

- 1 Text embedding models yield high-resolution insights
- 2 into conceptual knowledge from short multiple-choice

3 quizzes

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

¹Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

²Akili Interactive Labs

Boston, MA 02110, USA

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

Abstract

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

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

¹⁹ **Introduction**

²⁰ Suppose that a teacher had access to a complete, tangible “map” of everything a student knows.
²¹ Defining what such a map might even look like, let alone how it might be constructed or filled in, is
²² itself a non-trivial problem. But if a teacher *were* to gain access to such a map, how might it change
²³ their ability to teach that student? Perhaps they might start by checking how well the student
²⁴ knows the to-be-learned information already, or how much they know about related concepts.
²⁵ For some students, they could potentially optimize their teaching efforts to maximize efficiency
²⁶ by focusing primarily on not-yet-known content. For other students (or other content areas), it
²⁷ might be more effective to optimize for direct connections between already known content and
²⁸ new material. Observing how the student’s knowledge changed over time, in response to their
²⁹ teaching, could also help to guide the teacher towards the most effective strategy for that individual
³⁰ student.

³¹ A common approach to assessing a student’s knowledge is to present them with a set of quiz
³² questions, calculate the proportion they answer correctly, and provide them with feedback in the
³³ form of a simple numeric or letter grade. While such a grade can provide *some* indication of whether
³⁴ the student has mastered the to-be-learned material, any univariate measure of performance on a
³⁵ complex task sacrifices certain relevant information, risks conflating underlying factors, and so on.
³⁶ For example, consider the relative utility of the theoretical map described above that characterizes
³⁷ a student’s knowledge in detail, versus a single annotation saying that the student answered 85%
³⁸ of their quiz questions correctly, or that they received a ‘B’. Here we show that the same quiz data
³⁹ required to compute proportion-correct scores or letter grades can instead be used to obtain far
⁴⁰ more detailed insights into what a student knew at the time they took the quiz.

⁴¹ Designing and building procedures and tools for mapping out knowledge touches on deep
⁴² questions about what it means to learn. For example, how do we acquire conceptual knowledge?
⁴³ Memorizing course lectures or textbook chapters by rote can lead to the superficial *appearance*
⁴⁴ of understanding the underlying content, but achieving true conceptual understanding seems to
⁴⁵ require something deeper and richer. Does conceptual understanding entail connecting newly

46 acquired information to the scaffolding of one’s existing knowledge or experience [6, 11, 13, 15, 30,
47 65]? Or weaving a lecture’s atomic elements (e.g., its component words) into a structured network
48 that describes how those individual elements are related [40, 70]? Conceptual understanding
49 could also involve building a mental model that transcends the meanings of those individual
50 atomic elements by reflecting the deeper meaning underlying the gestalt whole [37, 41, 62, 69].

51 The difference between “understanding” and “memorizing,” as framed by researchers in ed-
52 ucation, cognitive psychology, and cognitive neuroscience [e.g., 23, 28, 33, 41, 62], has profound
53 analogs in the fields of natural language processing and natural language understanding. For
54 example, considering the raw contents of a document (e.g., its constituent symbols, letters, and
55 words) might provide some clues as to what the document is about, just as memorizing a passage
56 might provide some ability to answer simple questions about it. However, text embedding mod-
57 els [e.g., 7, 8, 10, 12, 16, 39, 51, 71] also attempt to capture the deeper meaning *underlying* those
58 atomic elements. These models consider not only the co-occurrences of those elements within and
59 across documents, but (in many cases) also patterns in how those elements appear across different
60 scales (e.g., sentences, paragraphs, chapters, etc.), their temporal and grammatical properties, and
61 other high-level characteristics of how they are used [42, 43]. To be clear, this is not to say that text
62 embedding models themselves are capable of “understanding” deep conceptual meaning in any
63 traditional sense. But rather, their ability to capture the underlying *structure* of text documents
64 beyond their surface-level contents provides a computational framework through which those
65 documents’ deeper conceptual meanings may be quantified, explored, and understood. Accord-
66 ing to these models, the deep conceptual meaning of a document may be captured by a feature
67 vector in a high-dimensional representation space, wherein nearby vectors reflect conceptually
68 related documents. A model that succeeds at capturing an analogue of “understanding” is able
69 to assign nearby feature vectors to two conceptually related documents *even when the specific words*
70 *contained in those documents have limited overlap*. In this way, “concepts” are defined implicitly by
71 the model’s geometry [e.g., how the embedding coordinate of a given word or document relates
72 to the coordinates of other text embeddings; 56].

