

¹ Text embedding models enable high-resolution insights
² into conceptual knowledge and learning in
³ classroom-like settings

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

¹Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

²Akili Interactive Labs

Boston, MA 02110, USA

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

5

Abstract

6 We develop a mathematical framework, based on natural language processing models, for track-
7 ing and characterizing the acquisition of conceptual knowledge in real-world educational con-
8 texts. Our approach embeds course content in a high-dimensional conceptual space, where
9 nearby coordinates reflect similar or related concepts. We test our approach using behavioral
10 data from participants who viewed two lecture videos from the Khan Academy platform, inter-
11 leaved between three short multiple-choice quizzes. We applied our framework to the lectures'
12 transcripts and the text of the quiz questions to quantify the conceptual content presented in each
13 moment of video and knowledge probed by each quiz question. We used these embeddings,
14 along with participants' quiz responses, to track how the learners' knowledge changed after
15 watching each video. Our findings demonstrate how a small set of quiz questions may be used
16 to obtain rich and meaningful high-resolution insights into individuals' knowledge, and how it
17 changes over time as they learn.

19

Introduction

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

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

40 The difference between “understanding” and “memorizing,” as framed by researchers in ed-
41 ucation, cognitive psychology, and cognitive neuroscience [e.g., 19, 25, 30, 38, 54] has profound
42 analogs in the fields of natural language processing and natural language understanding. For
43 example, considering the raw contents of a document (e.g., its constituent symbols, letters, and

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

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

68 Here we develop a framework for modeling how knowledge is acquired during learning. The
69 central idea is to use text embedding models to define the coordinate systems of two maps: (a) a
70 *knowledge map* that describes the extent to which each concept is currently known and (b) a *learning*
71 *map* that describes changes in knowledge over time. Each location on these maps represents

⁷² a single concept, and the maps' geometries are defined such that related concepts are located
⁷³ nearby in space. We use this framework to analyze and interpret behavioral data collected from
⁷⁴ an experiment that had participants watch and answer multiple-choice questions about a series of
⁷⁵ recorded course lectures.

⁷⁶ Our primary research goal is to advance our understanding of what it means to acquire deep,
⁷⁷ real-world conceptual knowledge. Traditional laboratory approaches to studying learning and
⁷⁸ memory (e.g., list learning studies) often draw little distinction between memorization and under-
⁷⁹ standing. Instead, these studies typically focus on whether information is effectively encoded or
⁸⁰ retrieved, rather than whether the information is *understood*. Approaches to studying conceptual
⁸¹ learning, such as category learning experiments, can begin to investigate the distinction between
⁸² memorization and understanding, often by training participants to distinguish arbitrary or ran-
⁸³ dom features in otherwise meaningless categorized stimuli. However the objective of real-world
⁸⁴ training, or learning from life experiences more generally, is often to develop new knowledge that
⁸⁵ may be applied in *useful* ways in the future. In this sense, the gap between modern learning theo-
⁸⁶ ries and modern pedagogical approaches and classroom learning strategies is enormous: most of
⁸⁷ our theories about *how* people learn are inspired by experimental paradigms and models that have
⁸⁸ only peripheral relevance to the kinds of learning that students and teachers actually seek [25, 38].
⁸⁹ To help bridge this gap, our study uses course materials from real online courses to inform, fit,
⁹⁰ and test models of real-world conceptual learning. We also provide a demonstration of how our
⁹¹ models can be used to construct “maps” of what students know, and how their knowledge changes
⁹² with training. In addition to helping to visualize knowledge (and changes in knowledge), we hope
⁹³ that such maps might lead to real-world tools for improving how we educate.

