asked participants whether they had specifically seen the
424 videos used in our experiment. Of the 45 participants who reported having having taken at least
425 one Khan Academy course in the past, 44 participants reported that they had not watched the *Four*
426 *Fundamental Forces* video, and 1 participant reported that they were not sure whether they had
427 watched it. All participants reported that they had not watched the *Birth of Stars* video. When
428 we asked participants about non-Khan Academy online courses, they reported having watched
429 or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test
430 preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 partic-
431 ipants), Computing (2 participants), and other categories not listed in our survey (18 participants).
432 Finally, we asked participants about in-person courses they had taken in different subject areas.
433 They reported taking courses in Mathematics (39 participants), Science and engineering (38 par-
434 ticipants), Arts and humanities (35 participants), Test preparation (27 participants), Economics
435 and finance (26 participants), Computing (15 participants), College and careers (7 participants), or
436 other courses not listed in our survey (6 participants).

437 **Experiment**

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

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

456 **Analysis**

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

458 We adapted an approach we developed in prior work [27] to embed each moment of the two
459 lectures and each question in our pool in a common representational space. Briefly, our approach
460 uses a topic model (Latent Dirichlet Allocation; 8), trained on a set of documents, to discover a
461 set of k “topics” or “themes.” Formally, each topic is defined as a set of weights over each word
462 in the model’s vocabulary (i.e., the union of all unique words, across all documents, excluding

463 “stop words.”). Conceptually, each topic is intended to give larger weights to words that are
464 conceptually related or that tend to co-occur in the same documents. After fitting a topic model,
465 each document in the training set, or any *new* document that contains at least some of the words in
466 the model’s vocabulary, may be represented as a k -dimensional vector describing how much the
467 document (most probably) reflects each topic. (Unless, otherwise noted, we used $k = 15$ topics.)

468 As illustrated in Figure 2A, we start by building up a corpus of documents using overlapping
469 sliding windows that span each video’s transcript. Khan Academy provides professionally created,
470 manual transcriptions of all videos for closed captioning. However, such transcripts would not
471 be readily available in all contexts to which our framework could potentially be applied. Khan
472 Academy videos are hosted on the YouTube platform, which additionally provides automated
473 captions. We opted to use these automated transcripts (which, in prior work, we have found are
474 sufficiently near-human quality yield reliable data in behavioral studies; 67) when developing our
475 framework in order to make it more easily extensible and adaptable by others in the future.

476 We fetched these automated transcripts using the `youtube-transcript-api` Python package
477 (**Jeremy can you add a citation for this? <https://github.com/jdepoix/youtube-transcript-api>**). The
478 transcripts consisted of one timestamped line of text for every few seconds (mean: 2.34s; standard
479 deviation: 0.83s) of spoken content in the video (i.e., corresponding to each individual caption that
480 would appear on-screen if viewing the lecture via YouTube, and when those lines would appear).
481 We defined a sliding window length of (up to) $w = 30$ transcript lines, and assigned each window
482 a timestamp corresponding to the midpoint between its first and last lines’ timestamps. These
483 sliding windows ramped up and down in length at the very beginning and end of the transcript,
484 respectively. In other words, the first sliding window covered only the first line from the transcript;
485 the second sliding window covered the first two lines; and so on. This insured that each line of
486 the transcript appeared in the same number (w) of sliding windows. After performing various
487 standard text preprocessing (e.g., normalizing case, lemmatizing, removing punctuation and stop-
488 words), we treated the text from each sliding window as a single “document,” and we combined
489 these documents across the two videos’ windows to create a single training corpus for the topic
490 model. The top words from each of the 15 discovered topics may be found in Table S2.

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

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

506 **Estimating dynamic knowledge traces**

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

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

509 where

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

510 and where mincorr and maxcorr are the minimum and maximum correlations between any
 511 lecture timepoint and question, taken over all timepoints and questions across both lectures and
 512 all five question used to estimate the knowledge trace **Note: make sure this is correct in results**
 513 **section; not sure i caught all instances**. We also define $f(s, \Omega)$ as the s^{th} topic vector from the set

514 of topic vectors Ω . Here t indexes the set of lecture topic vectors, L , and i and j index the topic
515 vectors of questions used to estimate the knowledge trace, Q . Note that “correct” denotes the set
516 of indices of the questions the participant answered correctly on the given quiz.