73 Given these insights, what form might a representation of the sum total of a person’s knowledge

74 take? First, we might require a means of systematically describing or representing (at least some
75 subset of) the nearly infinite set of possible things a person could know. Second, we might want to
76 account for potential associations between different concepts. For example, the concepts of “fish”
77 and “water” might be associated in the sense that fish live in water. Third, knowledge may have
78 a critical dependency structure, such that knowing about a particular concept might require first
79 knowing about a set of other concepts. For example, understanding the concept of a fish swimming
80 in water first requires understanding what fish and water *are*. Fourth, as we learn, our “current
81 state of knowledge” should change accordingly. Learning new concepts should both update our
82 characterizations of “what is known” and also unlock any now-satisfied dependencies of those
83 newly learned concepts so that they are “tagged” as available for future learning.

84 Here we develop a framework for modeling how conceptual knowledge is acquired during
85 learning. The central idea behind our framework is to use text embedding models to define the
86 coordinate systems of two maps: a *knowledge map* that describes the extent to which each concept is
87 currently known, and a *learning map* that describes changes in knowledge over time. Each location
88 on these maps represents a single concept, and the maps’ geometries are defined such that related
89 concepts are located nearby in space. We use this framework to analyze and interpret behavioral
90 data collected from an experiment that had participants answer sets of multiple-choice questions
91 about a series of recorded course lectures.

92 Our primary research goal is to advance our understanding of what it means to acquire deep,
93 real-world conceptual knowledge. Traditional laboratory approaches to studying learning and
94 memory (e.g., list-learning studies) often draw little distinction between memorization and under-
95 standing. Instead, these studies typically focus on whether information is effectively encoded or
96 retrieved, rather than whether the information is *understood*. Approaches to studying conceptual
97 learning, such as category learning experiments, can begin to investigate the distinction between
98 memorization and understanding, often by training participants to distinguish arbitrary or random
99 features in otherwise meaningless categorized stimuli [1, 20, 21, 24, 31, 59]. However, the objective
100 of real-world training, or learning from life experiences more generally, is often to develop new
101 knowledge that may be applied in *useful* ways in the future. In this sense, the gap between modern

learning theories and modern pedagogical approaches that inform classroom learning strategies is enormous: most of our theories about *how* people learn are inspired by experimental paradigms and models that have only peripheral relevance to the kinds of learning that students and teachers actually seek [28, 41]. To help bridge this gap, our study uses course materials from real online courses to inform, fit, and test models of real-world conceptual learning. We show that these models recover meaningful relationships between concepts presented during course lectures and tested by assessments, and that these relationships can be leveraged to predict students' success on individual quiz questions. We also provide a demonstration of how our models can be used to construct "maps" of what students know, and how their knowledge changes with training. In addition to helping to visually capture knowledge (and changes in knowledge), we hope that such maps might lead to real-world tools for improving how we educate. Taken together, our work shows that existing course materials and evaluative tools like short multiple-choice quizzes may be leveraged to gain highly detailed insights into what students know and how they learn.

