

¹ Text embedding models reveal high resolution insights
² into conceptual knowledge from short multiple choice
³ quizzes

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

¹Dartmouth College

²Akili Interactive

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

⁵ **Abstract**

⁶ We develop a mathematical framework, based on natural language processing models, for
⁷ tracking and characterizing the acquisition of conceptual knowledge. Our approach embeds each
⁸ concept in a high dimensional representation space, where nearby coordinates reflect similar or
⁹ related concepts. We test our approach using behavioral data collected from a group of college
¹⁰ students. In the experiment, we ask the participants to answer small sets of multiple choice quiz
¹¹ questions interleaved between watching two course videos from the Khan Academy platform.
¹² We applied our framework to the videos' transcripts, and to text of the quiz questions, to quantify
¹³ the content of each moment of video and each quiz question. We used these embeddings, along
¹⁴ with participants' quiz responses, to track how the learners' knowledge changed after watching
¹⁵ each video. Our findings show how a small set of quiz questions may be used to obtain rich and
¹⁶ meaningful high resolution insights into what each learner knows, and how their knowledge
¹⁷ changes over time as they learn.

¹⁸ **Keywords:** education, learning, knowledge, concepts, natural language processing

¹⁹ **Introduction**

²⁰ Suppose that a teacher had access to a complete “map” of everything their student knew. Defining
²¹ what such a map might even look like, let alone how it might be constructed or filled in, is itself
²² a non-trivial problem. But if a teacher *were* to gain access to such a map, how might that change
²³ their ability to teach the 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 any new material.
²⁸ Observing how the student’s knowledge was changing over time, in response to their training,
²⁹ could also help to guide the teacher.

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 to
³⁴ require something deeper and richer. Does conceptual understanding entail connecting newly
³⁵ acquired information to the scaffolding of one’s existing knowledge or experience [1, 5, 7, 8, 23]?
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 [15, 17, 22].

⁴⁰ The difference between “understanding” and “memorizing,” as framed by the researchers
⁴¹ in education, cognitive psychology, and cognitive neuroscience [10, 11, 14, 17, 22] 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
⁴⁵ a passage might be used to answer simple questions about the passage [e.g., whether it might

46 contain words related to furniture versus physics; 2, 3, 16]. However, modern natural language
47 processing models [e.g., 4, 6, 21] also attempt to capture the deeper meaning *underlying* those
48 atomic elements. These models consider not only the co-occurrences of those elements within
49 and across documents, but also patterns in how those elements appear across different scales (e.g.,
50 sentences, paragraphs, chapters, etc.), the temporal and grammatical properties of the elements,
51 and other high-level characteristics of how they are used [18, 19]. According to these models, the
52 deep 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 should also unlock any now-satisfied dependencies of that newly learned
67 concept so that they are “tagged” as available for future learning.

68 Here we develop a framework for modelling 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 has 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 start 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. To help
89 bridge this gap, our study uses course materials from real online courses to inform, fit, and test
90 models of real-world conceptual learning. We also provide a “proof of concept” demonstration
91 of how our models might be used to construct “maps” of what students know, and how their
92 knowledge changes with training. In addition to helping to visualize knowledge (and changes
93 in knowledge), we hope that such maps might lead to real-world tools for improving how we
94 educate.