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

525 **Creating knowledge and learning map visualizations**

526 An important feature of our approach is that, given a trained text embedding model and partic-
527 ipants’ quiz performance on each question, we can estimate their knowledge about *any* content
528 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-
529 tions or even appearing in the lectures. To visualize these estimates (Figs. 7, S2, S3, S4, S5, and S6),
530 we used Uniform Manifold Approximation and Projection (UMAP; 41) to construct a 2D projection
531 of the text embedding space. Sampling the original 100-dimensional space at high resolution to
532 obtain an adequate set of topic vectors spanning the embedding space would be computationally
533 intractable. However, sampling a 2D grid is much more feasible. We defined a rectangle enclosing
534 the 2D projections of the lectures’ and quizzes’ embeddings, and we sampled points from a regular
535 100×100 grid of coordinates that evenly tiled the enclosing rectangle. We sought to estimate
536 participants’ knowledge (and learning, i.e., changes in knowledge) at each of the resulting 10,000
537 coordinates.

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

541 by:

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

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

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

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

549 Author contributions

550 Conceptualization: PCF, ACH, and JRM. Methodology: PCF, ACH, and JRM. Software: PCF.
551 Validation: PCF. Formal analysis: PCF. Resources: PCF, ACH, and JRM. Data curation: PCF.
552 Writing (original draft): JRM. Writing (review and editing): PCF, ACH, and JRM. Visualization:
553 PCF and JRM. Supervision: JRM. Project administration: PCF. Funding acquisition: JRM.

554 Data and code availability

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

558 **Acknowledgements**

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

566 **References**

- 567 [1] Aldrup, K., Carstensen, B., and Klusmann, U. (2022). Is empathy the key to effective teaching?
568 A systematic review of its association with teacher-student interactions and student outcomes.
569 *Educational Psychology Review*, 34:1177001216.
- 570 [2] Alloway, T. P. (2012). Teachers' perceptions of classroom behaviour and working memory.
571 *Educational Research and Review*, 7(6):138–142.
- 572 [3] Anderson, J. R. and Skwarecki, E. (1986). The automated tutoring of introductory computer
573 programming. *Communications of the ACM*, 29(9):842–849.
- 574 [4] Anderton, R. S., Vitali, J., Blackmore, C., and Bakeberg, M. C. (2021). Flexible teaching and learn-
575 ing modalities in undergraduate science amid the COVID-19 pandemic. *Frontiers in Education*,
576 5:doi.org/10.3389/feduc.2020.609703.
- 577 [5] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and
578 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
579 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 580 [6] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-

- 581 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
582 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 583 [7] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*
584 *Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
585 Machinery.
- 586 [8] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
587 *Learning Research*, 3:993–1022.
- 588 [9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
589 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
590 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
591 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
592 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 593 [10] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
594 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 595 [11] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
596 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
597 sentence encoder. *arXiv*, 1803.11175.
- 598 [12] Clark, J. (2010). Powerpoint and pedagogy: maintaining student interest in university lectures.
599 *College Teaching*, 56(1):39–44.
- 600 [13] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
601 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 602 [14] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
603 Evidence for a new conceptualization of semantic representation in the left and right cerebral
604 hemispheres. *Cortex*, 40(3):467–478.

- 605 [15] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
606 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,
607 41(6):391–407.
- 608 [16] den Brok, P., van Tartwijk, J., Wubbels, T., and Veldman, I. (2010). The differential effect of
609 the teacher-student interpersonal relationship on student outcomes for students with different
610 ethnic backgrounds. *British Journal of Educational Psychology*, 80(2):199–221.
- 611 [17] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
612 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
613 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 614 [18] Englehart, J. M. (2009). Teacher-student interaction. In Saha, L. J. and Dworkin, A. G., editors,
615 *International Handbook of Research on Teachers and Teaching*. Springer International Handbooks of
616 Education.
- 617 [19] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical
618 Transactions of the Royal Society A*, 222(602):309–368.
- 619 [20] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
620 *School Science and Mathematics*, 100(6):310–318.
- 621 [21] Garran, A. M. and Rasmussen, B. M. (2014). Safety in the classroom: reconsidered. *Journal of
622 Teaching in Social Work*, 34(4):401–412.
- 623 [22] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of
624 Cognition and Development*, 13(1):19–37.
- 625 [23] Goode, S., Willis, R. A., Wolf, J. R., and Harris, A. L. (2007). Enhancing IS education with
626 flexible teaching and learning. *Journal of Information Systems Education*, 18(3):297–302.
- 627 [24] Halff, H. M. (1988). Curriculum and instruction in automated tutors. *Foundations of intelligent
628 tutoring systems*, pages 79–108.