Results

At its core, our main modeling approach is based around a simple assumption that we sought to test empirically: all else being equal, knowledge about a given concept is predictive of knowledge about similar or related concepts. From a geometric perspective, this assumption implies that knowledge is fundamentally "smooth." In other words, as one moves through a space representing an individual's knowledge (where similar concepts occupy nearby coordinates), their "level of knowledge" should change relatively gradually. To begin to test this smoothness assumption, we sought to track participants' knowledge and how it changed over time in response to training. Two overarching goals guide our approach. First, we want to gain detailed insights into what learners know at different points in their training. For example, rather than simply reporting on the proportions of questions participants answer correctly (i.e., their overall performance), we seek estimates of their knowledge about a variety of specific concepts. Second, we want our approach to be potentially scalable to large numbers of diverse concepts, courses, and students. This requires

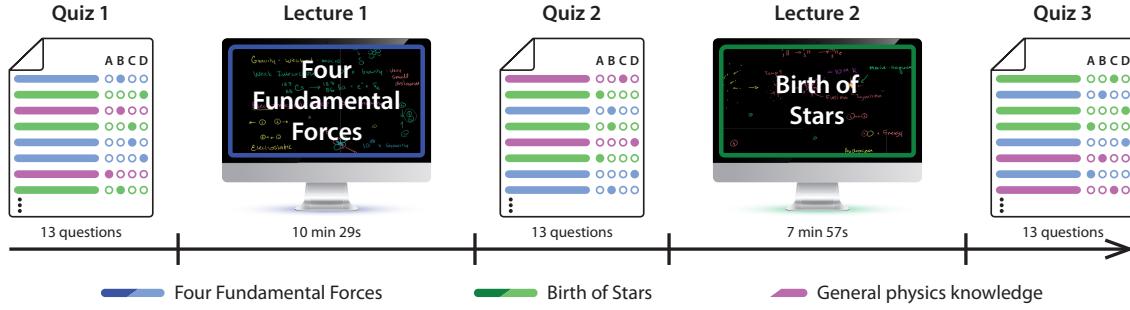


Figure 1: Experimental paradigm. Participants alternate between completing three 13-question multiple-choice quizzes and watching two Khan Academy lectures. Each quiz contains a mix of 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general physics knowledge. The specific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

128 that the conceptual content of interest be discovered *automatically*, rather than relying on manually
 129 produced ratings or labels.

130 We asked participants in our study to complete brief multiple-choice quizzes before, between,
 131 and after watching two lecture videos from the Khan Academy [36] platform (Fig. 1). The first
 132 lecture video, entitled *Four Fundamental Forces*, discussed the four fundamental forces in physics:
 133 gravity, strong and weak interactions, and electromagnetism. The second, entitled *Birth of Stars*,
 134 provided an overview of our current understanding of how stars form. We selected these particular
 135 lectures to satisfy three general criteria. First, we wanted both lectures to be accessible to a broad
 136 audience (i.e., with minimal prerequisite knowledge) so as to limit the impact of prior training
 137 on participants’ abilities to learn from the lectures. To this end, we selected two introductory
 138 videos that were intended to be viewed at the start of students’ training in their respective content
 139 areas. Second, we wanted the two lectures to have some related content so that we could test
 140 our approach’s ability to distinguish similar conceptual content. To this end, we chose two videos
 141 from the same Khan Academy course domain, “Cosmology and Astronomy.” Third, we sought to
 142 minimize dependencies and specific overlap between the videos. For example, we did not want
 143 participants’ abilities to understand one video to (directly) influence their abilities to understand the
 144 other. To satisfy this last criterion, we chose videos from two different lecture series (Lectures 1 and
 145 2 were from the “Scale of the Universe” and “Stars, Black Holes, and Galaxies” series, respectively).



Figure 2: Modeling course content. **A. Building a document pool from sliding windows of text.** We decompose each lecture’s transcript into a series of overlapping sliding windows. The full set of transcript snippets (across all windows) may be treated as a set of “documents” for training a text embedding model. **B. Constructing lecture content trajectories.** After training the model on the sliding windows from both lectures, we transform each lecture into a “trajectory” through text embedding space by joining the embedding coordinates of successive sliding windows parsed from its transcript. **C. Embedding multiple lectures and questions in a shared space.** We apply the same model (trained on the two lectures’ windows) to both lectures, along with the text of each question in our pool (Supp. Tab. 1), to project them into a shared text embedding space. This results in one trajectory per lecture and one coordinate for each question. Here, we have projected the 15-dimensional embeddings onto their first 3 principal components for visualization.