95 Results

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

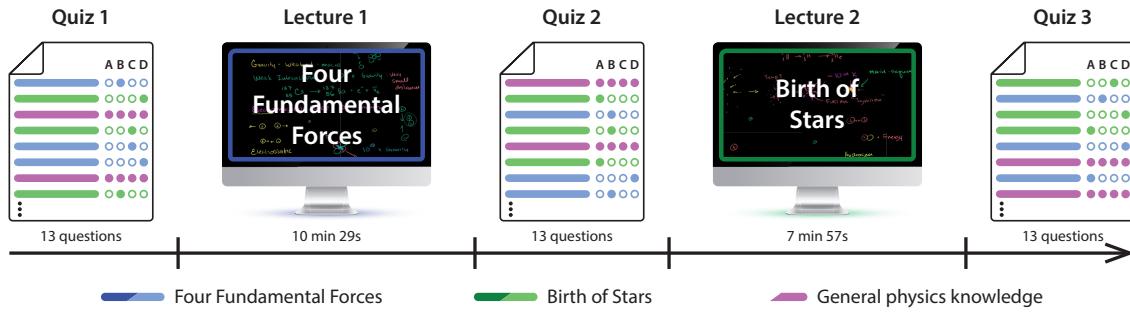


Figure 1: Experimental paradigm. Participants alternate between answering 13-question multiple choice quizzes and watching two Khan academy videos. Each quiz contains a mix of 5 questions about lecture 1, 5 questions about lecture 2, and 3 general physics knowledge questions. The specific questions reflected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

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

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

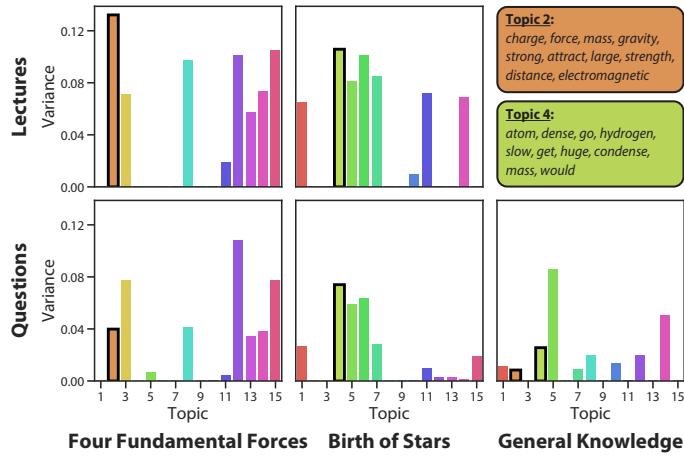


Figure 2: Lecture and question topic overlap. The bar plots display the variability in topic weights across lecture timepoints (top panels) and questions (bottom panels); colors denote topics. The top-weighted words from the most “expressive” (i.e., variable across observations) topics from each lecture are displayed in the upper right (orange: topic 2; yellow-green: topic 4). The top-weighted words from the full set of topics may be found in Table S2.

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

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

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

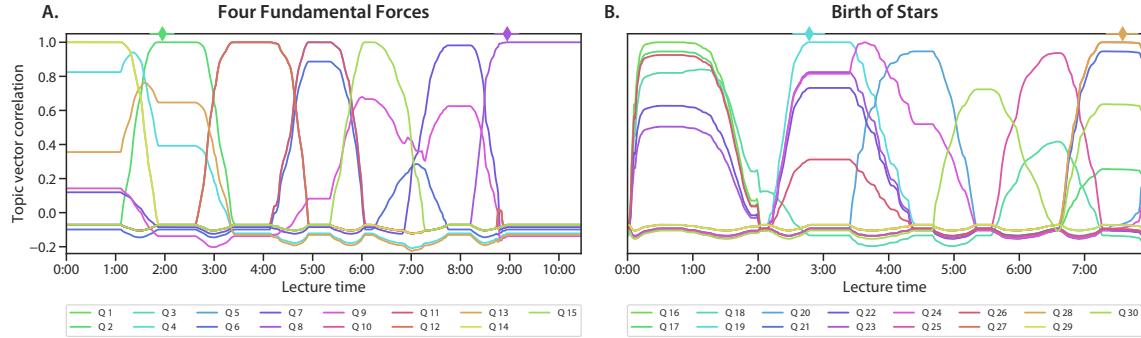


Figure 3: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel B.: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

as exact wording choices. If so, when we embed *new* text outside of the model’s training set, we should see a correspondance between the embeddings of the training data (i.e., snippets of text from the lectures’ transcripts) and other text that reflects related concepts (e.g., questions *about* each lecture). Further, although the content from any given moment from a lecture might stray from the average content (across all timepoints), we hoped that *variability* in each topic’s expression over timepoints within a lecture would match up with the variability in topic expressions for questions about that lecture. Intuitively, the variability in the expression of a given topic relates to how much “information” [9] the lecture (or questions) reflect about that topic. When we compared the variability in topic weights across each lecture’s timepoints with the variability in topic weights across each question set, we found a strong correspondence (Fig. —reffig:topics). The most variable topics from the *Four Fundamental Forces* lecture, and questions about that lecture, are 2, 3, 8, 12, 13, 14, and 15. The most variable topics from the **Birth of Stars** lecture, and questions about that lecture, are 1, 4, 5, 6, and 7. This strong overlap between the lectures and questions specifically about each lecture indicates that the topic model captures some of the underlying conceptual content.

