

1 **Geometric models reveal the hidden structure of**
2 **conceptual knowledge**

3 Paxton C. Fitzpatrick and Jeremy R. Manning*

Dartmouth College

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

4 **Abstract**

5 We develop a mathematical framework, based on natural language processing models, for
6 tracking and characterizing the acquisition of conceptual knowledge. Our approach embeds each
7 concept in a high dimensional representation space, where nearby coordinates reflect similar or
8 related concepts. We tested our approach using behavioral data collected from a group of
9 college students. In the experiment, we asked the participants to answer sets of quiz questions
10 interleaved between watching two course videos from the Khan Academy platform. We applied
11 our framework to the videos' transcripts, and to text of the quiz questions, to quantify the
12 content of each moment of video and each quiz question. We used these embeddings, along with
13 participants' quiz responses, to track how the learners' knowledge changed after watching each
14 video. Our findings show how a limited set of quiz questions may be used to construct rich and
15 meaningful representations of what each learner knows, and how their knowledge changes over
16 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 a 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 to
³³ require something deeper and richer. Does conceptual understanding entail connecting newly
³⁴ acquired information to the scaffolding of one’s existing knowledge or experience [1, 5, 7, 8, 22]?
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 [14, 16, 21].

³⁹ The difference between “understanding” and “memorizing,” as framed by the researchers
⁴⁰ in education, cognitive psychology, and cognitive neuroscience [9, 10, 13, 16, 21] 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; 2, 3, 15]. However, modern natural language
46 processing models [e.g., 4, 6, 20] 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 [17, 18]. 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 the extent to which each concept could be learned. Each location on these maps
71 represents a single concept, and the 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 an

73 experiment that has participants watch and answer conceptual questions about a series of recorded
74 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

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

103 We asked our participants to answer questions from several multiple choice quizzes and watch
104 two lecture videos from the *Khan Academy* platform (Fig. 1). One lecture video, entitled *Four*
105 *Fundamental Forces*, was about the four fundamental forces in physics: gravity, strong and weak
106 interactions, and electromagnetism. The second lecture video, entitled *Birth of Stars*, provides
107 an overview of our current understanding of how stars form. We selected both lessons to be (a)
108 accessible to a broad audience, e.g., by minimizing prerequisite knowledge, (b) largely independent
109 of each other, e.g., so that the two videos focused on different material and did not depend on
110 each other, and (c) related to each other, e.g., so that both videos contained at least *some* similar
111 or overlapping content. The two videos we selected are introductory, about different primary
112 concepts, but also touch on “physics” and “astronomy” themes. We also wrote a set of multiple
113 choice quiz questions that would enable us to test participants’ knoweldge about each individual
114 video and about related content not specifically presented in either video (Tab. S1). Participants
115 answered questions randomly drawn from each content area (lecture 1, lecture 2, and general
116 physics knowledge) across each of three quizzes. Quiz 1 was intended to assessed participants’
117 knowledge before training; quiz 2 assessed knowledge after watching the Four Fundamental Forces
118 video (i.e., lecture 1); and quiz 3 assessed knowledge after watching the Birth of Stars video (i.e.,
119 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. This
123 yielded, for each second of each video, and for each question, a single topic vector– i.e., a coordinate
124 in a text embedding space (Fig. 6). Intuitively, each dimension of the embedding space corresponds
125 to a “theme” or “topic” reflected in some part(s) of the videos (Tab. S2), and the coordinates in
126 embedding space denote the blend of themes reflected by a particular excerpt of text (e.g., from