146 We also wrote a set of multiple-choice quiz questions that we hoped would enable us to
 147 evaluate participants’ knowledge about each individual lecture, along with related knowledge
 148 about physics concepts not specifically presented in either video (see Supp. Tab. 1 for the full list
 149 of questions in our stimulus pool). Participants answered questions randomly drawn from each
 150 content area (Lecture 1, Lecture 2, and general physics knowledge) on each of the three quizzes.
 151 Quiz 1 was intended to assess participants’ “baseline” knowledge before training, Quiz 2 assessed
 152 knowledge after watching the *Four Fundamental Forces* video (i.e., Lecture 1), and Quiz 3 assessed
 153 knowledge after watching the *Birth of Stars* video (i.e., Lecture 2).

154 To study in detail how participants’ conceptual knowledge changed over the course of the
 155 experiment, we first sought to model the conceptual content presented to them at each moment
 156 throughout each of the two lectures. We adapted an approach we developed in prior work [29]
 157 to identify the latent themes in the lectures using a topic model [8]. Briefly, topic models take
 158 as input a collection of text documents, and learn a set of “topics” (i.e., latent themes) from their
 159 contents. Once fit, a topic model can be used to transform arbitrary (potentially new) documents
 160 into sets of “topic proportions” describing the weighted blend of learned topics reflected in their

161 texts. We parsed automatically generated transcripts of the two lectures into overlapping sliding
162 windows, where each window contained the text of the lecture transcript from a particular time
163 span. We treated the set of text snippets (across all of these windows) as documents to fit the model
164 (Fig. 2A; see *Constructing text embeddings of multiple lectures and questions*). Transforming the text
165 from every sliding window with the model yielded a number-of-windows by number-of-topics
166 (15) topic-proportions matrix describing the unique mixture of broad themes from both lectures
167 reflected in each window’s text. Each window’s “topic vector” (i.e., column of the topic-proportions
168 matrix) is analogous to a coordinate in a 15-dimensional space whose axes are topics discovered
169 by the model. Within this space, each lecture’s sequence of topic vectors (i.e., corresponding to its
170 transcript’s overlapping text snippets across sliding windows) forms a *trajectory* that captures how
171 its conceptual content unfolds over time (Fig. 2B). We resampled these trajectories to a resolution
172 of one topic vector for each second of video (i.e., 1 Hz).

173 We hypothesized that a topic model trained on transcripts of the two lectures should also
174 capture the conceptual knowledge probed by each quiz question. If indeed the topic model could
175 capture information about the deeper conceptual content of the lectures (i.e., beyond surface-level
176 details such as particular word choices), then we should be able to recover a correspondence
177 between each lecture and questions *about* each lecture. Importantly, such a correspondence could
178 not arise solely from superficial text matching between lecture transcripts and questions, since
179 the lectures and questions often used different words (Supp. Fig. 10) and phrasings. Simply
180 comparing the average topic weights from each lecture and question set (averaging across time
181 and questions, respectively) reveals a striking correspondence (Supp. Fig. 2). Specifically, the
182 average topic weights from Lecture 1 are strongly correlated with the average topic weights from
183 questions about Lecture 1 ($r(13) = 0.809$, $p < 0.001$, 95% confidence interval (CI) = [0.633, 0.962]),
184 and the average topic weights from Lecture 2 are strongly correlated with the average topic weights
185 from questions about Lecture 2 ($r(13) = 0.728$, $p = 0.002$, 95% CI = [0.456, 0.920]). At the same
186 time, the average topic weights from the two lectures are *negatively* correlated with the average
187 topic weights from their non-matching question sets (Lecture 1 video vs. Lecture 2 questions:
188 $r(13) = -0.547$, $p = 0.035$, 95% CI = [-0.812, -0.231]; Lecture 2 video vs. Lecture 1 questions:



Figure 3: Lecture and question topic overlap. A. Topic weight variability. The bar plots display the variance in each topic's weight across lecture timepoints (top row) and questions (bottom row); colors denote topics. The top-weighted words from the most “expressive” (i.e., variable across observations) topic from each lecture are displayed in the upper right (orange: topic 2; yellow-green: topic 4). The top-weighted words from the full set of topics may be found in Supplementary Table 2. **B. Relationships between topic weight variability.** Pairwise correlations between the distributions of topic weight variance for each lecture and question set. Each row and column corresponds to a bar plot in Panel A.

