

¹ Text embedding models reveal 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, *}

¹Dartmouth College

²Akili Interactive

*Corresponding author: jeremy.r.manning@dartmouth.edu

⁵ **Abstract**

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

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

¹⁸ **Introduction**

¹⁹ Suppose that a teacher had access to a complete “map” of everything their student knew. Defining
²⁰ what such a map might even look like, let alone how it might be constructed or filled in, is itself
²¹ a non-trivial problem. But if a teacher *were* to gain access to such a map, how might that change
²² their ability to teach the 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 any new material.
²⁷ Observing how the student’s knowledge was changing over time, in response to their training,
²⁸ could also help to guide the teacher.

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 the researchers
⁴⁰ in education, cognitive psychology, and cognitive neuroscience [19, 25, 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
⁴⁴ a passage might be used to answer simple questions about the passage [e.g., whether it might

45 contain words related to furniture versus physics; 7, 8, 37]. However, modern natural language
46 processing models [e.g., 9, 11, 45] also attempt to capture the deeper meaning *underlying* those
47 atomic elements. These models consider not only the co-occurrences of those elements within
48 and across documents, but also patterns in how those elements appear across different scales (e.g.,
49 sentences, paragraphs, chapters, etc.), the temporal and grammatical properties of the elements,
50 and other high-level characteristics of how they are used [39, 40]. According to these models, the
51 deep conceptual meaning of a document may be captured by a feature vector in a high-dimensional
52 representation space, where nearby vectors reflect conceptually related documents. A model that
53 succeeds at capturing an analog of “understanding” is able to assign nearby feature vectors to
54 two conceptually related documents, *even when the words contained in those documents have very little*
55 *overlap*.

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

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

73 an experiment that has participants watch and answer multiple choice questions about a series of
74 recorded course lectures.

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

94 Results

95 At its core, our main modeling approach is based around a simple assumption that we sought to test
96 empirically: all else being equal, knowledge about a given concept is predictive of knowledge about
97 similar or related concepts. From a geometric perspective, this assumption implies that knowledge
98 is fundamentally “smooth.” In other words, as one moves through a space representing someone’s

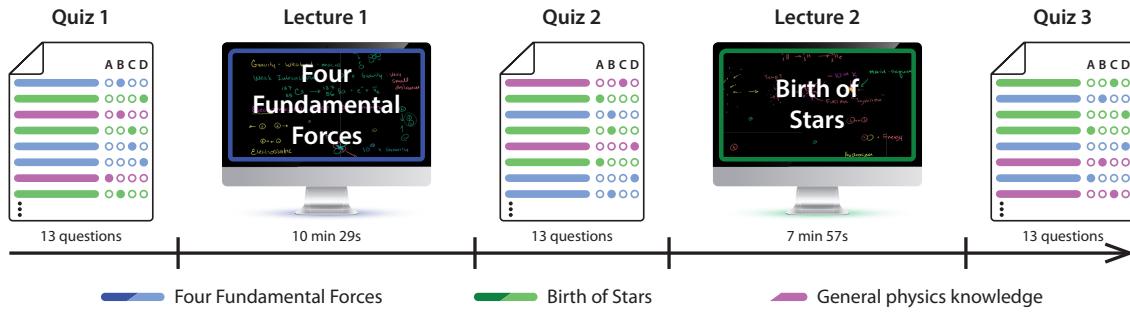


Figure 1: Experimental paradigm. Participants alternate between answering 13-question multiple choice quizzes and watching two Khan academy videos. Each quiz contains a mix of 5 questions about lecture 1, 5 questions about lecture 2, and 3 general physics knowledge questions. The specific questions reflected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

knowledge (where similar concepts occupy nearby coordinates), their “level of knowledge” should change relatively gradually throughout that space. To begin to test this smoothness assumption, we sought to track our participants’ knowledge and how it changed over time in response to training.

We asked our participants to answer questions from several multiple choice quizzes and watch two lecture videos from the *Khan Academy* platform (Fig. 1). One lecture video, entitled *Four Fundamental Forces*, was about the four fundamental forces in physics: gravity, strong and weak interactions, and electromagnetism. The second lecture video, entitled *Birth of Stars*, provides an overview of our current understanding of how stars form. We selected both lessons to be (a) accessible to a broad audience, e.g., by minimizing prerequisite knowledge, (b) largely independent of each other, e.g., so that the two videos focused on different material and did not depend on each other, and (c) related to each other, e.g., so that both videos contained at least *some* similar or overlapping content. The two videos we selected are introductory (i.e., minimizing specific prerequisite knowledge), and are about different primary concepts, but they both also touch on “physics” and “astronomy” themes. We also wrote a set of multiple choice quiz questions that would enable us to test participants’ knowledge about each individual video and about related content not specifically presented in either video (Tab. S1). Participants answered questions randomly drawn from each content area (lecture 1, lecture 2, and general physics knowledge) across

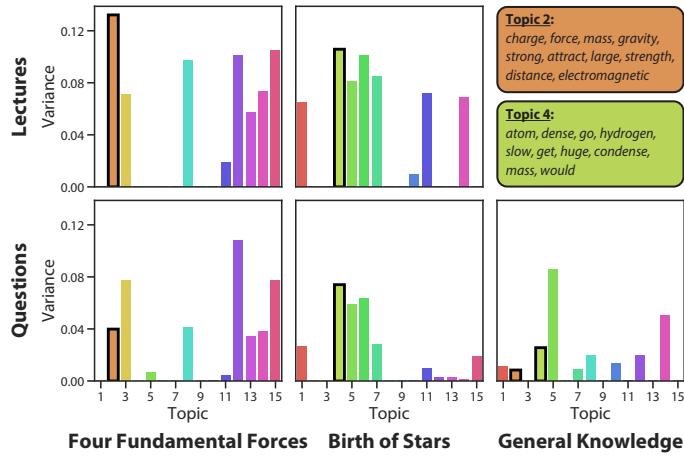


Figure 2: Lecture and question topic overlap. The bar plots display the variability in topic weights across lecture timepoints (top panels) and questions (bottom panels); colors denote topics. The top-weighted words from the most “expressive” (i.e., variable across observations) topics from each lecture 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.