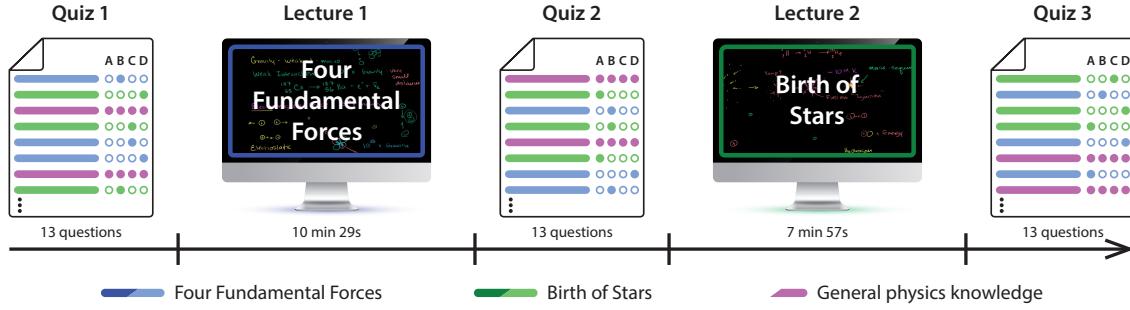


Figure 1: Experimental paradigm. Participants alternate between answering 13-question multiple choice quizzes and watching two Khan academy videos. Each quiz contains a mix of 5 questions about lecture 1, 5 questions about lecture 2, and 3 general physics knowledge questions. The specific questions reflected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

¹²⁷ part of a video’s transcript, from a question, etc.).

¹²⁸ Although a single lecture may be organized a single broad theme at a coarse scale, at a finer
¹²⁹ scale each moment of a lecture typically covers a narrower range of content. We wondered whether
¹³⁰ a text embedding model trained on the lectures’ transcripts might capture some of this finer scale
¹³¹ content. For example, if a particular question asks about the content from one small part of a
¹³² lecture, we wondered whether our text embedding model could be used to automatically identify
¹³³ the “matching” moment(s) in the lecture. When we correlated each question’s topic vector with
¹³⁴ the topic vectors for each second of the lectures, we found some evidence that each question is
¹³⁵ temporally specific (Fig. 2). In particular, most questions’ topic vectors were maximally correlated
¹³⁶ with a well-defined range of timepoints from their corresponding lectures, and the correlations fell
¹³⁷ off sharply outside of that range. We also examined the best-matching intervals for each question
¹³⁸ qualitatively by comparing the text of the question to the text of the most-correlated parts of the
¹³⁹ lectures. Despite that the questions were excluded from the text embedding model’s training set,
¹⁴⁰ in general we found a close correspondence between the conceptual content that each question
¹⁴¹ covered and the content covered by the best-matching moments of the lectures. Two representative
¹⁴² examples are shown at the bottom of Fig. 2.

¹⁴³ The ability to quantify how much each question is “asking about” each moment of the lectures
¹⁴⁴ could enable high-resolution insights into participants’ knowledge. Traditional approaches to

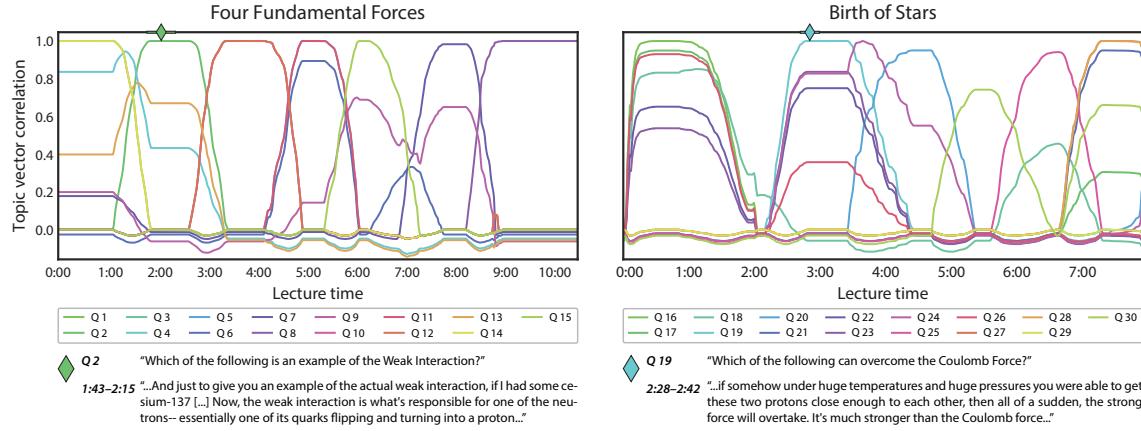


Figure 2: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question’s topic vector correlates with each video timepoint’s topic vector. The left panel displays these correlations for the *Four Fundamental Forces* lecture and associated questions, and the right panels displays these 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.