189 $r(13) = -0.612, p = 0.015, 95\% \text{ CI} = [-0.874, -0.281]$), indicating that the topic model also exhibits
190 some degree of specificity. The full set of pairwise comparisons between average topic weights for
191 the lectures and question sets is reported in Supplementary Figure 2.

192 Another, more sensitive, way of summarizing the conceptual content of the lectures and ques-
193 tions is to look at *variability* in how topics are weighted over time and across different questions
194 (Fig. 3). Intuitively, the variability in the expression of a given topic relates to how much “infor-
195 mation” [22] the lecture (or question set) reflects about that topic. For example, suppose a given topic
196 is weighted on heavily throughout a lecture. That topic might be characteristic of some aspect or
197 property of the lecture *overall* (conceptual or otherwise), but unless the topic’s weights change in
198 meaningful ways over time, it would be a poor indicator of any *specific* conceptual content in the
199 lecture. We therefore also compared the variances in topic weights (over time and across questions)
200 between the lectures and questions. The variability in topic expression was similar for the Lecture 1
201 video and questions ($r(13) = 0.824, p < 0.001, 95\% \text{ CI} = [0.696, 0.973]$), and for the Lecture 2 video
202 and questions ($r(13) = 0.801, p < 0.001, 95\% \text{ CI} = [0.539, 0.958]$). Simultaneously, as reported

203 in Figure 3B, the variabilities in topic expression across *different* videos and lecture-specific ques-
204 tions (i.e., Lecture 1 video vs. Lecture 2 questions; Lecture 2 video vs. Lecture 1 questions) were
205 negatively correlated, and neither video’s topic variability was reliably correlated with the topic
206 variability across general physics knowledge questions. Taken together, the analyses reported in
207 Figure 3 and Supplementary Figure 2 indicate that a topic model fit to the videos’ transcripts can
208 also reveal correspondences (at a coarse scale) between the lectures and questions.

209 An individual lecture may be organized around a single broad theme at a coarse scale, but at
210 a finer scale, each moment of a lecture typically covers a narrower range of content. Given the
211 correspondence we found between the variabilities in topic expression across moments of each
212 lecture and questions from its corresponding set (Fig. 3), we wondered whether the text embedding
213 model might additionally capture these conceptual relationships at a finer scale. For example, if a
214 particular question asks about the content from one small part of a lecture, we wondered whether
215 the text embeddings could be used to automatically identify the “matching” moment(s) in the
216 lecture. To explore this, we computed the correlation between each question’s topic weights
217 and the topic weights for each second of its corresponding lecture, and found that each question
218 appeared to be temporally specific (Fig. 4). In particular, most questions’ topic vectors were
219 maximally correlated with a well-defined (and relatively narrow) range of timepoints from their
220 corresponding lectures, outside of which the correlations fell off sharply (Supp. Figs. 3, 4). We also
221 qualitatively examined the best-matching intervals for each question by comparing the questions’
222 text to the transcribed text from the most-correlated parts of the lectures (Supp. Tab. 3). Despite
223 that the questions were excluded from the text embedding model’s training set, in general we
224 found (through manual inspection) a close correspondence between the conceptual content that
225 each question probed and the content covered by the best-matching moments of the lectures. Two
226 representative examples are shown at the bottom of Figure 4.