⁹⁴ Results

⁹⁵ At its core, our main modeling approach is based around a simple assumption that we sought to
⁹⁶ test empirically: all else being equal, knowledge about a given concept is predictive of knowledge
⁹⁷ about similar or related concepts. From a geometric perspective, this assumption implies that

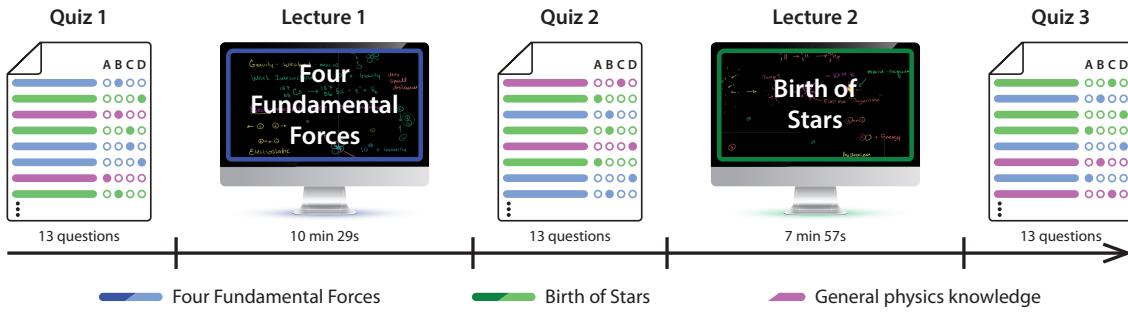


Figure 1: Experimental paradigm. Participants alternate between answering 13-question multiple-choice quizzes and watching two Khan Academy lectures. Each quiz contains a mix of 5 questions about lecture 1, 5 questions about lecture 2, and 3 questions about general physics knowledge. The specific questions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

knowledge is fundamentally “smooth.” In other words, as one moves through a space representing an individual’s 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 participants’ knowledge and how it changed over time in response to training.

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 lessons to be (a) accessible to a broad audience, i.e., requiring minimal prerequisite knowledge to understand; (b) conceptually related to each other, i.e., covering at least *some* similar or overlapping content; and (c) largely independent of each other, i.e., focused on sufficiently different material that understanding one did not require having seen the other. The two videos we selected are introductory lectures that both belong to Khan Academy’s “Cosmology and Astronomy” course domain, but are taken from different lecture series (“Scale of the Universe” and “Stars, Black Holes, and Galaxies” for the first and second lectures, respectively).

We then created a pool of multiple-choice quiz questions that would enable us to test partici-

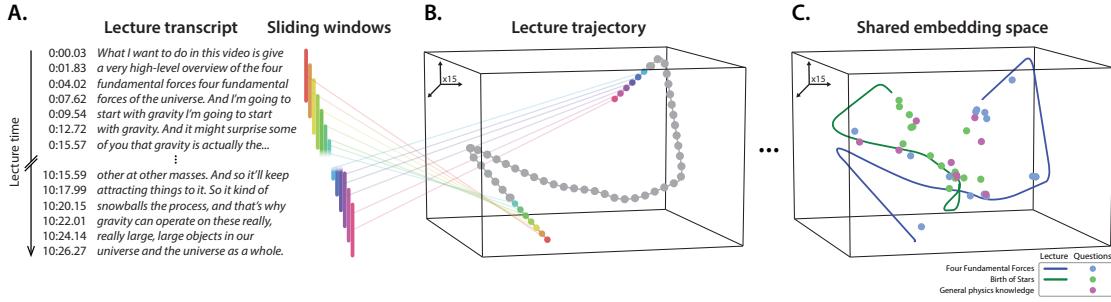


Figure 2: Modeling course content. **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding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

116 pants’ knowledge about each individual lecture, as well as related content not specifically presented
 117 in either video (see Tab. S1). Participants answered questions randomly drawn from each content
 118 area (lecture 1, lecture 2, and general physics knowledge) on each of the three quizzes. Quiz 1 was
 119 intended to assess participants’ “baseline” knowledge before training, quiz 2 assessed knowledge
 120 after watching the *Four Fundamental Forces* video (i.e., lecture 1), and quiz 3 assessed knowledge
 121 after watching the *Birth of Stars* video (i.e., lecture 2).

122 To study how participants’ conceptual knowledge changed over the course of the experiment,
 123 we first sought to characterize the abstract concepts presented to them in each of the two lectures.
 124 We adapted an approach we developed in prior work [27] to extract the latent themes from the
 125 lectures’ contents using a topic model [8]. Briefly, topic models take as input a collection of text doc-
 126 uments and learn a set of “topics” (i.e., latent themes) from their contents. Once fit, a topic model
 127 can be used to transform arbitrary (potentially new) documents into sets of “topic proportions,” de-
 128 scribing the weighted blend of learned topics reflected in their texts. We parsed automatically gen-
 129 erated transcripts of the two lectures into overlapping sliding windows, which we treated as doc-
 130 uments to fit our model (Fig. 2A; see *Constructing text embeddings of multiple lectures and questions*).