Although a single lecture may be organized around a single broad theme at a coarse scale, at a

149 finer scale each moment of a lecture typically covers a narrower range of content. We wondered
150 whether a text embedding model trained on the lectures' transcripts might capture some of this
151 finer scale content. For example, if a particular question asks about the content from one small
152 part of a lecture, we wondered whether our text embedding model could be used to automatically
153 identify the "matching" moment(s) in the lecture. When we correlated each question's topic vector
154 with the topic vectors for each second of the lectures, we found some evidence that each question is
155 temporally specific (Fig. 3). In particular, most questions' topic vectors were maximally correlated
156 with a well-defined (and relatively narrow) range of timepoints from their corresponding lectures,
157 and the correlations fell off sharply outside of that range. We also examined the best-matching
158 intervals for each question qualitatively by comparing the text of the question to the text of the most-
159 correlated parts of the lectures. Despite that the questions were excluded from the text embedding
160 model's training set, in general we found (through manual inspection) a close correspondence
161 between the conceptual content that each question covered and the content covered by the best-
162 matching moments of the lectures. Two representative examples are shown at the bottom of
163 Fig. 3.

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

175 We developed a simple formula (Eqn. 1) for using a participant's responses to a small set
176 of multiple choice questions to estimate how much the participant "knows" about the concept

reflected by any arbitrary coordinate, x , in text embedding space (e.g., the content reflected by any moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially, the estimated knowledge at the coordinate is given by the weighted average proportion of quiz questions the participant answered correctly, where the weights reflect how much each question is “about” the content at x . When we apply this approach to estimate the participant’s knowledge about the content presented in each moment of each lecture, we can obtain a detailed timecourse describing how much “knowledge” the participant has about any part of the lecture. As shown in Figure 4, we can also apply this approach separately for the questions from each quiz the participants took throughout the experiment. From just 13 questions per quiz, we obtain a high-resolution snapshot (at the time each quiz was taken) of what the participants knew about any moment’s content, from either of the two lectures they watched (comprising a total of 1106 samples across the two lectures).

Of course, even though the timecourses in Figure 4A and C provide detailed *estimates* about participants’ knowlege, those estimates are only *useful* to the extent that they accurately reflect what participants actually know. As one sanity check, we anticipated that the knowledge estimates should show a content-specific “boost” in participants’ knowledge after watching each lecture. In other words, if participants learn about each lecture’s content when they watch each lecture, the knowledge estimates should reflect that. After watching the *Four Fundamental Forces* lecture, participants should show more knowledge for the content of that lecture than they had before, and that knowledge should persist for the remainder of the experiment. Specifically, knowledge about that lecture’s content should be relatively low when estimated using Quiz 1 responses, but should increase when estimated using Quiz 2 or 3 responses (Fig. 4B). Indeed, we found that participants’ estimated knowledge about the content of the *Four Fundamental Forces* was substantially higher on Quiz 2 versus Quiz 1 ($t(49) = 8.764, p < 0.001$) and on Quiz 3 versus Quiz 1 ($t(49) = 10.519, p < 0.001$). We found no reliable differences in estimated knowledge about that lecture’s content on Quiz 2 versus 3 ($t(49) = 0.160, p = 0.874$). Similarly, we hypothesized (and subsequently confirmed) that participants should show more estimated knowledge about the content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 4D). Specifically, since