117 each of three quizzes. Quiz 1 was intended to assessed participants’ knowledge before training;
 118 quiz 2 assessed knowledge after watching the *Four Fundamental Forces* video (i.e., lecture 1); and
 119 quiz 3 assessed knowledge after watching the *Birth of Stars* video (i.e., lecture 2).

120 We trained a text embedding model using sliding windows of text from the two videos’ trans-
 121 scripts (see *Constructing text embeddings of multiple videos and questions*). We also used the same
 122 model (i.e., trained on the videos’ transcripts) to embed the text of each question in our pool.
 123 This yielded, for each second of each video, and for each question, a single 15-dimensional topic
 124 vector—i.e., a coordinate in a text embedding space (Fig. 7). Intuitively, each dimension of the em-
 125 bedding space corresponds to a “theme” or “topic” reflected in some part(s) of the videos (Tab. S2),
 126 and the coordinates in embedding space denote the blend of themes reflected by a particular
 127 excerpt of text (e.g., from part of a video’s transcript, from a question, etc.).

128 Given that we trained the text embedding model using the video transcripts, we wondered
 129 whether the questions that were (ostensibly, by design) “about” the content of each lecture would
 130 “match up” correctly with the lectures. In other words, we hoped that the text embeddings would
 131 capture something about the deeper conceptual content of the lectures, beyond surface details such

as exact wording choices. If so, when we embed *new* text outside of the model’s training set, we should see a correspondance between the embeddings of the training data (i.e., snippets of text from the lectures’ transcripts) and other text that reflects related concepts (e.g., questions *about* each lecture). Further, although the content from any given moment from a lecture might stray from the average content (across all timepoints), we hoped that *variability* in each topic’s expression over timepoints within a lecture would match up with the variability in topic expressions for questions about that lecture. Intuitively, the variability in the expression of a given topic relates to how much “information” [18] the lecture (or questions) reflect about that topic. When we compared the variability in topic weights across each lecture’s timepoints with the variability in topic weights across each question set, we found a strong correspondence (Fig. 2). The most variable topics from the *Four Fundamental Forces* lecture, and questions about that lecture, are 2, 3, 8, 12, 13, 14, and 15. The most variable topics from the *Birth of Stars* lecture, and questions about that lecture, are 1, 4, 5, 6, and 7. This strong overlap between the lectures and questions specifically about each lecture indicates that the topic model captures some of the underlying conceptual content.

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 of a lecture, we wondered whether our text embedding model 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 for each second of the lectures, we found some evidence that each question is temporally specific (Fig. 3). 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 examined the best-matching intervals for each question qualitatively 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 covered and the content covered by the best-

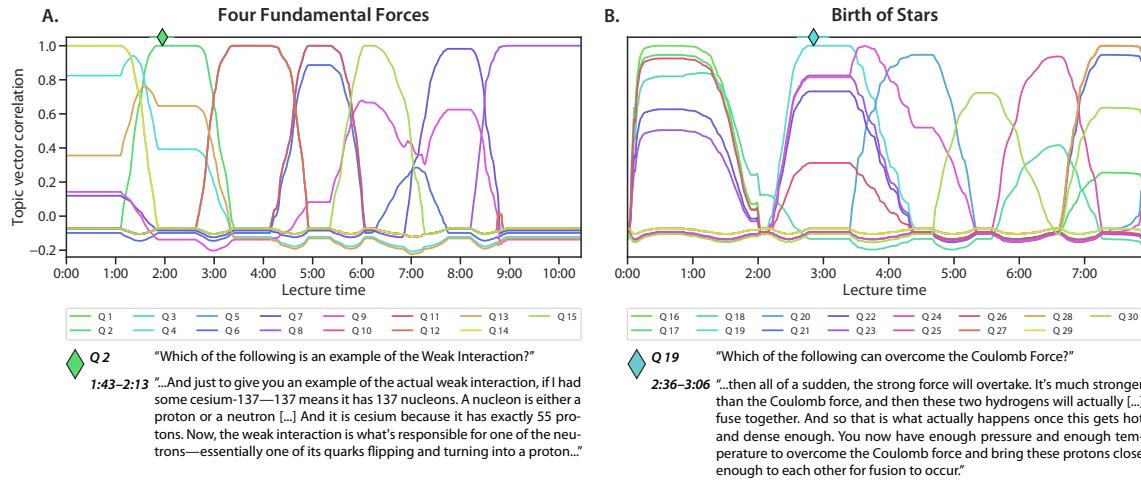


Figure 3: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel B.: 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 questions, in the indicated lectures. The associated questions’ text, and snippets of the lectures’ transcripts in the best-matching sliding windows, are displayed at the bottom of the figure.

160 matching moments of the lectures. Two representative examples are shown at the bottom of
161 Figure 3.

162 The ability to quantify how much each question is “asking about” the content from each moment
163 of the lectures could enable high-resolution insights into participants’ knowledge. Traditional
164 approaches to estimating how much a student “knows” about the content of a given lecture entail
165 computing the proportion of correctly answered questions. But if two students receive identical
166 scores on an exam, might our modeling framework help us to gain more nuanced insights into the
167 *specific* content that each student has mastered (or failed to master)? For example, a student who
168 misses three questions that were all about the same concept (e.g., concept *A*) will have gotten the
169 same *proportion* of questions correct as another student who missed three questions about three
170 *different* concepts (e.g., *A*, *B*, and *C*). But if we wanted to fill in the “gaps” in the two students’
171 understandings, we might do well to focus on concept *A* for the first student, but to also add in
172 materials pertaining to concepts *B* and *C* for the second student.