145 estimating how much a student “knows” about the content of a given lecture entail computing
 146 the proportion of correctly answered questions. But if two students receive identical scores on an
 147 exam, might our modeling framework help us to gain more nuanced insights into the *specific* content
 148 that each student has mastered (or failed to master)? For example, a student who misses three
 149 questions that were all about the same concept (e.g., concept *A*) will have gotten the same *proportion*
 150 of questions correct as another student who missed three questions about three *different* concepts
 151 (e.g., *A*, *B*, and *C*). But if we wanted to fill in the “gaps” in the two students’ understandings, we
 152 might do well to focus on concept *A* for the first student, but to also add in materials pertaining to
 153 concepts *B* and *C* for the second student.

154 We developed a simple formula (Eqn. 1) for using a participant’s responses to a small set
 155 of multiple choice questions to estimate how much the participant “knows” about the concept
 156 reflected by any arbitrary coordinate, x , in text embedding space (e.g., the content reflected by
 157 any moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially,
 158 the estimated knowledge at the coordinate is given by the weighted average proportion of quiz

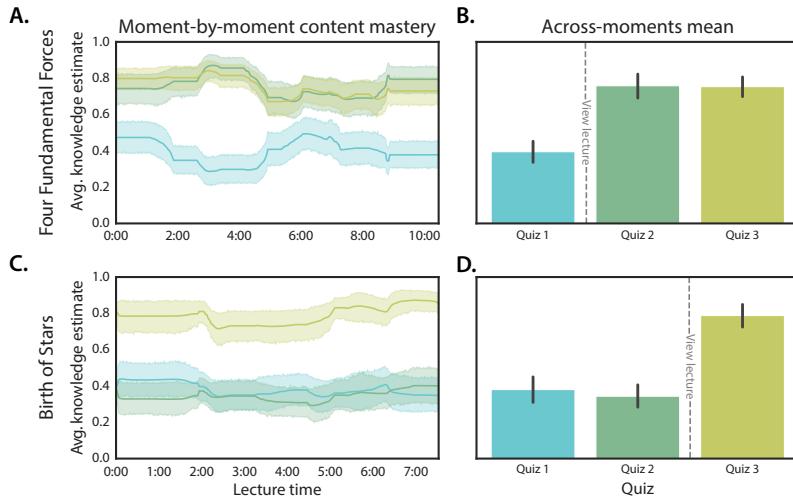


Figure 3: Estimating moment-by-moment knowledge acquisition. **A. Moment-by-moment knowledge about the *Four Fundamental Forces*.** Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (see *Estimating dynamic knowledge 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.

159 questions the participant answered correctly, where the weights reflect how much each question
 160 is “about” the content at x . When we apply this approach to estimate the participant’s knowledge
 161 about the content presented in each moment of each lecture, we can obtain a detailed timecourse
 162 describing how much “knowledge” the participant has about any part of the lecture. As shown in
 163 Figures 3A and C), we can also apply this approach separately for the questions from each quiz
 164 the participants took throughout the experiment. From just 13 questions per quiz, we obtain a
 165 high-resolution snapshot (at the time each quiz was taken) of what the participants knew about
 166 any moment’s content, from either of the two lectures they watched (comprising a total of 1106
 167 samples across the two lectures).

168 Of course, even though the timecourses in Figure 3A and C provide detailed *estimates* about
 169 participants’ knowledge, those estimates are only *useful* to the extent that they accurately reflect what

170 participants actually know. As one sanity check, we anticipated that the knowledge estimates
171 should show a content-specific “boost” in participants’ knowledge after watching each lecture.
172 In other words, if participants learn about each lecture’s content when they watch each lecture,
173 the knowledge estimates should reflect that. After watching the *Four Fundamental Forces* lecture,
174 participants should show more knowledge for the content of that lecture than they had before, and
175 that knowledge should persist for the remainder of the experiment. Specifically, knowledge about
176 that lecture’s content should be relatively low when estimated using Quiz 1 responses, but should
177 increase when estimated using Quiz 2 or 3 responses (Fig. 3B). Indeed, we found that participants’
178 estimated knowledge about the content of the *Four Fundamental Forces* was substantially higher on
179 Quiz 2 versus Quiz 1 ($t(XX) = XX, p = XX$) and on Quiz 3 versus Quiz 1 ($t(XX) = XX, p = XX$). We
180 found no reliable differences in estimated knowledge about that lecture’s content on Quiz 2 versus 3
181 ($t(XX) = XX, p = XX$). Similarly, we hypothesized (and subsequently confirmed) that participants
182 should show more estimated knowledge about the content of the *Birth of Stars* lecture after (versus
183 before) watching it (Fig. 3D). Specifically, since participants watched that lecture after taking Quiz
184 2 (but before Quiz 3), we hypothesized that their knowledge estimates should be relatively low on
185 Quizzes 1 and 2, but should show a “boost” on Quiz 3. Consistent with this prediction, we found
186 no reliable differences in estimated knowledge about the *Birth of Stars* lecture content on Quizzes
187 1 versus 2 ($t(XX) = XX, p = XX$), but the estimated knowledge was substantially higher on Quiz 3
188 versus 2 ($t(XX) = XX, p = XX$) and Quiz 3 versus 1 ($t(XX) = XX, p = XX$).