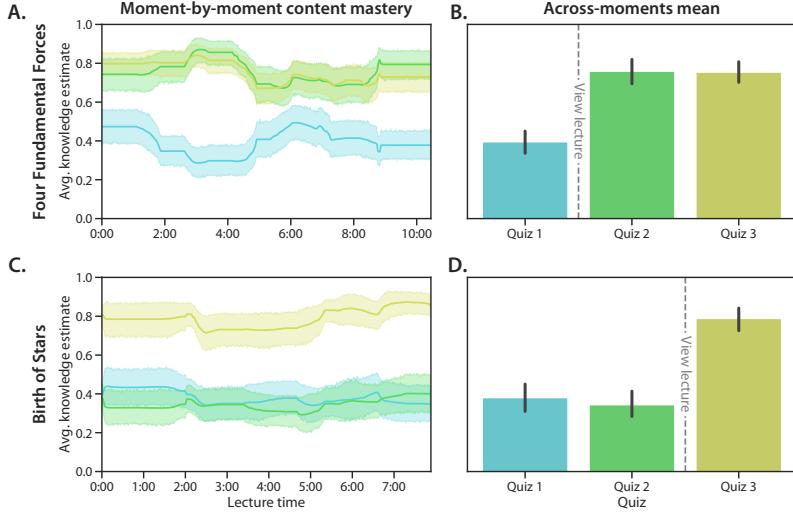


Figure 4: Estimating moment-by-moment knowledge acquisition. **A. Moment-by-moment knowledge about the *Four Fundamental Forces*.** Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (see *Estimating dynamic knowledge traces*), using responses from one quiz (color). The traces are averaged across participants. **B. Average estimated knowledge about the *Four Fundamental Forces*.** Each bar displays the across-timepoint average knowledge, estimated using the responses to one quiz's questions. **C. Moment-by-moment knowledge about the *Birth of Stars*.** The panel is in the same format as Panel A, but here the knowledge estimates are for the moment-by-moment content of the *Birth of Stars* lecture. **D. Average estimated knowledge about the *Birth of Stars*.** The panel is in the same format as Panel B, but here the knowledge estimates are for the content of the *Birth of Stars* lecture. All panels: error ribbons and error bars denote 95% confidence intervals, estimated across participants.

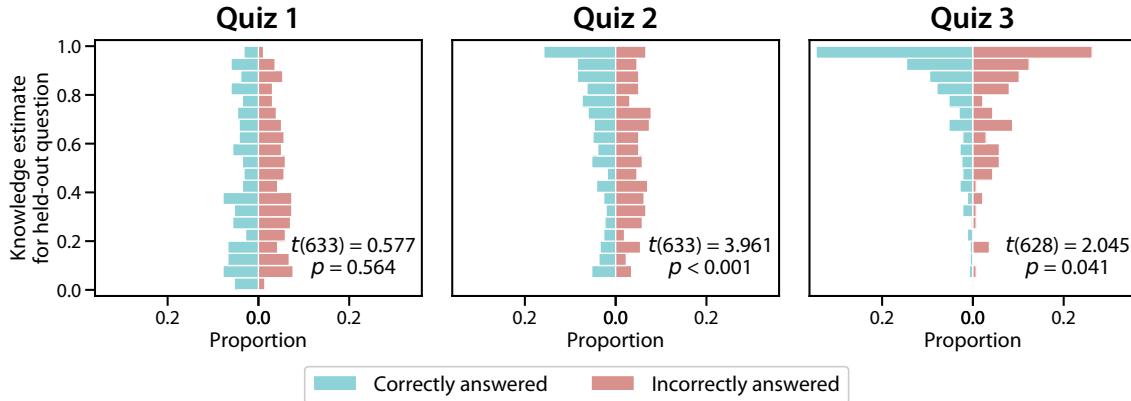


Figure 5: Estimating knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held out correctly (blue) or incorrectly (red) answered question. The t -tests reported in each panel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

205 participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their
 206 knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a “boost” on
 207 Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge
 208 about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ($t(49) = 1.013, p = 0.316$), but the
 209 estimated knowledge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and
 210 Quiz 3 versus 1 ($t(49) = 8.969, p < 0.001$).