173 We developed a simple formula (Eqn. 1) for using a participant’s responses to a small set
174 of multiple choice questions to estimate how much the participant “knows” about the concept
175 reflected by any arbitrary coordinate, x , in text embedding space (e.g., the content reflected by
176 any moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially,
177 the estimated knowledge at the coordinate is given by the weighted average proportion of quiz
178 questions the participant answered correctly, where the weights reflect how much each question
179 is “about” the content at x . When we apply this approach to estimate the participant’s knowledge
180 about the content presented in each moment of each lecture, we can obtain a detailed timecourse
181 describing how much “knowledge” the participant has about any part of the lecture. As shown
182 in Figure 4, we can also apply this approach separately for the questions from each quiz the
183 participants took throughout the experiment. From just 13 questions per quiz, we obtain a high-
184 resolution snapshot (at the time each quiz was taken) of what the participants knew about any
185 moment’s content, from either of the two lectures they watched (comprising a total of 1106 samples
186 across the two lectures).

187 Of course, even though the timecourses in Figure 4A and C provide detailed *estimates* about

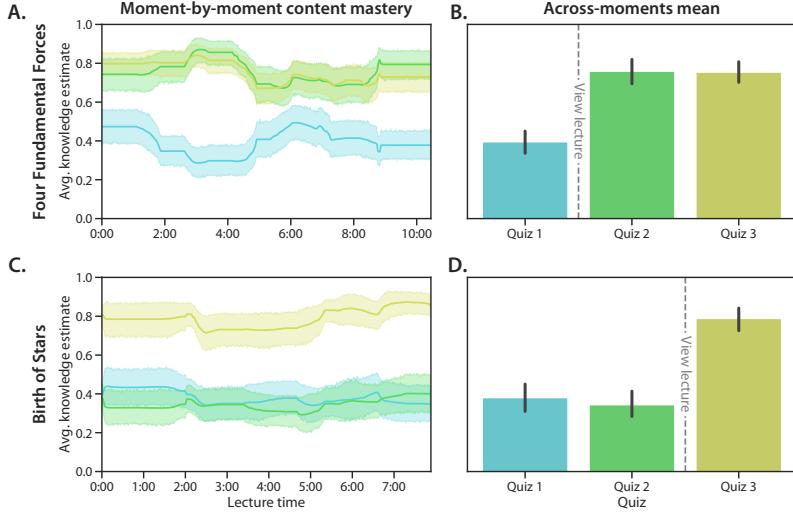


Figure 4: Estimating moment-by-moment knowledge acquisition. **A. Moment-by-moment knowledge about the *Four Fundamental Forces*.** Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (see *Estimating dynamic knowledge 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.

188 participants' knowledge, those estimates are only *useful* to the extent that they accurately reflect what
189 participants actually know. As one sanity check, we anticipated that the knowledge estimates
190 should show a content-specific "boost" in participants' knowledge after watching each lecture.
191 In other words, if participants learn about each lecture's content when they watch each lecture,
192 the knowledge estimates should reflect that. After watching the *Four Fundamental Forces* lecture,
193 participants should show more knowledge for the content of that lecture than they had before,
194 and that knowledge should persist for the remainder of the experiment. Specifically, knowledge
195 about that lecture's content should be relatively low when estimated using Quiz 1 responses,
196 but should increase when estimated using Quiz 2 or 3 responses (Fig. 4B). Indeed, we found
197 that participants' estimated knowledge about the content of the *Four Fundamental Forces* was
198 substantially higher on Quiz 2 versus Quiz 1 ($t(49) = 8.764, p < 0.001$) and on Quiz 3 versus Quiz
199 1 ($t(49) = 10.519, p < 0.001$). We found no reliable differences in estimated knowledge about
200 that lecture's content on Quiz 2 versus 3 ($t(49) = 0.160, p = 0.874$). Similarly, we hypothesized
201 (and subsequently confirmed) that participants should show more estimated knowledge about the
202 content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 4D). Specifically, since
203 participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their
204 knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a "boost" on
205 Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge
206 about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ($t(49) = 1.013, p = 0.316$), but the
207 estimated knowledge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and
208 Quiz 3 versus 1 ($t(49) = 8.969, p < 0.001$).

209 If we are able to accurately estimate a participant's knowledge about the content tested by a
210 given question, the estimated knowledge should have some predictive information about whether
211 the participant is likely to answer the question correctly or incorrectly. For each question in turn, for
212 each participant, we used Equation 1 to estimate (using all *other* questions from the same quiz, from
213 the same participant) the participant's knowledge at the held-out question's embedding coordinate.
214 For each quiz, we grouped these estimates into two distributions: one for the estimated knowledge
215 at the coordinates of each *correctly* answered question, and another for the estimated knowledge at

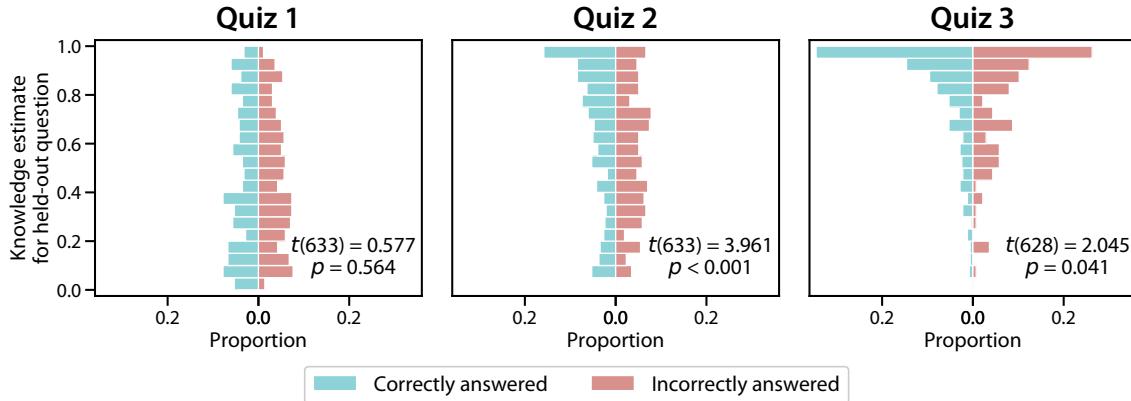


Figure 5: Estimating knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held out correctly (blue) or incorrectly (red) answered question. The t -tests reported in each panel are between the distributions of estimated knowledge at the coordinates of correctly versus incorrectly answered held-out questions.