227 The ability to quantify how much each question is “asking about” the content from each moment
228 of the lectures could enable high-resolution insights into participants’ knowledge. Traditional
229 approaches to estimating how much a student “knows” about the content of a given lecture
230 entail administering some form of assessment (e.g., a quiz) and computing the proportion of

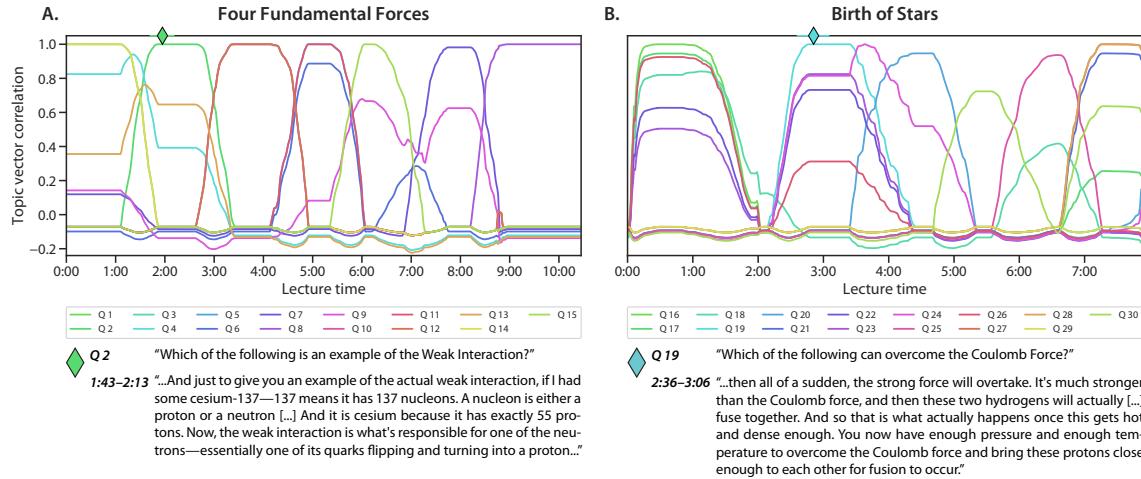


Figure 4: Which parts of each lecture are captured by each question? Each panel displays time series plots showing how each question’s topic vector correlates with each video timepoint’s topic vector (Panel A.: correlations for the *Four Fundamental Forces* lecture and associated questions; Panel B.: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

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

We developed a simple formula (Eqn. 1) for using a participant’s responses to a small set

244 of multiple-choice questions to estimate how much that participant “knows” about the concept
245 reflected by any arbitrary coordinate x in text embedding space (e.g., the content reflected by
246 any moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially,
247 the estimated knowledge at coordinate x is given by the weighted proportion of quiz questions
248 the participant answered correctly, where the weights reflect how much each question is “about”
249 the content at x . When we apply this approach to estimate the participant’s knowledge about
250 the content presented in each moment of each lecture, we can obtain a detailed time course
251 describing how much “knowledge” that participant has about the content presented at any part of
252 the lecture. As shown in Figure 5A and C, we can apply this approach separately for the questions
253 from each quiz participants took throughout the experiment. From just a few questions per quiz
254 (see *Estimating dynamic knowledge traces*), we obtain a high-resolution snapshot (at the time each
255 quiz was taken) of what participants knew about any moment’s content, from either of the two
256 lectures they watched (comprising a total of 1,100 samples across the two lectures).

257 While the time courses in Figure 5A and C provide detailed *estimates* about participants’ knowl-
258 ege, these estimates are of course only *useful* to the extent that they accurately reflect what partic-
259 ipants actually know. As one sanity check, we anticipated that the knowledge estimates should
260 reflect a content-specific “boost” in participants’ knowledge after watching each lecture. In other
261 words, if participants learn about each lecture’s content upon watching it, the knowledge esti-
262 mates should capture that. After watching the *Four Fundamental Forces* lecture, participants should
263 exhibit more knowledge for the content of that lecture than they had before, and that knowledge
264 should persist for the remainder of the experiment. Specifically, knowledge about that lecture’s
265 content should be relatively low when estimated using Quiz 1 responses, but should increase
266 when estimated using Quiz 2 or 3 responses (Fig. 5B). Indeed, we found that participants’ esti-
267 mated knowledge about the content of *Four Fundamental Forces* was substantially higher on Quiz 2
268 versus Quiz 1 ($t(49) = 8.764, p < 0.001$) and on Quiz 3 versus Quiz 1 ($t(49) = 10.519, p < 0.001$).
269 We found no reliable differences in estimated knowledge about that lecture’s content on Quiz 2
270 versus 3 ($t(49) = 0.160, p = 0.874$). Similarly, we hypothesized (and subsequently confirmed) that
271 participants should show greater estimated knowledge about the content of the *Birth of Stars* lec-