211 If we are able to accurately estimate a participant’s knowledge about the content tested by a
 212 given question, the estimated knowledge should have some predictive information about whether
 213 the participant is likely to answer the question correctly or incorrectly. For each question in turn,
 214 for each participant, we used Equation 1 to estimate (using all *other* questions from the same quiz,
 215 from the same participant) the participant’s knowledge at the held-out question’s embedding
 216 coordinate. For each quiz, we aggregated these estimates into two distributions: one for the
 217 estimated knowledge at the coordinates of each *correctly* answered question, and another for the
 218 estimated knowledge at the coordinates of each *incorrectly* answered question (Fig. 5). We then used
 219 independent samples t -tests to compare the means of these distributions of estimated knowledge.

220 For the initial quizzes participants took (prior to watching either lecture), participants’ estimated

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

Knowledge estimates need not be limited to the content of the lectures. As illustrated in Figure 6, our general approach to estimating knowledge from a small number of quiz questions may be applied to *any* content, given its text embedding coordinate. To visualize how knowledge "spreads" through text embedding space to content beyond the lectures participants watched, we first fit a new topic model to the lectures' sliding windows with $k = 100$ topics. We hoped that increasing the number of topics from 15 to 100 might help us to generalize the knowledge predictions. (Aside from increasing the number of topics from 15 to 100, all other procedures and model parameters were carried over from the preceding analyses.) As in our other analyses, we resampled each lecture's topic trajectory to 1 Hz and also projected each question into a shared text embedding space.

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

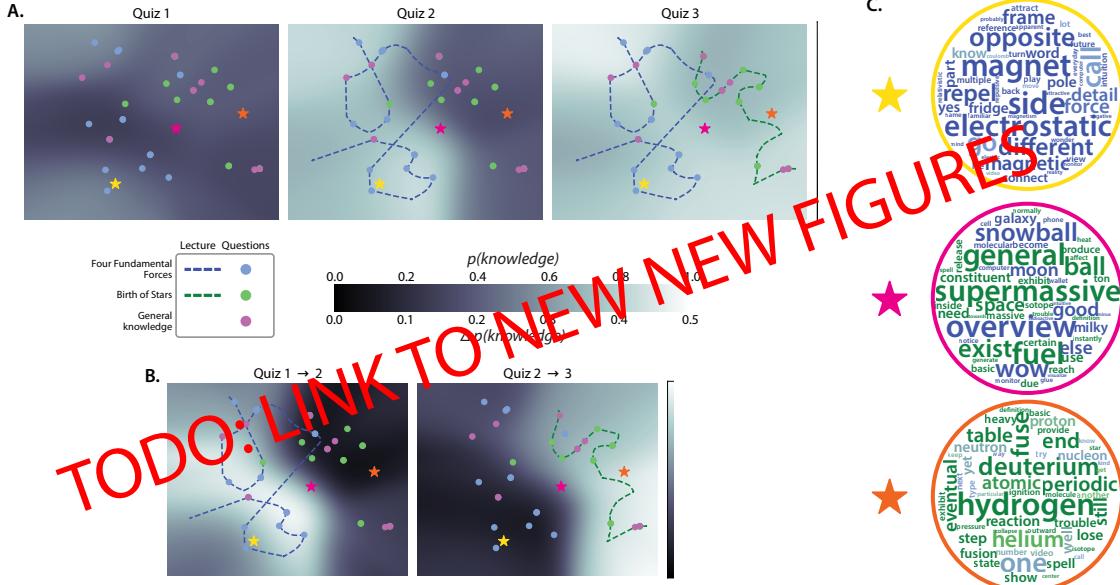


Figure 6: Mapping out the geometry of knowledge and learning. **A.** Average “knowledge maps” estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by all regions of topic space (see *Creating knowledge and learning map visualizations*). The topic trajectories of each lecture and the coordinates of each question are indicated by dotted lines and dots. Each map reflects an average across all participants. For individual participants’ maps, see Figures S1, S2, and S3. **B.** Average “learning maps” estimated between each successive pair of quizzes. The learning maps are in the same general format as the knowledge maps in Panel A, but each coordinate in the learning maps indicates the difference between the corresponding coordinates in the indicated pair of knowledge maps—i.e., how much the estimated knowledge “changed” across the two quizzes. Each map reflects an average across all participants. For individual participants’ maps, see Figures S4 and S5. **C.** Word clouds for sampled points in topic space. Each word cloud displays the relative weights of each word reflected by the blend of topics represented at the locations of the stars in the maps. The words’ colors indicate how much each word is weighted on average across all timepoints’ topic vectors in the *Four Fundamental Forces* (blue) and *Birth of Stars* (green) videos, respectively.