216 the coordinates of each *incorrectly* answered question (Fig. 5). We then used independent samples
 217 t -tests to compare the means of these distributions of estimated knowledge.

218 For the initial quizzes participants took (prior to watching either lecture), participants' estimated
 219 knowledge tended to be low overall, and relatively unstructured (Fig. 5, left panel). When we held
 220 out individual questions and estimated their knowledge at the held-out questions' embedding
 221 coordinates, we found no reliable differences in the estimates when the held-out question had
 222 been correctly versus incorrectly answered ($t(633) = 0.577, p = 0.564$). After watching the first
 223 video, estimated knowledge for held-out correctly answered questions (from the second quiz;
 224 Fig. 5, middle panel) exhibited a positive shift relative to held-out incorrectly answered questions
 225 ($t(633) = 3.961, p < 0.001$). After watching the second video, estimated knowledge (from the
 226 third quiz; Fig. 5, right panel) for *all* questions exhibited a positive shift. However, the increase
 227 in estimated knowledge for held-out correctly answered questions was larger than for held-out
 228 incorrectly answered questions ($t(628) = 2.045, p = 0.041$).

229 Knowledge estimates need not be limited to the content of the lectures. As illustrated in
 230 Figure 6, our general approach to estimating knowledge from a small number of quiz questions
 231 may be applied to *any* content, given its text embedding coordinate. To visualize how knowledge

232 “spreads” through text embedding space to content beyond the lectures participants watched,
233 we first fit a new topic model to the lectures’ sliding windows with $k = 100$ topics. We hoped
234 that increasing the number of topics from 15 to 100 might help us to generalize the knowledge
235 predictions. (Aside from increasing the number of topics from 15 to 100, all other procedures and
236 model parameters were carried over from the preceding analyses.) As in our other analyses, we
237 resampled each lecture’s topic trajectory to 1 Hz and also projected each question into a shared
238 text embedding space.

239 We projected the resulting 100-dimensional topic vectors (for each second of video and for each
240 question) into a shared 2-dimensional space (see *Creating knowledge and learning map visualizations*).
241 Next, we sampled points evenly from a 100×100 grid of coordinates that evenly tiled a rectangle
242 enclosing the 2D projections of the videos and questions. We used Equation 4 to estimate partici-
243 pants’ knowledge at each of these 10K sampled locations, and we averaged these estimates across
244 participants to obtain an estimated average *knowledge map* (Fig. 6). Intuitively, the knowledge map
245 constructed from a given quiz’s responses provides a visualization of how “much” participants
246 know about any content expressible by the fitted text embedding model.

247 Several features of the resulting knowledge maps are worth noting. The average knowledge
248 map estimated from Quiz 1 responses (Fig. 6, leftmost map) shows that participants tended to
249 have relatively little knowledge about any parts of the text embedding space (i.e., the shading
250 is relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a
251 marked increase in knowledge on the left side of the map (around roughly the same range of
252 coordinates covered by the *Four Fundamental Forces* lecture, indicated by the dotted blue line).
253 In other words, participants’ estimated increase in knowledge is localized to conceptual content
254 that is nearby (i.e., related to) the content from the lecture they watched prior to taking Quiz
255 2. This localization is non-trivial: the knowledge estimates are informed only by the embedded
256 coordinates of the *quiz questions*, not by the embeddings of either lecture (Eqn. 4). Finally, the
257 knowledge map estimated from Quiz 3 responses shows a second increase in knowledge, localized
258 to the region surrounding the embedding of the *Birth of Stars* lecture participants watched prior to
259 taking Quiz 3.

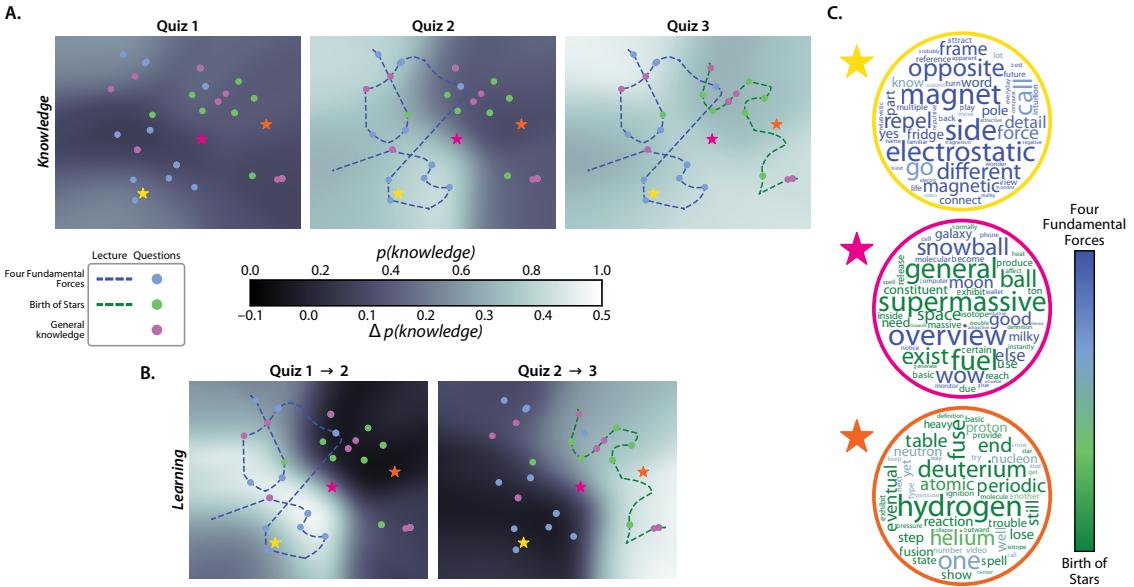


Figure 6: Mapping out the geometry of knowledge and learning. **A.** Average “knowledge maps” estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*). 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 S1, S2, and S3. **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 S4 and S5. **C.** Word clouds for sampled points in topic space. Each word cloud displays the relative weights of each word 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.