131 Transforming these windows with our model yielded a number-of-windows by number-of-topics
132 (15) topic-proportions matrix, denoting the unique mixture of broad themes from both lectures
133 reflected in each window’s content. Intuitively, each window’s “topic vector” (i.e., column of the
134 topic-proportions matrix) is analogous to a coordinate in a 15-dimensional space (whose axes are
135 topics discovered by the model). Within this space, each lecture’s sequence of topic vectors (i.e.,
136 corresponding to sliding windows parsed from its transcript) forms a *trajectory* that captures how
137 its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution
138 of 1 topic vector for each second of video.

139 Next, we sought to characterize what information participants’ performance on each of the
140 three quizzes could provide about their conceptual knowledge at that point in time. Traditional
141 approaches to evaluating students’ performance on short-form knowledge assessments, such as
142 those in our study, entail computing the proportion of correctly answered questions to assign a
143 simple numeric score. While this may afford some measure of the *extent* of a learner’s knowledge,
144 such a score provides little information to either the teacher or the learner about the particular
145 *contents* of their knowledge—in other words, raw “proportion-correct” measures may capture *how*
146 *much* a student knows, but not *what* they know. For instance, suppose a participant in our study was
147 highly knowledgeable about fundamental physical forces (i.e., lecture 1 content) but unfamiliar
148 with how stars are formed (i.e., lecture 2 content), while a second participant was unfamiliar
149 with fundamental forces but had extensive prior knowledge about star formation. Since quiz
150 1 (completed before viewing either lecture) contained an equal number of questions about each
151 lecture’s content (see Fig. 1), these two participants could easily achieve the same proportion-correct
152 score despite their conceptual knowledge differing substantially. How might we distinguish these
153 individuals?

154 We hypothesized that our text embedding model, which we had trained on transcripts of the two
155 lectures to characterize their conceptual contents, should also allow us to capture the conceptual
156 knowledge probed by each quiz question. If our model successfully represented information about
157 the deeper conceptual content of the lectures (i.e., beyond surface-level details such as particular
158 word choices) then we should be able to recover a correspondence between their embeddings and

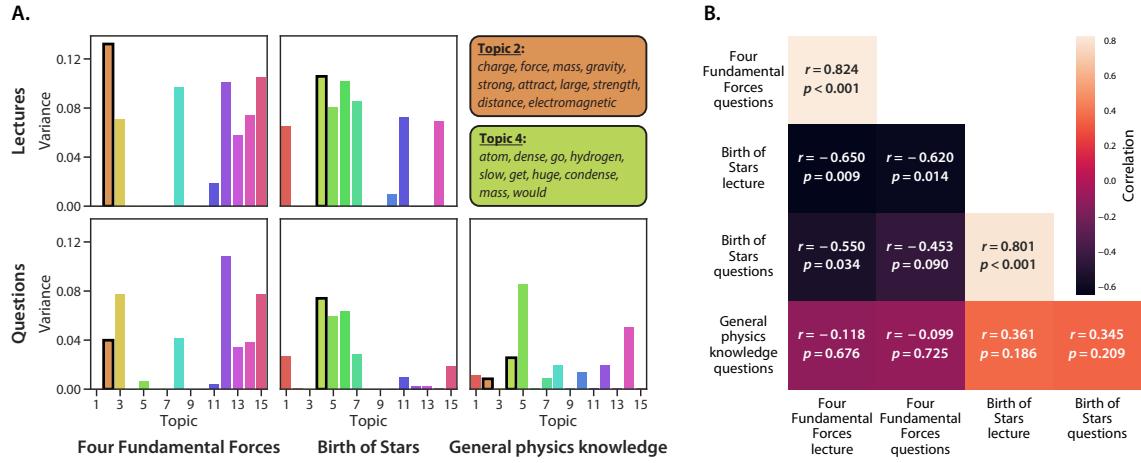


Figure 3: Lecture and question topic overlap. A. Topic weight variability. The bar plots display the variance of each topic's weight across lecture timepoints (top row) and questions (bottom row); colors denote topics. The top-weighted words from the most “expressive” (i.e., variable across observations)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