189 If we are able to accurately estimate a participant’s knowledge about the content tested by a
190 given question, the estimated knowledge should have some predictive information about whether
191 the participant is likely to answer the question correctly or incorrectly. For each question in turn,
192 for each participant, we used Equation 1 to estimate (using all *other* questions from the same quiz,
193 from the same participant) the participant’s knowledge at the held-out question’s embedding
194 coordinate. For each quiz, we aggregated these estimates into two distributions: one for the
195 estimated knowledge at the coordinates of each *correctly* answered question, and another for the
196 estimated knowledge at the coordinates of each *incorrectly* answered question (Fig. 4). We then used
197 independent samples t -tests to compare the means of these distributions of estimated knowledge.

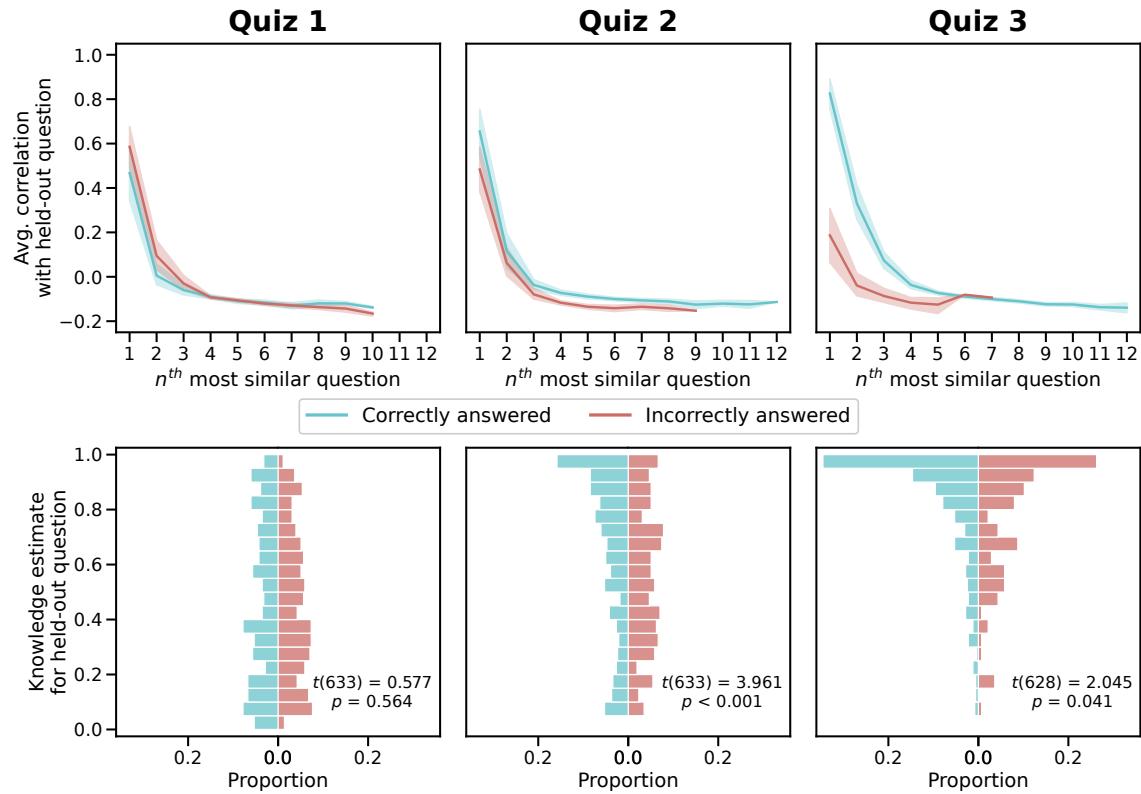


Figure 4: Estimating knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held out correctly (blue) or incorrectly (red) answered question. The t -tests reported in each panel are between the distributions of estimated knowledge at the coordinates of correctly versus incorrectly answered held-out questions.