260 Another way of visualizing these content-specific increases in knowledge (apparently driven
261 by watching each lecture) is displayed in Figure 6B. Taking the point-by-point difference between
262 the knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map*
263 that describes the *change* in knowledge estimates from one quiz to the next. These learning maps
264 highlight that the estimated knowledge increases we observed across maps were specific to the
265 regions around the embeddings of each lecture in turn.

266 Because the 2D projection we used to construct the knowledge and learning maps is (partially)
267 invertible, we may gain additional insights into the estimates by reconstructing the original high-
268 dimensional topic vectors for any point(s) in the maps we are interested in. For example, this
269 could serve as a useful tool for an instructor looking to better understand which content areas
270 a student (or a group of students) knows well (or poorly). As a demonstration, we show the
271 top-weighted words from the blends of topics reconstructed from three example locations on the
272 maps (Fig. 6C): one point near the *Four Fundamental Forces* embedding (yellow); a second point
273 near the *Birth of Stars* embedding (orange), and a third point somewhere in between the two
274 lectures' embeddings (pink). As shown in the word clouds in the Panel, the top-weighted words
275 at the example coordinate near the *Four Fundamental Forces* embedding also tended to be weighted
276 heavily by the topics expressed in that lecture. Similarly, the top-weighted words at the example
277 coordinate near the *Birth of Stars* embedding tended to be weighted most heavily by the topics
278 expressed in *that* lecture. And the top-weighted words at the example coordinate between the
279 two lectures' embeddings show a roughly even mix of words most strongly associated with each
280 lecture.

281 Discussion

282 Teaching, like effective writing and speaking, is fundamentally about empathy [1, 44, 59]. Great
283 teachers consider students' interests [12, 60], backgrounds [15, 47, 53], and working memory capac-
284 ities [2], and flexibly optimize their teaching strategies within those constraints [4, 22, 28]. In the
285 classroom, empathizing with students also means maintaining open lines of communication [65] by

286 fostering an environment in which all students feel comfortable speaking up if they have an exciting
287 new idea, or if they are having trouble understanding something [20, 61]. In-person instruction
288 also often entails dynamic student-teacher and student-student interactions. These in-person in-
289 teractions can provide the instructor with valuable information about students' understanding of
290 the course material, beyond what they can glean solely from exams or assignments [17, 24, 62].
291 In turn, this can allow the instructor to adapt their teaching approaches on-the-fly according to
292 students' questions and behaviors. But what does great teaching look like in asynchronous online
293 courses, when the instructor typically prepares course lectures and materials without knowing
294 who will ultimately be learning from them? Can the empathetic side of teaching be automated
295 and scaled?

296 The notion of empathy also related to "theory of mind" of other individuals [21, 29, 42].
297 Considering others' unique perspectives, prior experiences, knowledge, goals, etc., can help us
298 to more effectively interact and communicate [51, 55, 58]. The knowledge and learning maps
299 we estimate in our study (Fig. 6) hint at one potential form that an automated "empathetic"
300 teacher might take. We imagine automated content delivery systems that adapt lessons on the
301 fly according to continually updated estimates of what students know and how quickly they are
302 learning different conceptual content [e.g., building on ideas such as 3, 23, 36, 64, and others].

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

314 Alongside these questions, there is a growing desire to extend existing theories beyond the
315 domain of lab testing rooms and into real classrooms [31]. In part, this has led to a recent
316 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better
317 reflect more ethologically valid phenomena that are more directly relevant to real-world situations
318 and behaviors [48]. In turn, this has brought new challenges in data analysis and interpretation. A
319 key step towards solving these challenges will be to build explicit models of real-world scenarios
320 and how people behave in them (e.g., models of how people learn conceptual content from real-
321 world courses, as in our current study). A second key step will be to understand which sorts
322 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 5,
323 16, 43, 49, 50] might help to inform these models. A third major step will be to develop and
324 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic
325 paradigms.

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

336 **Materials and methods**

337 **Participants**

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

343 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
344 years). A total of 15 participants reported their gender as male and 35 participants reported their
345 gender as female. A total of 49 participants reported their native language as "English" and 1
346 reported having another native language. A total of 47 participants reported their ethnicity as
347 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants
348 reported their races as White (32 participants), Asian (14 participants), Black or African American
349 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
350 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