embeddings of other text that reflects related concepts (e.g., quiz questions that were ostensibly, by design, “about” one of the two lecture) despite differences in exact verbiage. Intuitively, while the content of each given lecture will naturally vary from moment to moment, the particular combination of dimensions along which its trajectory varies across moments correspond to topics that encode meaningful “information” [18] about that lecture’s dynamic conceptual content. We therefore expected that quiz questions pertaining to those same concepts might express the same set of topics.

We used our model to transform the text of each question in our pool into the same embedding space as the two lectures’ trajectories (Fig. 2C). This yielded a single 15-dimensional coordinate (i.e., topic vector) for each question. We then computed the variance of each topic’s weight across timepoints of each lecture, and across questions from each category (i.e., lecture 1-related, lecture 2-related, and general physics knowledge; Fig. 3A). Visual inspection of Figure 3A reveals a strong correspondence between the sets of topics expressed by each lecture and its related questions, as well as limited overlap in the topics expressed by non-matched lecture-question

set pairs. To quantify these apparent similarities and differences, we computed the correlation between each pair of topic-weight variance distributions (Fig. 3B). We found that our model captured the contents of lecture-related questions using a weighted blend of topic dimensions that was highly similar to that of their reference lecture (*Four Fundamental Forces* (FFF) questions vs. lecture: Pearson’s $r(13) = 0.824$, $p < 0.001$, 95% confidence interval (CI) = [0.696, 0.973]; *Birth of Stars* (BoS) questions vs. lecture: $r(13) = 0.801$, $p < 0.001$, 95% CI = [0.539, 0.958]) but diverged significantly from that of the non-reference lecture (FFF questions vs. BoS lecture: $r(13) = -0.620$, $p = 0.014$, 95% CI = [-0.871, -0.326]; BoS questions vs. FFF lecture: $r(13) = -0.550$, $p = 0.034$, 95% CI = [-0.803, -0.246]). This indicated that our model was sufficiently sensitive to the conceptual content tested by different quiz questions to differentiate between those relating to one lecture versus the other, and allow us to

sensitive to the underlying conceptual content tested by different question, and should allow us to differentially assess participants’ knowledge for the two lectures based on the *specific* questions they answered correctly or incorrectly on each quiz.

Although an individual 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. 4). In particular, most questions’ topic vectors were maximally correlated with a well-defined (and relatively narrow) range of timepoints from their corresponding lectures, and the correlations fell off sharply outside of that range. We also 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

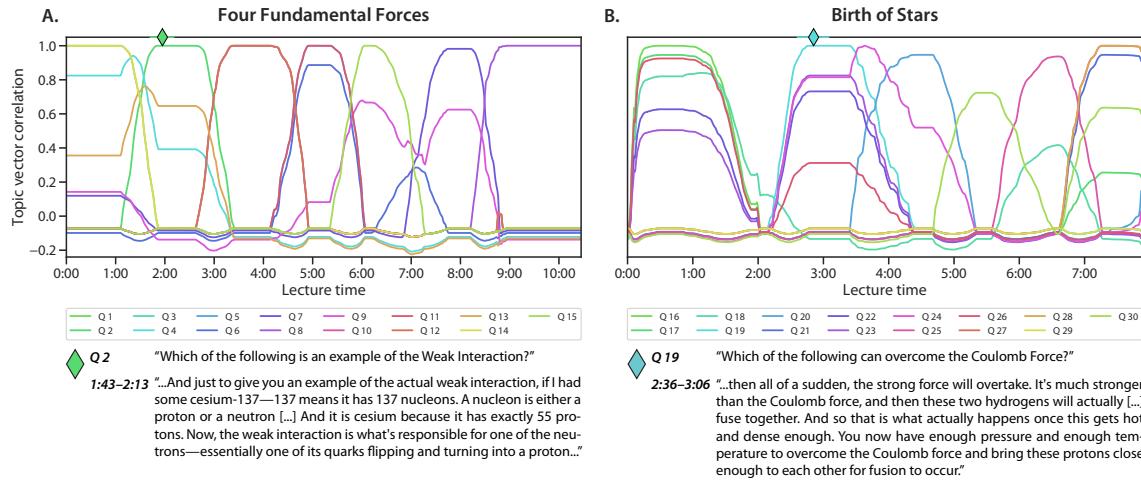


Figure 4: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel B.: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

201 the content covered by the best-matching moments of the lectures. Two representative examples
202 are shown at the bottom of Figure 4.