198 For the initial quizzes participants took (prior to watching either lecture), participants' estimated
199 knowledge tended to be low overall, and relatively unstructured (Fig. 4, left panel). When we held
200 out individual questions and estimated their knowledge at the held-out questions' embedding
201 coordinates, we found no reliable differences in the estimates when the held-out question had
202 been correctly versus incorrectly answered ($t(633) = 0.577, p = 0.564$). After watching the first
203 video, estimated knowledge for held-out correctly answered questions (from the second quiz;
204 Fig. 4, middle panel) exhibited a positive shift relative to held-out incorrectly answered questions
205 ($t(633) = 3.961, p < 0.001$). After watching the second video, estimated knowledge (from the
206 third quiz; Fig. 4, right panel) for *all* questions exhibited a positive shift. However, the increase
207 in estimated knowledge for held-out correctly answered questions was larger than for held-out
208 incorrectly answered questions ($t(628) = 2.045, p = 0.041$).

209 Knowledge estimates need not be limited to the content of the lectures. As illustrated in
210 Figure 5, our general approach to estimating knowledge from a small number of quiz questions
211 may be applied to *any* content, given its text embedding coordinate. To visualize how knowledge
212 "spreads" through text embedding space to content beyond the lectures participants watched,
213 we first fit a new topic model to the lectures' sliding windows with $k = 100$ topics. We hoped
214 that increasing the number of topics from 15 to 100 might help us to generalize the knowledge
215 predictions. (Aside from increasing the number of topics from 15 to 100, all other procedures and
216 model parameters were carried over from the preceding analyses.) We resampled each lecture's
217 topic trajectory to 1 Hz and also projected each question into the same 100-dimensional topic space.

218 We projected the resulting 100-dimensional topic vectors (for each second of video and for each
219 question) into a shared 2-dimensional space (see *Creating knowledge and learning map visualizations*).
220 Next, we sampled points evenly from a 100×100 grid of coordinates that evenly tiled a rectangle
221 enclosing the 2D projections of the video trajectories and questions. We used Equation 4 to
222 estimate participants' knowledge at each of these 10K sampled locations, and we averaged these
223 estimates across participants to obtain an estimated average *knowledge map* (Fig. 5). Intuitively,
224 the knowledge map constructed from a given quiz's responses provides a visualization of how
225 "much" participants know about any content expressible by the fitted model.

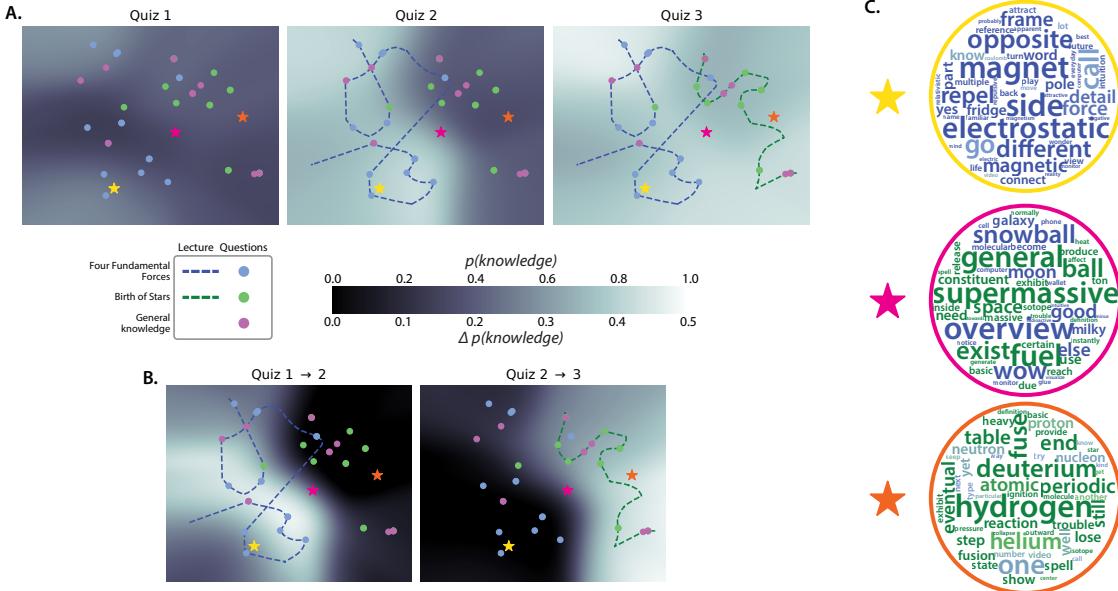


Figure 5: Mapping out the geometry of knowledge and learning. **A.** Average “knowledge maps” estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by all regions of topic space (see *Creating knowledge and learning map visualizations*). The 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.