Figure 5: Estimating knowledge about the content presented at each moment of each lecture. **A. Knowledge about the time-varying content of *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 a single quiz (color). The traces are averaged across participants. **B. Average estimated knowledge about *Four Fundamental Forces*.** Each bar displays the across-timepoint average knowledge, estimated using the responses to one quiz's questions. **C. Knowledge about the time-varying content of *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 *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.

ture after (versus before) watching it (Fig. 5D). Specifically, since participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a “boost” on Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge about the *Birth of Stars* lecture content on Quiz 1 versus 2 ($t(49) = 1.013, p = 0.316$), but estimated knowledge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and Quiz 3 versus 1 ($t(49) = 8.969, p < 0.001$).

If we are able to accurately estimate a participant’s knowledge about the content tested by a given question, our estimates of their knowledge should carry some predictive information about whether they are likely to answer that question correctly or incorrectly. We developed a statistical approach to test this claim. For each quiz question a participant answered, in turn, we used Equation 1 to estimate their knowledge at the given question’s embedding-space coordinate based on other questions that participant answered on the same quiz. We repeated this for all participants, and for each of the three quizzes. Then, separately for each quiz, we fit a generalized linear mixed model (GLMM) with a logistic link function to explain the probability of correctly answering a question as a function of estimated knowledge at its embedding coordinate, while accounting for varied effects of individual participants and questions (see *Generalized linear mixed models*). To assess the predictive value of the knowledge estimates, we compared each GLMM to an analogous (i.e., nested) “null” model that assumed these estimates carried no predictive information using parametric bootstrap likelihood-ratio tests.

We carried out three different versions of the analyses described above, wherein we considered different sources of information in our estimates of participants’ knowledge for each quiz question. First, we estimated knowledge at each question’s embedding coordinate using *all* other questions answered by the same participant on the same quiz (“All questions”; Fig. 6, top row). This test was intended to assess the overall predictive power of our approach. Second, we estimated knowledge for each question about a given lecture using only the other questions (from the same participant and quiz) about that *same* lecture (“Within-lecture”; Fig. 6, middle rows). This test was intended to assess the *specificity* of our approach by asking whether our predictions could distinguish between questions about different content covered by the same lecture. Third, we estimated knowledge



Figure 6: Predicting success on held-out questions using estimated knowledge. We used generalized linear mixed models (GLMMs) to model the probability of correctly answering a quiz question as a function of estimated knowledge for its embedding coordinate (see *Generalized linear mixed models*). Separately for each quiz (column), we examined this relationship based on three different sets of knowledge estimates: knowledge for each question based on all other questions the same participant answered on the same quiz (“All questions”; top row), knowledge for each question about one lecture based on all other questions (from the same participant and quiz) about the *same* lecture (“Within-lecture”; middle rows), and knowledge for each question about one lecture based on all questions (from the same participant and quiz) about the *other* lecture (“Across-lecture”; bottom rows). The backgrounds in each panel display kernel density estimates of the relative observed proportions of correctly (blue) versus incorrectly (red) answered questions, for each level of estimated knowledge along the x -axis. The black curves display the (population-level) GLMM-predicted probabilities of correctly answering a question as a function of estimated knowledge. Error ribbons denote 95% confidence intervals.

300 for each question about one lecture using only the questions (from the same participant and quiz)
301 about the *other* lecture (“Across-lecture”; Fig. 6, bottom rows). This test was intended to assess the
302 *generalizability* of our approach by asking whether our predictions could extend across the content
303 areas of the two lectures. When estimating participants’ knowledge, we used a rebalancing
304 procedure to ensure that (for a given participant and quiz) their knowledge estimates for correctly
305 and incorrectly answered questions were computed from the same underlying proportion of
306 correctly answered questions (see *Generalized linear mixed models*).