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

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

228 Of course, even though the timecourses in Figure 5A and C provide detailed *estimates* about



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.

229 participants' knowledge, those estimates are only *useful* to the extent that they accurately reflect what
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



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.

257 the coordinates of each *incorrectly* answered question (Fig. 6). We then used independent samples
 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 each
281 question) into a shared 2-dimensional space (see *Creating knowledge and learning map visualizations*).
282 Next, we sampled points evenly from a 100×100 grid of coordinates that evenly tiled a rectangle
283 enclosing the 2D projections of the videos and questions. We used Equation 4 to estimate partici-
284 pants’ knowledge at each of these 10K sampled locations, and we averaged these estimates across
285 participants to obtain an estimated average *knowledge map* (Fig. 7). Intuitively, the knowledge map
286 constructed from a given quiz’s responses provides a visualization of how “much” participants
287 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.



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 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.

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

322 Discussion

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

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

337 The notion of empathy also related to "theory of mind" of other individuals [21, 29, 42].
338 Considering others' unique perspectives, prior experiences, knowledge, goals, etc., can help us
339 to more effectively interact and communicate [51, 55, 58]. The knowledge and learning maps
340 we estimate in our study (Fig. 7) hint at one potential form that an automated "empathetic"
341 teacher might take. We imagine automated content delivery systems that adapt lessons on the
342 fly according to continually updated estimates of what students know and how quickly they are
343 learning different conceptual content [e.g., building on ideas such as 3, 23, 36, 64, and others].

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

355 Alongside these questions, there is a growing desire to extend existing theories beyond the
356 domain of lab testing rooms and into real classrooms [31]. In part, this has led to a recent
357 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better
358 reflect more ethologically valid phenomena that are more directly relevant to real-world situations
359 and behaviors [48]. In turn, this has brought new challenges in data analysis and interpretation. A
360 key step towards solving these challenges will be to build explicit models of real-world scenarios
361 and how people behave in them (e.g., models of how people learn conceptual content from real-
362 world courses, as in our current study). A second key step will be to understand which sorts
363 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 5,
364 16, 43, 49, 50] might help to inform these models. A third major step will be to develop and
365 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic
366 paradigms.

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

³⁷⁷ **Materials and methods**

³⁷⁸ **Participants**

³⁷⁹ We enrolled a total of 50 Dartmouth undergraduate students in our study. Participants received
³⁸⁰ course credit for enrolling. We asked each participant to fill out a demographic survey that included
³⁸¹ questions about their age, gender, native spoken language, ethnicity, race, hearing, color vision,
³⁸² sleep, coffee consumption, level of alertness, and several aspects of their educational background
³⁸³ and prior coursework.

³⁸⁴ Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
³⁸⁵ years). A total of 15 participants reported their gender as male and 35 participants reported their
³⁸⁶ gender as female. A total of 49 participants reported their native language as "English" and 1
³⁸⁷ reported having another native language. A total of 47 participants reported their ethnicity as
³⁸⁸ "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants
³⁸⁹ reported their races as White (32 participants), Asian (14 participants), Black or African American
³⁹⁰ (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
³⁹¹ Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

³⁹² A total of 49 participants reporting having normal hearing and 1 participant reported having
³⁹³ some hearing impairment. A total of 49 participants reported having normal color vision and 1
³⁹⁴ participant reported being color blind. Participants reported having had, on the night prior to
³⁹⁵ testing, 2–4 hours of sleep (1 participant), 4–6 hours of sleep (9 participants), 6–8 hours of sleep (35
³⁹⁶ participants), or 8+ hours of sleep (5 participants). They reported having consumed, on the same
³⁹⁷ day and leading up to their testing session, 0 cups of coffee (38 participants), 1 cup of coffee (10
³⁹⁸ participants), 3 cups of coffee (1 participant), or 4+ cups of coffee (1 participant).