Several features of the resulting knowledge maps are worth noting. The average knowledge map estimated from Quiz 1 responses (Fig. 6, 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 (Eqn. 4). 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 taking Quiz 3.

Another way of visualizing these content-specific increases in knowledge (apparently driven by watching each lecture) is displayed in Figure 6B. 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 example locations on the maps (Fig. 6C): 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

277 at the example coordinate near the *Four Fundamental Forces* embedding also tended to be weighted
278 heavily by the topics expressed in that lecture. Similarly, the top-weighted words at the example
279 coordinate near the *Birth of Stars* embedding tended to be weighted most heavily by the topics
280 expressed in *that* lecture. And the top-weighted words at the example coordinate between the
281 two lectures' embeddings show a roughly even mix of words most strongly associated with each
282 lecture.

283 Discussion

284 Materials and methods

285 Participants

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

291 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
292 years). A total of 15 participants reported their gender as male and 35 participants reported their
293 gender as female. A total of 49 participants reported their native language as "English" and 1
294 reported having another native language. A total of 47 participants reported their ethnicity as
295 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants
296 reported their races as White (32 participants), Asian (14 participants), Black or African American
297 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
298 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

299 A total of 49 participants reporting having normal hearing and 1 participant reported having
300 some hearing impairment. A total of 49 participants reported having normal color vision and 1

301 participant reported being color blind. Participants reported having had, on the night prior to
302 testing, 2–4 hours of sleep (1 participant), 4–6 hours of sleep (9 participants), 6–8 hours of sleep (35
303 participants), or 8+ hours of sleep (5 participants). They reported having consumed, on the same
304 day and leading up to their testing session, 0 cups of coffee (38 participants), 1 cup of coffee (10
305 participants), 3 cups of coffee (1 participant), or 4+ cups of coffee (1 participant).

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

311 Participants reported their undergraduate major(s) as Social Sciences (28 participants), Natural
312 sciences (16), Professional (e.g., pre-med or pre-law; 8 participants), Mathematics and engineering
313 (7 participants), Humanities (4 participants), or Undecided (3 participants). Note that some par-
314 ticipants selected multiple categories for their undergraduate major. We also asked participants
315 about the courses they had taken. In total, 46 participants reported having taken at least one Khan
316 academy course in the past or being familiar with the Khan academy, and 4 reported not having
317 taken any Khan academy courses. Of the participants who reported having watched at least one
318 Khan academy course, 1 participant declined to report the number of courses they had watched;
319 7 participants reported having watched 1–2 courses; 11 reported having watched 3–5 courses; 8
320 reported having watched 5–10 courses; and 19 reported having watched 10 or more courses. We
321 also asked participants about the specific courses they had watched, categorized under different
322 subject areas. In the “Mathematics” area participants reported having watched videos on AP
323 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-
324 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry
325 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential
326 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),
327 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other
328 videos not listed in our survey (6 participants). In the “Science and engineering” area participants

329 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 participants);
330 Physics, AP Physics I, or AP Physics II (15 participants); Biology, AP Biology; or High
331 school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed in
332 our survey (20 participants). We also asked participants if they had specifically seen the videos
333 used in our experiment. When we asked about the *Four Fundamental Forces* video, 45 participants
334 reported not having watched it before, 1 participant reported that they were not sure if they had
335 watched it before, and 4 participants declined to respond. When we asked about the *Birth of*
336 *Stars* video, 46 participants reported not having watched it before and 4 participants declined to
337 respond. When we asked participants about non-Khan academy online courses, they reported
338 having watched or taken courses on Mathematics (15 participants), Science and engineering (11
339 participants), Test preparation (9 participants), Economics and finance (3 participants), Arts and
340 humanities (2 participants), Computing (2 participants), and other categories not listed in our
341 survey (18 participants). Finally, we asked participants about in-person courses they had taken in
342 different subject areas. They reported taking courses in Mathematics (39 participants), Science and
343 engineering (38 participants), Arts and humanities (35 participants), Test preparation (27 participants),
344 Economics and finance (26 participants), Computing (15 participants), College and careers
345 (7 participants), or other courses not listed in our survey (6 participants).