Several features of the resulting knowledge maps are worth noting. The average knowledge map estimated from Quiz 1 responses (Fig. 5, leftmost map) shows that participants tended to have relatively little knowledge about any parts of the text embedding space (i.e., the shading is relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked increase in knowledge on the left side of the map (around roughly the same range of coordinates covered by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words, participants' estimated increase in knowledge is localized to conceptual content that is nearby (i.e., related to) the content from the lecture they watched prior to taking Quiz 2. This localization is non-trivial: the knowledge estimates are informed only by the embedded coordinates of the *quiz questions*, not by the embeddings of either lecture. Finally, the knowledge map estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region surrounding the embedding of the *Birth of Stars* lecture participants watched prior to taken Quiz 3.

Another way of visualizing these content-specific increases in knowledge (apparently driven by watching each lecture) is displayed in Figure 5B. Taking the point-by-point difference between the knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map* that describes the *change* in knowledge estimates from one quiz to the next. These learning maps highlight that the estimated knowledge increases we observed across maps were specific to the regions around the embeddings of each lecture in turn.

Because the 2D projection we used to construct the knowledge and learning maps is (partially) invertable, we may gain additional insights into the estimates by reconstructing the original high-dimensional topic vectors for any point(s) in the maps we are interested in. For example, this could serve as a useful tool for an instructor looking to better understand which content areas a student (or a group of students) knows well (or poorly). As a demonstration, we show the top-weighted words from the blends of topics reconstructed from three locations on the maps (Fig. 5C): one point near the *Four Fundamental Forces* embedding (yellow); a second point near the *Birth of Stars* embedding (orange), and a third point somewhere in between the two lectures' embeddings (pink). As shown in the word clouds in the Panel, the top-weighted words at the coordinate near the *Four Fundamental Forces* embedding also tended to be weighted heavily by the topics expressed in that

lecture. Similarly, the top-weighted words at the coordinate near the *Birth of Stars* embedding tended to be weighted most heavily by the topics expressed in *that* lecture. And the top-weighted words at the coordinate between the two lectures' embeddings show a roughly even mix of words most strongly associated with each lecture.

Discussion

Materials and methods

Participants

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

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

A total of 49 participants reporting having normal hearing and 1 participant reported having some hearing impairment. A total of 49 participants reported having normal color vision and 1 participant reported being color blind. Participants reported having had, on the night prior to testing, 2 – 4 hours of sleep (1 participant), 4 – 6 hours of sleep (9 participants), 6 – 8 hours of sleep

278 (35 participants), or 8+ hours of sleep (5 participants). They reported having consumed, on the
279 same day and leading up to their testing session, 0 cups of coffee (38 participants), 1 cup of coffee
280 (10 participants), 3 cups of coffee (1 participant), or 4+ cups of coffee (1 participant).

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

286 Participants reported their undergraduate major(s) as Social Sciences (28 participants), Natural
287 sciences (16), Professional (e.g., pre-med or pre-law; 8 participants), Mathematics and engineering
288 (7 participants), Humanities (4 participants), or Undecided (3 participants). Note that some par-
289 ticipants selected multiple categories for their undergraduate major. We also asked participants
290 about the courses they had taken. In total, 46 participants reported having taken at least one Khan
291 academy course in the past or being familiar with the Khan academy, and 4 reported not having
292 taken any Khan academy courses. Of the participants who reported having watched at least one
293 Khan academy course, 1 participant declined to report the number of courses they had watched;
294 7 participants reported having watched 1–2 courses; 11 reported having watched 3–5 courses; 8
295 reported having watched 5–10 courses; and 19 reported having watched 10 or more courses. We
296 also asked participants about the specific courses they had watched, categorized under different
297 subject areas. In the “Mathematics” area participants reported having watched videos on AP
298 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-
299 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry
300 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential
301 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),
302 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other
303 videos not listed in our survey (6 participants). In the “Science and engineering” area participants
304 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-
305 ipants); Physics, AP Physics I, or AP Physics II (15 participants); Biology, AP Biology; or High