- 629 [25] Hall, J. K. and Walsh, M. (2002). Teacher-student interaction and language learning. *Annual*
630 *Review of Applied Linguistics*, 22:186–203.
- 631 [26] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
632 learning, pages 212–221. Sage Publications.
- 633 [27] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal be-
634 havioral and neural signatures of transforming naturalistic experiences into episodic memories.
635 *Nature Human Behavior*, 5:905–919.
- 636 [28] Johnston, S. (2002). Introducing and supporting change towards more flexible teaching ap-
637 proaches. In *The convergence of distance and conventional education*. Routledge.
- 638 [29] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
639 empathy and theory of mind. In *Social Behavior from Rodents to Humans*, pages 193–206. Springer.
- 640 [30] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
641 Columbia University Press.
- 642 [31] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
643 326(7382):213–216.
- 644 [32] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
645 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International*
646 *Journal of Environmental Research and Public Health*, 18(5):2672.
- 647 [33] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 648 [34] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 649 [35] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
650 *The Chronicle of Higher Education*, 21:1–5.
- 651 [36] Kumar, A. N. (2005). Generation of problems, answers, grade, and feedback—case study of a
652 fully automated tutor. *Journal on Educational Resources in Computing*, 5(3):1–25.

- 653 [37] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic
654 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
655 104:211–240.
- 656 [38] MacLellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of
657 Educational Studies*, 53(2):129–147.
- 658 [39] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
659 *Handbook of Human Memory*. Oxford University Press.
- 660 [40] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
661 function? *Psychological Review*, 128(4):711–725.
- 662 [41] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
663 projection for dimension reduction. *arXiv*, 1802(03426).
- 664 [42] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
665 mind. In *The Wiley-Blackwell handbook of childhood cognitive development*. Wiley-Blackwell.
- 666 [43] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
667 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
668 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 669 [44] Meyers, S., Rowell, K., Wells, M., and Smith, B. C. (2019). Teacher empathy: a model of
670 empathy for teaching for student success. *College Teaching*, 67(3):160–168.
- 671 [45] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
672 tations in vector space. *arXiv*, 1301.3781.
- 673 [46] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
674 from a national survey of language educators. *System*, 97:102431.
- 675 [47] Muijs, D. and Reynolds, D. (2003). Student background and teacher effects on achievement and
676 attainment in mathematics: a longitudinal study. *Educational Research and Evaluation*, 9(3):289–
677 314.

- 678 [48] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
679 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 680 [49] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
681 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*
682 *Neuroscience*, 17(4):367–376.
- 683 [50] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
684 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
685 7:43916.
- 686 [51] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*
687 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 688 [52] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
689 higher education: unmasking power and raising questions about the movement’s democratic
690 potential. *Educational Theory*, 63(1):87–110.
- 691 [53] Rosenshine, B. (1976). Recent research on teaching behaviors and student achievement. *Journal*
692 *of Teacher Education*, 27(1):61–64.
- 693 [54] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
694 Student conceptions and conceptual learning in science. Routledge.
- 695 [55] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
696 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*
697 *tion in Nursing*, 22:32–42.
- 698 [56] Shim, T. E. and Lee, S. Y. (2020). College students’ experience of emergency remote teaching
699 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 700 [57] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
701 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*
702 *Mathematics Education*, 35(5):305–329.

- 703 [58] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal*
704 *Medicine*, 21:524–530.
- 705 [59] Stojiljković, S., Djigić, G., and Zlatković, B. (2012). Empathy and teachers' roles. *Procedia –*
706 *Social and Behavioral Sciences*, 69:960–966.
- 707 [60] Swarat, S., Ortony, A., and Revelle, W. (2012). Activity matters: understanding student interest
708 in school science. *Journal of Research in Science Teaching*, 49(4):515–537.
- 709 [61] Turner, S. and Braine, M. (2015). Unravelling the 'safe' concept in teaching: what can we learn
710 from teachers' understanding? *Pastoral Care in Education*, 33(1):47–62.
- 711 [62] van de Pol, J., Volman, M., and Beishuizen, J. (2010). Scaffolding in teacher-student interaction:
712 a decade of research. *Educational Psychology Review*, 22:271–296.
- 713 [63] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
714 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 715 [64] Wolz, U., McKeown, K., and Kaiser, G. E. (1988). Automated tutoring in interactive environ-
716 ments: a task centered approach. Technical report, Columbia University.
- 717 [65] Wulff, S. S. and Wulff, D. H. (2004). "of course i'm communicating; I lecture every day":
718 enhancing teaching and learning in introductory statistics. *Communication Education*, 53(1):92–
719 103.
- 720 [66] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
721 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 722 [67] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is
723 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*
724 *Research Methods*, 50:2597–2605.