346 Experiment

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

355 Participants began the main experiment by answering a battery of 13 randomly selected ques-

356 tions (chosen from the full set of 39). Then they watched the *The Four Fundamental Forces* video.
357 Next, they answered a second set of 13 questions (chosen at random from the remaining 26 ques-
358 tions). Fourth, participants watch the *Birth of Stars* video, and finally they answered the remaining
359 13 questions. Our experimental procedure is diagramed in Figure 1. We used the experiment to
360 develop and test our computational framework for estimating knowledge and learning.

361 Analysis

362 Constructing text embeddings of multiple videos and questions

363 We extended an approach developed by [13] to construct text embeddings for each moment of
364 each lecture, and of each question in our pool. Briefly, our approach uses a topic model [3], trained
365 on a set of documents, to discover a set of k “topics” or “themes.” Formally, each topic is defined
366 as a set of weights over each word in the model’s vocabulary (i.e., the union of all unique words,
367 across all documents, excluding “stop words.”). Conceptually, each topic is intended to give larger
368 weights to words that are conceptually related or that tend to co-occur in the same documents.
369 After fitting a topic model, each document in the training set, or any *new* document that contains at
370 least some of the words in the model’s vocabulary, may be represented as a k -dimensional vector
371 describing how much the document (most probably) reflects each topic. (Unless, otherwise noted,
372 we used $k = 15$ topics.)

373 As illustrated in Figure 7A, we start by building up a corpus of documents using overlapping
374 sliding windows that span each video’s transcript. Khan Academy videos are hosted on the
375 YouTube platform, and all YouTube videos are run through Google’s speech-to-text API [12] to
376 derive a timestamped transcript of any detected speech in the video. The resulting transcripts
377 contain one timestamped row per line, and each line generally corresponds to a few seconds of
378 spoken content from the video. We defined a sliding window length of (up to) $w = 30$ transcript
379 lines, and we assigned each window a timestamp according to the midpoint between its first
380 and last lines’ timestamps. These sliding windows ramped up and down in length at the very
381 beginning and end of the transcript, respectively. In other words, the first sliding window covered

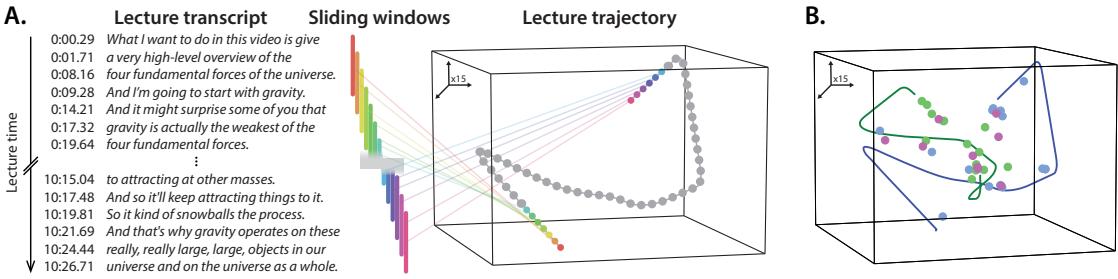


Figure 7: Constructing video content *trajectories*. **A. Building a document pool from sliding windows of text.** We decompose each video’s transcript into a series of overlapping sliding windows. The set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding model. After training a text embedding model using the two videos’ sliding windows, along with the text from each question in our pool (Tab. S1), we construct “trajectories” through text embedding space by joining the embedding coordinates of successive sliding windows from each video. **B. Embedding multiple videos and questions.** Applying the same text embedding approach to each video, along with the text of each question, results in one trajectory per video and one embedding coordinate (dot) per question (blue: *Four Fundamental Forces*; green: *Birth of Stars*; pink: general physics knowledge). Here we have projected the 15-dimensional embeddings into a 3D space using Uniform Manifold Approximation and Projection [UMAP; 20].

only the first line from the transcript; the second sliding window covered the first two lines; and so on. This insured that each line of the transcript appeared in the same number (w) of sliding windows. We treated the text from each sliding window as a single “document,” and we combined these documents across the two videos’ windows to create a single training corpus for the topic model. The top words from each of the 15 discovered topics may be found in Table S2.