306 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed in
307 our survey (20 participants). We also asked participants if they had specifically seen the videos
308 used in our experiment. When we asked about the *Four Fundamental Forces* video, 45 participants
309 reported not having watched it before, 1 participant reported that they were not sure if they had
310 watched it before, and 4 participants declined to respond. When we asked about the *Birth of*
311 *Stars* video, 46 participants reported not having watched it before and 4 participants declined to
312 respond. When we asked participants about non-Khan academy online courses, they reported
313 having watched or taken courses on Mathematics (15 participants), Science and engineering (11
314 participants), Test preparation (9 participants), Economics and finance (3 participants), Arts and
315 humanities (2 participants), Computing (2 participants), and other categories not listed in our
316 survey (18 participants). Finally, we asked participants about in-person courses they had taken in
317 different subject areas. They reported taking courses in Mathematics (39 participants), Science and
318 engineering (38 participants), Arts and humanities (35 participants), Test preparation (27 participants),
319 Economics and finance (26 participants), Computing (15 participants), College and careers
320 (7 participants), or other courses not listed in our survey (6 participants).

321 **Experiment**

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

330 Participants began the main experiment by answering a battery of 13 randomly selected ques-
331 tions (chosen from the full set of 39). Then they watched the *The Four Fundamental Forces* video.
332 Next, they answered a second set of 13 questions (chosen at random from the remaining 26 ques-

333 tions). Fourth, participants watch the *Birth of Stars* video, and finally they answered the remaining
334 13 questions. Our experimental procedure is diagramed in Figure 1. We used the experiment to
335 develop and test our computational framework for estimating knowledge and learning maps.

336 **Analysis**

337 **Constructing text embeddings of multiple videos and questions**

338 We extended an approach developed by [12] to construct text embeddings for each moment of each
339 lecture, and of each question in our pool. Briefly, our approach uses a topic model [3], trained on a
340 set of documents, to discover a set of k “topics” or “themes.” Formally, each topic is defined as a set
341 of weights over each word in the model’s vocabulary (i.e., the union of all unique words, across all
342 documents, excluding “stop words.”). Conceptually, each topic is intended to give larger weights
343 to set of words that appear conceptually related or that tend to co-occur in the same documents.
344 After fitting a topic model, each document in the training set, or any *new* document that contains at
345 least some of the words in the model’s vocabulary, may be represented as a k -dimensional vector
346 describing how much the document (most probably) reflects each topic. (Unless, otherwise noted,
347 we used $k = 15$ topics.)

348 As illustrated in Figure 6A, we start by building up a corpus of documents using overlapping
349 sliding windows that span each video’s transcript. Khan Academy videos are hosted on the
350 YouTube platform, and all YouTube videos are run through Google’s speech-to-text API [11] to
351 derive a timestamped transcript of any detected speech in the video. The resulting transcripts
352 contain one timestamped row per line, and each line generally corresponds to a few seconds of
353 spoken content from the video. We defined a sliding window length of (up to) $w = 30$ transcript
354 lines, and we assigned each window a timestamp according to the midpoint between its first
355 and last lines’ timestamps. These sliding windows ramped up and down in length at the very
356 beginning and end of the transcript, respectively. In other words, the first sliding window covered
357 only the first line from the transcript; the second sliding window covered the first two lines; and
358 so on. This insured that each line of the transcript appeared in the same number (w) of sliding

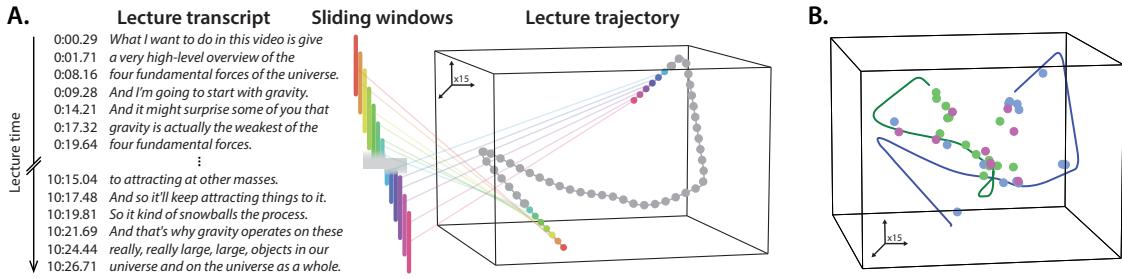


Figure 6: Constructing video content *trajectories*. A. Building a document pool from sliding windows of text. We decompose each video’s transcript into a series of overlapping sliding windows. The set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text 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; 19].