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

363 Participants reported their undergraduate major(s) as Social Sciences (28 participants), Natural
364 sciences (16), Professional (e.g., pre-med or pre-law; 8 participants), Mathematics and engineering
365 (7 participants), Humanities (4 participants), or Undecided (3 participants). Note that some par-
366 ticipants selected multiple categories for their undergraduate major. We also asked participants
367 about the courses they had taken. In total, 46 participants reported having taken at least one Khan
368 academy course in the past or being familiar with the Khan academy, and 4 reported not having
369 taken any Khan academy courses. Of the participants who reported having watched at least one
370 Khan academy course, 1 participant declined to report the number of courses they had watched;
371 7 participants reported having watched 1–2 courses; 11 reported having watched 3–5 courses; 8
372 reported having watched 5–10 courses; and 19 reported having watched 10 or more courses. We
373 also asked participants about the specific courses they had watched, categorized under different
374 subject areas. In the “Mathematics” area participants reported having watched videos on AP
375 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-
376 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry
377 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential
378 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),
379 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other
380 videos not listed in our survey (6 participants). In the “Science and engineering” area participants
381 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-
382 ipants); Physics, AP Physics I, or AP Physics II (15 participants); Biology, AP Biology; or High
383 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed in
384 our survey (20 participants). We also asked participants if they had specifically seen the videos
385 used in our experiment. When we asked about the *Four Fundamental Forces* video, 45 participants
386 reported not having watched it before, 1 participant reported that they were not sure if they had
387 watched it before, and 4 participants declined to respond. When we asked about the *Birth of*
388 *Stars* video, 46 participants reported not having watched it before and 4 participants declined to
389 respond. When we asked participants about non-Khan academy online courses, they reported
390 having watched or taken courses on Mathematics (15 participants), Science and engineering (11

391 participants), Test preparation (9 participants), Economics and finance (3 participants), Arts and
392 humanities (2 participants), Computing (2 participants), and other categories not listed in our
393 survey (18 participants). Finally, we asked participants about in-person courses they had taken in
394 different subject areas. They reported taking courses in Mathematics (39 participants), Science and
395 engineering (38 participants), Arts and humanities (35 participants), Test preparation (27 participants),
396 Economics and finance (26 participants), Computing (15 participants), College and careers
397 (7 participants), or other courses not listed in our survey (6 participants).

398 **Experiment**

399 We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces*
400 (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force;
401 duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed;
402 duration: 7 minutes and 57 seconds). We hand-wrote 39 multiple choice questions: 15 about the
403 conceptual content of *Four Fundamental Forces*, another 15 about the conceptual content of *Birth*
404 of *Stars*, and 9 other questions that tested for general conceptual knowledge about basic physics
405 (covering material that was not presented in either video). The full set of questions may be found
406 in Table S1.

407 Participants began the main experiment by answering a battery of 13 randomly selected ques-
408 tions (chosen from the full set of 39). Then they watched the *The Four Fundamental Forces* video.
409 Next, they answered a second set of 13 questions (chosen at random from the remaining 26 ques-
410 tions). Fourth, participants watch the *Birth of Stars* video, and finally they answered the remaining
411 13 questions. Our experimental procedure is diagramed in Figure 1. We used the experiment to
412 develop and test our computational framework for estimating knowledge and learning.

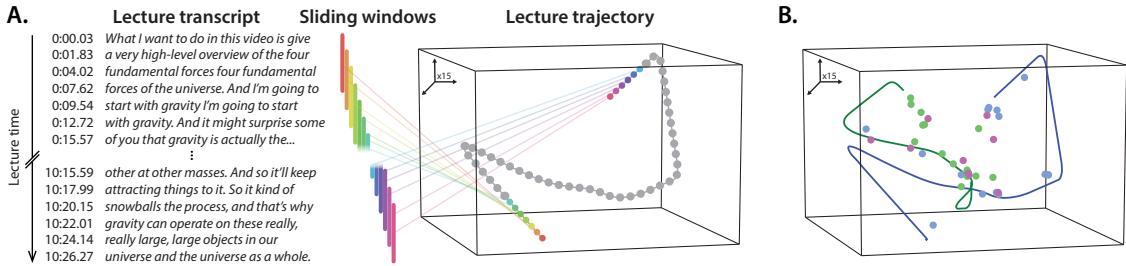


Figure 7: Constructing video content trajectories. **A. Building a document pool from sliding windows of text.** We decompose each video’s transcript into a series of overlapping sliding windows. The set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding model. After training a text embedding model using the two videos’ sliding windows, along with the text from each question in our pool (Tab. S1), we construct “trajectories” through text embedding space by joining the embedding coordinates of successive sliding windows from each video. **B. Embedding multiple videos and questions.** Applying the same text embedding approach to each video, along with the text of each question, results in one trajectory per video and one embedding coordinate (dot) per question (blue: *Four Fundamental Forces*; green: *Birth of Stars*; pink: general physics knowledge). Here we have projected the 15-dimensional embeddings into a 3D space using Uniform Manifold Approximation and Projection [UMAP; 41].

413 Analysis

414 Constructing text embeddings of multiple videos and questions

415 We extended an approach developed by [27] to construct text embeddings for each moment of
 416 each lecture, and of each question in our pool. Briefly, our approach uses a topic model [8], trained
 417 on a set of documents, to discover a set of k “topics” or “themes.” Formally, each topic is defined
 418 as a set of weights over each word in the model’s vocabulary (i.e., the union of all unique words,
 419 across all documents, excluding “stop words.”). Conceptually, each topic is intended to give larger
 420 weights to words that are conceptually related or that tend to co-occur in the same documents.
 421 After fitting a topic model, each document in the training set, or any *new* document that contains at
 422 least some of the words in the model’s vocabulary, may be represented as a k -dimensional vector
 423 describing how much the document (most probably) reflects each topic. (Unless, otherwise noted,
 424 we used $k = 15$ topics.)

425 As illustrated in Figure 7A, we start by building up a corpus of documents using overlapping
 426 sliding windows that span each video’s transcript. Khan Academy videos are hosted on the

427 YouTube platform, and all YouTube videos are run through Google’s speech-to-text API [26] to
428 derive a timestamped transcript of any detected speech in the video. The resulting transcripts
429 contain one timestamped row per line, and each line generally corresponds to a few seconds of
430 spoken content from the video. We defined a sliding window length of (up to) $w = 30$ transcript
431 lines, and we assigned each window a timestamp according to the midpoint between its first
432 and last lines’ timestamps. These sliding windows ramped up and down in length at the very
433 beginning and end of the transcript, respectively. In other words, the first sliding window covered
434 only the first line from the transcript; the second sliding window covered the first two lines; and
435 so on. This insured that each line of the transcript appeared in the same number (w) of sliding
436 windows. We treated the text from each sliding window as a single “document,” and we combined
437 these documents across the two videos’ windows to create a single training corpus for the topic
438 model. The top words from each of the 15 discovered topics may be found in Table S2.