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

We transformed each sliding window’s text into a topic vector, and then used linear interpolation (independently for each topic dimension) to resample the resulting timeseries to once per second. This yielded a single topic vector for each second of each video. We also used the fitted model to obtain topic vectors for each question in our pool (Tab. S1). Taken together, we obtained

397 a *trajectory* for each video, describing its path through topic space, and a single coordinate for each
 398 question (Fig. 7B). Embedding both videos and all of the questions using a common model enables
 399 us to compare the content from different moments of videos, compare the content across videos,
 400 and estimate potential associations between specific questions and specific moments of video.

401 **Estimating dynamic knowledge traces**

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

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

404 where

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

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

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

419 **Creating knowledge and learning map visualizations**

420 An important feature of our approach is that, given a trained text embedding model and partic-
421 ipants’ quiz performance on each question, we can estimate their knowledge about *any* content
422 expressable by the embedding model— not solely the content explicitly probed by the quiz ques-
423 tions. To visualize these estimates (Figs. 6, S1, S2, S3, S4, and S5), we used UMAP [20] to define a
424 2D projection of the text embedding space. Sampling the original 100-dimensional space at high
425 resolution to obtain an adequate set of topic vectors spanning the embedding space would be
426 computationally intractable. However, sampling a 2D grid is much more feasible. We defined a
427 rectangle enclosing the 2D projections of the lectures’ and quizzes’ embeddings, and we sampled
428 points from a regular 100×100 grid of coordinates that evenly tiled the enclosing rectangle. We
429 sought to estimate participants’ knoweldge (and learning—i.e., changes in knowledge) at each of
430 the resulting 10000 coordinates.

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

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

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

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

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

⁴⁴² **References**

- ⁴⁴³ [1] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in children: distinguishing response flexibility from conceptual flexibility; the protracted development of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.
- ⁴⁴⁶ [2] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing Machinery.
- ⁴⁴⁹ [3] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- ⁴⁵¹ [4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- ⁴⁵⁶ [5] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- ⁴⁵⁸ [6] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *arXiv*, 1803.11175.
- ⁴⁶¹ [7] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- ⁴⁶³ [8] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004). Evidence for a new conceptualization of semantic representation in the left and right cerebral hemispheres. *Cortex*, 40(3):467–478.

- 466 [9] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*
467 *Transactions of the Royal Society A*, 222(602):309–368.
- 468 [10] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
469 *School Science and Mathematics*, 100(6):310–318.
- 470 [11] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
471 learning, pages 212–221. Sage Publications.
- 472 [12] Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml, M. (2016).
473 Contextual prediction models for speech recognition. In *Interspeech*, pages 2338–2342.
- 474 [13] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal be-
475 havioral and neural signatures of transforming naturalistic experiences into episodic memories.
476 *Nature Human Behavior*, 5:905–919.
- 477 [14] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
478 Columbia University Press.
- 479 [15] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 480 [16] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic
481 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
482 104:211–240.
- 483 [17] Macellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of*
484 *Educational Studies*, 53(2):129–147.
- 485 [18] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
486 *Handbook of Human Memory*. Oxford University Press.
- 487 [19] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
488 function? *Psychological Review*, 128(4):711–725.

- 489 [20] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
490 projection for dimension reduction. *arXiv*, 1802(03426).
- 491 [21] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
492 tations in vector space. *arXiv*, 1301.3781.
- 493 [22] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
494 Student conceptions and conceptual learning in science. Routledge.
- 495 [23] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
496 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in
497 Mathematics Education*, 35(5):305–329.