³⁹⁹ No participants reported that their focus was currently impaired (e.g., by drugs or alcohol).
⁴⁰⁰ Participants reported their current level of alertness, and we converted their responses to numerical
⁴⁰¹ scores as follows: "very sluggish" (-2), "a little sluggish" (-1), "neutral" (0), "fairly alert" (1), and
⁴⁰² "very alert" (2). Across all participants, a range of alertness levels were reported (range: -2 – 1;
⁴⁰³ mean: -0.10; standard deviation: 0.84).

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

432 ipants), Computing (2 participants), and other categories not listed in our survey (18 participants).
433 Finally, we asked participants about in-person courses they had taken in different subject areas.
434 They reported taking courses in Mathematics (39 participants), Science and engineering (38 par-
435 ticipants), Arts and humanities (35 participants), Test preparation (27 participants), Economics
436 and finance (26 participants), Computing (15 participants), College and careers (7 participants), or
437 other courses not listed in our survey (6 participants).

438 Experiment

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

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

457 **Analysis**

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

459 Following Heusser et al. [27]

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

466 We adapted an approach we developed in prior work [27] to embed each moment of the two
467 lectures and each question in our pool in a common representational space. Briefly, our approach
468 uses a topic model [Latent Dirichlet Allocation; 8], trained on a set of documents, to discover a
469 set of k “topics” or “themes.” Formally, each topic is defined as a set of weights over each word
470 in the model’s vocabulary (i.e., the union of all unique words, across all documents, excluding
471 “stop words.”). Conceptually, each topic is intended to give larger weights to words that are
472 conceptually related or that tend to co-occur in the same documents. After fitting a topic model,
473 each document in the training set, or any *new* document that contains at least some of the words in
474 the model’s vocabulary, may be represented as a k -dimensional vector describing how much the
475 document (most probably) reflects each topic. (Unless, otherwise noted, we used $k = 15$ topics.)

476 As illustrated in Figure 2A, we first

477 As illustrated in Figure 2A, we start by building up a corpus of documents using overlapping
478 sliding windows that span each video’s transcript. Khan Academy provides professionally created,
479 manual transcriptions of all videos for closed captioning. However, such transcripts would not
480 be readily available in all contexts to which our framework could potentially be applied. Khan
481 Academy videos are hosted on the YouTube platform, which additionally provides automated
482 transcripts

483 Khan Academy videos are hosted on the YouTube platform, and all YouTube videos are run

484 through Google’s speech-to-text API [26] to derive a timestamped transcript of any detected speech
485 in the video. The resulting transcripts contain one timestamped row per line, and each line
486 generally corresponds to a few seconds of spoken content from the video. We defined a sliding
487 window length of (up to) $w = 30$ transcript lines, and we assigned each window a timestamp
488 according to the midpoint between its first and last lines’ timestamps. These sliding windows
489 ramped up and down in length at the very beginning and end of the transcript, respectively. In
490 other words, the first sliding window covered only the first line from the transcript; the second
491 sliding window covered the first two lines; and so on. This insured that each line of the transcript
492 appeared in the same number (w) of sliding windows. We treated the text from each sliding window
493 as a single “document,” and we combined these documents across the two videos’ windows to
494 create a single training corpus for the topic model. The top words from each of the 15 discovered
495 topics may be found in Table S2.

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

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

510 **Estimating dynamic knowledge traces**

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

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

513 where

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

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

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

528 **Creating knowledge and learning map visualizations**

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

532 tions. To visualize these estimates (Figs. 7, S1, S2, S3, S4, and S5), we used UMAP [41] to define a
 533 2D projection of the text embedding space. Sampling the original 100-dimensional space at high
 534 resolution to obtain an adequate set of topic vectors spanning the embedding space would be
 535 computationally intractable. However, sampling a 2D grid is much more feasible. We defined a
 536 rectangle enclosing the 2D projections of the lectures' and quizzes' embeddings, and we sampled
 537 points from a regular 100×100 grid of coordinates that evenly tiled the enclosing rectangle. We
 538 sought to estimate participants' knowledge (and learning—i.e., changes in knowledge) at each of
 539 the resulting 10000 coordinates.