439 After fitting a topic model to each videos’ transcripts, we could use the trained model to
440 transform arbitrary (potentially new) documents into k -dimensional topic vectors. A convenient
441 property of these topic vectors is that documents that reflect similar blends of topics (i.e., documents
442 that reflect similar themes, according to the model) will yield similar (in terms of Euclidean distance,
443 correlation, etc.) topic vectors. In general, the similarity between different documents’ topic vectors
444 may be used to characterize the similarity in conceptual content between the documents.

445 We transformed each sliding window’s text into a topic vector, and then used linear interpo-
446 lation (independently for each topic dimension) to resample the resulting timeseries to once per
447 second. This yielded a single topic vector for each second of each video. We also used the fitted
448 model to obtain topic vectors for each question in our pool (Tab. S1). Taken together, we obtained
449 a *trajectory* for each video, describing its path through topic space, and a single coordinate for each
450 question (Fig. 7B). Embedding both videos and all of the questions using a common model enables
451 us to compare the content from different moments of videos, compare the content across videos,
452 and estimate potential associations between specific questions and specific moments of video.

453 **Estimating dynamic knowledge traces**

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

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

456 where

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

457 and where mincorr and maxcorr are the minimum and maximum correlations between any lecture
458 timepoint and question, taken over all timepoints and questions across both lectures and all three
459 question sets. We also define $f(s, \Omega)$ as the s^{th} topic vector from the set of topic vectors Ω . Here
460 t indexes the set of lecture topic vectors, L , and i and j index the topic vectors of questions in the
461 quiz’s question set, Q . Note that “correct” denotes the set of indices of the questions the participant
462 answered correctly on the given quiz.

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

471 **Creating knowledge and learning map visualizations**

472 An important feature of our approach is that, given a trained text embedding model and partic-
473 ipants’ quiz performance on each question, we can estimate their knowledge about *any* content
474 expressible by the embedding model– not solely the content explicitly probed by the quiz ques-

475 tions. To visualize these estimates (Figs. 6, S1, S2, S3, S4, and S5), we used UMAP [41] to define a
 476 2D projection of the text embedding space. Sampling the original 100-dimensional space at high
 477 resolution to obtain an adequate set of topic vectors spanning the embedding space would be
 478 computationally intractable. However, sampling a 2D grid is much more feasible. We defined a
 479 rectangle enclosing the 2D projections of the lectures' and quizzes' embeddings, and we sampled
 480 points from a regular 100×100 grid of coordinates that evenly tiled the enclosing rectangle. We
 481 sought to estimate participants' knowledge (and learning—i.e., changes in knowledge) at each of
 482 the resulting 10000 coordinates.

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

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

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

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

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

494 **Author contributions**

495 Conceptualization: PCF, ACH, and JRM. Methodology: PCF, ACH, and JRM. Software: PCF.
 496 Validation: PCF. Formal analysis: PCF. Resources: PCF, ACH, and JRM. Data curation: PCF.

497 Writing (original draft): JRM. Writing (review and editing): PCF, ACH, and JRM. Visualization:
498 PCH and JRM. Supervision: JRM. Project administration: PCH. Funding acquisition: JRM.

499 Data and code availability

500 All of the data analyzed in this manuscript, along with all of the code for running our experiment
501 and carrying out the analyses may be found at <https://github.com/ContextLab/efficient-learning-khan>.
502

503 Acknowledgements

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

511 References

- 512 [1] Aldrup, K., Carstensen, B., and Klusmann, U. (2022). Is empathy the key to effective teaching?
513 A systematic review of its association with teacher-student interactions and student outcomes.
514 *Educational Psychology Review*, 34:1177001216.
- 515 [2] Alloway, T. P. (2012). Teachers' perceptions of classroom behaviour and working memory.
516 *Educational Research and Review*, 7(6):138–142.
- 517 [3] Anderson, J. R. and Skwarecki, E. (1986). The automated tutoring of introductory computer
518 programming. *Communications of the ACM*, 29(9):842–849.

- 519 [4] Anderton, R. S., Vitali, J., Blackmore, C., and Bakeberg, M. C. (2021). Flexible teaching and learn-
520 ing modalities in undergraduate science amid the COVID-19 pandemic. *Frontiers in Education*,
521 5:doi.org/10.3389/feduc.2020.609703.
- 522 [5] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and
523 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
524 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 525 [6] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
526 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
527 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 528 [7] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*
529 *Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
530 Machinery.
- 531 [8] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
532 *Learning Research*, 3:993–1022.
- 533 [9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
534 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
535 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
536 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
537 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 538 [10] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
539 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 540 [11] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
541 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
542 sentence encoder. *arXiv*, 1803.11175.

- 543 [12] Clark, J. (2010). Powerpoint and pedagogy: maintaining student interest in university lectures.
- 544 *College Teaching*, 56(1):39–44.
- 545 [13] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
- 546 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 547 [14] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
- 548 Evidence for a new conceptualization of semantic representation in the left and right cerebral
- 549 hemispheres. *Cortex*, 40(3):467–478.
- 550 [15] den Brok, P., van Tartwijk, J., Wubbels, T., and Veldman, I. (2010). The differential effect of
- 551 the teacher-student interpersonal relationship on student outcomes for students with different
- 552 ethnic backgrounds. *British Journal of Educational Psychology*, 80(2):199–221.
- 553 [16] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
- 554 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
- 555 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 556 [17] Englehart, J. M. (2009). Teacher-student interaction. In Saha, L. J. and Dworkin, A. G., editors,
- 557 *International Handbook of Research on Teachers and Teaching*. Springer International Handbooks of
- 558 Education.
- 559 [18] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*
- 560 *Transactions of the Royal Society A*, 222(602):309–368.
- 561 [19] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
- 562 *School Science and Mathematics*, 100(6):310–318.
- 563 [20] Garran, A. M. and Rasmussen, B. M. (2014). Safety in the classroom: reconsidered. *Journal of*
- 564 *Teaching in Social Work*, 34(4):401–412.
- 565 [21] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*
- 566 *Cognition and Development*, 13(1):19–37.