359 windows. We treated the text from each sliding window as a single “document,” and we combined
 360 these documents across the two videos’ windows to create a single training corpus for the topic
 361 model. The top words from each of the 15 discovered topics may be found in Table S2.

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

368 We transformed each sliding window’s text into a topic vector, and then used linear interpo-
 369 lation (independently for each topic dimension) to resample the resulting timeseries to once per
 370 second. This yielded a single topic vector for each second of each video. We also used the fitted
 371 model to obtain topic vectors for each question in our pool (Tab. S1). Taken together, we obtained
 372 a *trajectory* for each video, describing its path through topic space, and a single coordinate for each
 373 question (Fig. 6B). Embedding both videos and all of the questions using a common model enables

374 us to compare the content from different moments of videos, compare the content across videos,
375 and estimate potential associations between specific questions and specific moments of video.

376 **Estimating dynamic knowledge traces**

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

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

379 where

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

380 and where mincorr and maxcorr are the minimum and maximum correlations between any lecture
381 timepoint and question, taken over all timepoints and questions across both lectures and all three
382 question sets.

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

391 **Creating knowledge and learning map visualizations**

392 An important feature of our approach is that, given a trained text embedding model and partic-
393 ipants' quiz performance on each question, we can estimate their knowledge about *any* content
394 expressable by the embedding model– not solely the content explicitly probed by the quiz ques-
395 tions. To visualize these estimates (Figs. 5, S1, S2, S3, S4, and S5), we used UMAP [19] to define a

396 2D projection of the text embedding space. Sampling the original 100-dimensional space at high
 397 resolution to obtain an adequate set of topic vectors spanning the embedding space would be
 398 computationally intractable. However, sampling a 2D grid is much more feasible. We defined a
 399 rectangle enclosing the 2D projections of the lectures' and quizzes' embeddings, and we sampled
 400 points from a regular 100×100 grid of coordinates that evenly tiled the enclosing rectangle. We
 401 sought to estimate participants' knowledge (and learning— i.e., changes in knowledge) at each of
 402 the resulting 10000 coordinates.

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

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

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

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

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

414 References

- 415 [1] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
 416 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
 417 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.

- 418 [2] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*
419 *Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
420 Machinery.
- 421 [3] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
422 *Learning Research*, 3:993–1022.
- 423 [4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
424 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
425 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
426 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
427 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 428 [5] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
429 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 430 [6] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
431 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
432 sentence encoder. *arXiv*, 1803.11175.
- 433 [7] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
434 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 435 [8] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
436 Evidence for a new conceptualization of semantic representation in the left and right cerebral
437 hemispheres. *Cortex*, 40(3):467–478.
- 438 [9] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge. *School*
439 *Science and Mathematics*, 100(6):310–318.
- 440 [10] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
441 learning, pages 212–221. Sage Publications.

- 442 [11] Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml, M. (2016).
443 Contextual prediction models for speech recognition. In *Interspeech*, pages 2338–2342.
- 444 [12] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal be-
445 havioral and neural signatures of transforming naturalistic experiences into episodic memories.
446 *Nature Human Behavior*, 5:905–919.
- 447 [13] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
448 Columbia University Press.
- 449 [14] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 450 [15] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic
451 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
452 104:211–240.
- 453 [16] MacLellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of
454 Educational Studies*, 53(2):129–147.
- 455 [17] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
456 *Handbook of Human Memory*. Oxford University Press.
- 457 [18] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
458 function? *Psychological Review*, 128(4):711–725.
- 459 [19] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
460 projection for dimension reduction. *arXiv*, 1802(03426).
- 461 [20] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
462 tations in vector space. *arXiv*, 1301.3781.
- 463 [21] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
464 Student conceptions and conceptual learning in science. Routledge.

- ⁴⁶⁵ [22] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
⁴⁶⁶ conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in*
⁴⁶⁷ *Mathematics Education*, 35(5):305–329.