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

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

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

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

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

551 **Author contributions**

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

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

556 Data and code availability

557 All of the data analyzed in this manuscript, along with all of the code for running our experiment
558 and carrying out the analyses may be found at <https://github.com/ContextLab/efficient-learning-khan>.
559

560 Acknowledgements

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

568 References

- 569 [1] Aldrup, K., Carstensen, B., and Klusmann, U. (2022). Is empathy the key to effective teaching?
570 A systematic review of its association with teacher-student interactions and student outcomes.
571 *Educational Psychology Review*, 34:1177001216.
- 572 [2] Alloway, T. P. (2012). Teachers' perceptions of classroom behaviour and working memory.
573 *Educational Research and Review*, 7(6):138–142.
- 574 [3] Anderson, J. R. and Skwarecki, E. (1986). The automated tutoring of introductory computer
575 programming. *Communications of the ACM*, 29(9):842–849.

- 576 [4] Anderton, R. S., Vitali, J., Blackmore, C., and Bakeberg, M. C. (2021). Flexible teaching and learn-
577 ing modalities in undergraduate science amid the COVID-19 pandemic. *Frontiers in Education*,
578 5:doi.org/10.3389/feduc.2020.609703.
- 579 [5] Bevilacque, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and
580 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
581 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 582 [6] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
583 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
584 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- 585 [7] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*
586 *Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
587 Machinery.
- 588 [8] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
589 *Learning Research*, 3:993–1022.
- 590 [9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
591 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
592 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
593 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
594 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 595 [10] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
596 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 597 [11] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
598 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
599 sentence encoder. *arXiv*, 1803.11175.

- 600 [12] Clark, J. (2010). Powerpoint and pedagogy: maintaining student interest in university lectures.
- 601 *College Teaching*, 56(1):39–44.
- 602 [13] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
- 603 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 604 [14] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
- 605 Evidence for a new conceptualization of semantic representation in the left and right cerebral
- 606 hemispheres. *Cortex*, 40(3):467–478.
- 607 [15] den Brok, P., van Tartwijk, J., Wubbels, T., and Veldman, I. (2010). The differential effect of
- 608 the teacher-student interpersonal relationship on student outcomes for students with different
- 609 ethnic backgrounds. *British Journal of Educational Psychology*, 80(2):199–221.
- 610 [16] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
- 611 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
- 612 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 613 [17] Englehart, J. M. (2009). Teacher-student interaction. In Saha, L. J. and Dworkin, A. G., editors,
- 614 *International Handbook of Research on Teachers and Teaching*. Springer International Handbooks of
- 615 Education.
- 616 [18] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*
- 617 *Transactions of the Royal Society A*, 222(602):309–368.
- 618 [19] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
- 619 *School Science and Mathematics*, 100(6):310–318.
- 620 [20] Garran, A. M. and Rasmussen, B. M. (2014). Safety in the classroom: reconsidered. *Journal of*
- 621 *Teaching in Social Work*, 34(4):401–412.
- 622 [21] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*
- 623 *Cognition and Development*, 13(1):19–37.

- 624 [22] Goode, S., Willis, R. A., Wolf, J. R., and Harris, A. L. (2007). Enhancing IS education with
625 flexible teaching and learning. *Journal of Information Systems Education*, 18(3):297–302.
- 626 [23] Halff, H. M. (1988). Curriculum and instruction in automated tutors. *Foundations of intelligent
627 tutoring systems*, pages 79–108.
- 628 [24] Hall, J. K. and Walsh, M. (2002). Teacher-student interaction and language learning. *Annual
629 Review of Applied Linguistics*, 22:186–203.
- 630 [25] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
631 learning, pages 212–221. Sage Publications.
- 632 [26] Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml, M. (2016).
633 Contextual prediction models for speech recognition. In *Interspeech*, pages 2338–2342.
- 634 [27] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal be-
635 havioral and neural signatures of transforming naturalistic experiences into episodic memories.
636 *Nature Human Behavior*, 5:905–919.
- 637 [28] Johnston, S. (2002). Introducing and supporting change towards more flexible teaching ap-
638 proaches. In *The convergence of distance and conventional education*. Routledge.
- 639 [29] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
640 empathy and theory of mind. In *Social Behavior from Rodents to Humans*, pages 193–206. Springer.
- 641 [30] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
642 Columbia University Press.
- 643 [31] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
644 326(7382):213–216.
- 645 [32] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
646 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International
647 Journal of Environmental Research and Public Health*, 18(5):2672.