- 567 [22] Goode, S., Willis, R. A., Wolf, J. R., and Harris, A. L. (2007). Enhancing IS education with
568 flexible teaching and learning. *Journal of Information Systems Education*, 18(3):297–302.
- 569 [23] Halff, H. M. (1988). Curriculum and instruction in automated tutors. *Foundations of intelligent
570 tutoring systems*, pages 79–108.
- 571 [24] Hall, J. K. and Walsh, M. (2002). Teacher-student interaction and language learning. *Annual
572 Review of Applied Linguistics*, 22:186–203.
- 573 [25] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
574 learning, pages 212–221. Sage Publications.
- 575 [26] Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml, M. (2016).
576 Contextual prediction models for speech recognition. In *Interspeech*, pages 2338–2342.
- 577 [27] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal be-
578 havioral and neural signatures of transforming naturalistic experiences into episodic memories.
579 *Nature Human Behavior*, 5:905–919.
- 580 [28] Johnston, S. (2002). Introducing and supporting change towards more flexible teaching ap-
581 proaches. In *The convergence of distance and conventional education*. Routledge.
- 582 [29] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
583 empathy and theory of mind. In *Social Behavior from Rodents to Humans*, pages 193–206. Springer.
- 584 [30] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
585 Columbia University Press.
- 586 [31] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
587 326(7382):213–216.
- 588 [32] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
589 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International
590 Journal of Environmental Research and Public Health*, 18(5):2672.

- 591 [33] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 592 [34] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 593 [35] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
- 594 *The Chronicle of Higher Education*, 21:1–5.
- 595 [36] Kumar, A. N. (2005). Generation of problems, answers, grade, and feedback—case study of a
596 fully automated tutor. *Journal on Educational Resources in Computing*, 5(3):1–25.
- 597 [37] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic
598 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
599 104:211–240.
- 600 [38] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of
601 Educational Studies*, 53(2):129–147.
- 602 [39] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
603 *Handbook of Human Memory*. Oxford University Press.
- 604 [40] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
605 function? *Psychological Review*, 128(4):711–725.
- 606 [41] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
607 projection for dimension reduction. *arXiv*, 1802(03426).
- 608 [42] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
609 mind. In *The Wiley-Blackwell handbook of childhood cognitive development*. Wiley-Blackwell.
- 610 [43] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
611 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
612 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 613 [44] Meyers, S., Rowell, K., Wells, M., and Smith, B. C. (2019). Teacher empathy: a model of
614 empathy for teaching for student success. *College Teaching*, 67(3):160–168.

- 615 [45] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
616 tations in vector space. *arXiv*, 1301.3781.
- 617 [46] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
618 from a national survey of language educators. *System*, 97:102431.
- 619 [47] Muijs, D. and Reynolds, D. (2003). Student background and teacher effects on achievement and
620 attainment in mathematics: a longitudinal study. *Educational Research and Evaluation*, 9(3):289–
621 314.
- 622 [48] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
623 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 624 [49] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
625 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective
626 Neuroscience*, 17(4):367–376.
- 627 [50] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
628 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
629 7:43916.
- 630 [51] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of
631 Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 632 [52] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
633 higher education: unmasking power and raising questions about the movement’s democratic
634 potential. *Educational Theory*, 63(1):87–110.
- 635 [53] Rosenshine, B. (1976). Recent research on teaching behaviors and student achievement. *Journal
636 of Teacher Education*, 27(1):61–64.
- 637 [54] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
638 Student conceptions and conceptual learning in science. Routledge.

- 639 [55] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
640 based empathy training improves the communication skills of neonatal nurses. *Clinical Simulation in Nursing*, 22:32–42.
- 642 [56] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching
643 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 644 [57] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
645 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in
646 Mathematics Education*, 35(5):305–329.
- 647 [58] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal
648 Medicine*, 21:524–530.
- 649 [59] Stojiljković, S., Djigić, G., and Zlatković, B. (2012). Empathy and teachers' roles. *Procedia –
650 Social and Behavioral Sciences*, 69:960–966.
- 651 [60] Swarat, S., Ortony, A., and Revelle, W. (2012). Activity matters: understanding student interest
652 in school science. *Journal of Research in Science Teaching*, 49(4):515–537.
- 653 [61] Turner, S. and Braine, M. (2015). Unravelling the 'safe' concept in teaching: what can we learn
654 from teachers' understanding? *Pastoral Care in Education*, 33(1):47–62.
- 655 [62] van de Pol, J., Volman, M., and Beishuizen, J. (2010). Scaffolding in teacher-student interaction:
656 a decade of research. *Educational Psychology Review*, 22:271–296.
- 657 [63] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
658 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 659 [64] Wolz, U., McKeown, K., and Kaiser, G. E. (1988). Automated tutoring in interactive environ-
660 ments: a task centered approach. Technical report, Columbia University.
- 661 [65] Wulff, S. S. and Wulff, D. H. (2004). "of course i'm communicating; I lecture every day":
662 enhancing teaching and learning in introductory statistics. *Communication Education*, 53(1):92–
663 103.

- 664 [66] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
665 free courses. *The Chronicle of Higher Education*, 19(7):1–4.