- 648 [33] Khan, S. (2004). *The Khan Academy*. Salman Khan.
- 649 [34] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 650 [35] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
- 651 *The Chronicle of Higher Education*, 21:1–5.
- 652 [36] Kumar, A. N. (2005). Generation of problems, answers, grade, and feedback—case study of a
653 fully automated tutor. *Journal on Educational Resources in Computing*, 5(3):1–25.
- 654 [37] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic
655 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
656 104:211–240.
- 657 [38] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of
658 Educational Studies*, 53(2):129–147.
- 659 [39] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
660 *Handbook of Human Memory*. Oxford University Press.
- 661 [40] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
662 function? *Psychological Review*, 128(4):711–725.
- 663 [41] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
664 projection for dimension reduction. *arXiv*, 1802(03426).
- 665 [42] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
666 mind. In *The Wiley-Blackwell handbook of childhood cognitive development*. Wiley-Blackwell.
- 667 [43] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
668 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
669 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 670 [44] Meyers, S., Rowell, K., Wells, M., and Smith, B. C. (2019). Teacher empathy: a model of
671 empathy for teaching for student success. *College Teaching*, 67(3):160–168.

- 672 [45] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
673 tations in vector space. *arXiv*, 1301.3781.
- 674 [46] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
675 from a national survey of language educators. *System*, 97:102431.
- 676 [47] Muijs, D. and Reynolds, D. (2003). Student background and teacher effects on achievement and
677 attainment in mathematics: a longitudinal study. *Educational Research and Evaluation*, 9(3):289–
678 314.
- 679 [48] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
680 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 681 [49] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
682 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective
683 Neuroscience*, 17(4):367–376.
- 684 [50] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
685 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
686 7:43916.
- 687 [51] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of
688 Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.
- 689 [52] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
690 higher education: unmasking power and raising questions about the movement’s democratic
691 potential. *Educational Theory*, 63(1):87–110.
- 692 [53] Rosenshine, B. (1976). Recent research on teaching behaviors and student achievement. *Journal
693 of Teacher Education*, 27(1):61–64.
- 694 [54] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
695 Student conceptions and conceptual learning in science. Routledge.

- 696 [55] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
697 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-*
698 *tion in Nursing*, 22:32–42.
- 699 [56] Shim, T. E. and Lee, S. Y. (2020). College students' experience of emergency remote teaching
700 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 701 [57] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
702 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*
703 *Mathematics Education*, 35(5):305–329.
- 704 [58] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal*
705 *Medicine*, 21:524–530.
- 706 [59] Stojiljković, S., Djigić, G., and Zlatković, B. (2012). Empathy and teachers' roles. *Procedia –*
707 *Social and Behavioral Sciences*, 69:960–966.
- 708 [60] Swarat, S., Ortony, A., and Revelle, W. (2012). Activity matters: understanding student interest
709 in school science. *Journal of Research in Science Teaching*, 49(4):515–537.
- 710 [61] Turner, S. and Braine, M. (2015). Unravelling the 'safe' concept in teaching: what can we learn
711 from teachers' understanding? *Pastoral Care in Education*, 33(1):47–62.
- 712 [62] van de Pol, J., Volman, M., and Beishuizen, J. (2010). Scaffolding in teacher-student interaction:
713 a decade of research. *Educational Psychology Review*, 22:271–296.
- 714 [63] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
715 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 716 [64] Wolz, U., McKeown, K., and Kaiser, G. E. (1988). Automated tutoring in interactive environ-
717 ments: a task centered approach. Technical report, Columbia University.
- 718 [65] Wulff, S. S. and Wulff, D. H. (2004). "of course i'm communicating; I lecture every day":
719 enhancing teaching and learning in introductory statistics. *Communication Education*, 53(1):92–
720 103.

⁷²¹ [66] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
⁷²² free courses. *The Chronicle of Higher Education*, 19(7):1–4.