307 When we fit a GLMM to estimates of participants’ knowledge for each Quiz 1 question based on
308 all other Quiz 1 questions, we found that higher estimated knowledge for a given question predicted
309 a greater likelihood of answering it correctly (odds ratio (OR) = 8.126, 95% CI = [3.116, 20.123],
310 likelihood-ratio test statistic (λ_{LR}) = 17.002, $p < 0.001$). This relationship held when we repeated
311 this analysis for Quiz 2 (OR = 14.902, 95% CI = [4.976, 39.807], λ_{LR} = 25.408, $p < 0.001$) and again
312 for Quiz 3 (OR = 37.409, 95% CI = [10.425, 107.145], λ_{LR} = 40.948, $p < 0.001$). Taken together,
313 these results suggest that our knowledge estimates can reliably predict participants’ performance
314 on individual questions when they incorporate information from all (other) quiz content.

315 We observed a similar set of results when we restricted our estimates of participants’ knowl-
316 edge for questions about each lecture to consider only their performance on other questions
317 about the *same* lecture. Specifically, for Quiz 1, participants’ knowledge of *Four Fundamental Forces*-
318 related questions, estimated from their performance on other *Four Fundamental Forces*-related ques-
319 tions, was predictive of their ability to answer those questions correctly (OR = 15.934, 95% CI =
320 [5.173, 38.005], λ_{LR} = 40.971, $p = 0.001$). The same was true of participants’ estimated knowledge
321 for *Birth of Stars*-related questions based on their performance on other *Birth of Stars*-related ques-
322 tions (OR = 9.775, 95% CI = [2.93, 25.08], λ_{LR} = 13.924, $p = 0.001$). These within-lecture knowl-
323 edge estimates also predicted success on questions about both lectures when we computed them
324 analogously for Quiz 2 (*Four Fundamental Forces*: OR = 35.126, 95% CI = [5.113, 123.868], λ_{LR} =
325 32.251, $p < 0.001$; *Birth of Stars*: OR = 4.717, 95% CI = [2.021, 9.844], λ_{LR} = 16.788, $p < 0.001$).
326 For Quiz 3, we found that within-lecture knowledge estimates predicted participants’ success on
327 *Birth of Stars*-related questions (OR = 16.902, 95% CI = [3.353, 53.265], λ_{LR} = 23.233, $p < 0.001$)

328 but not on *Four Fundamental Forces*-related questions ($OR = 2.485$, 95% CI = [0.724, 8.366], $\lambda_{LR} =$
329 1.984, $p = 0.170$). This may indicate that the within-lecture knowledge estimates are susceptible
330 to ceiling effects in participants' quiz performance. On Quiz 3, after viewing both lectures, no
331 participant answered more than three *Four Fundamental Forces*-related questions incorrectly, and
332 all but five participants (out of 50) answered two or fewer incorrectly. (This was the only subset
333 of questions about either lecture, across all three quizzes, for which this was true.) Because of
334 this, for 90% of participants, our within-lecture estimates of their knowledge for *Four Fundamen-*
335 *tal Forces*-related questions that they answered incorrectly leveraged information from at most a
336 single other question they were *not* able to correctly answer. This likely hampered our ability to
337 accurately characterize the specific (and by the time they took Quiz 3, relatively few) aspects of the
338 lecture content these participants did *not* know about, and successfully distinguish them from the
339 far more numerous aspects of the lecture content they now *did* know about. Taken together, these
340 results suggest that our knowledge estimates can reliably distinguish between questions about
341 different content covered by a single lecture, provided there is sufficient diversity in participants'
342 quiz responses to extract meaningful information about both what they know and what they do
343 not know.

344 Finally, we estimated participants' knowledge for each question about one lecture using their
345 performance on questions (from the same quiz) about the *other* lecture. This is an especially
346 stringent test of our approach. Our primary assumption in constructing our knowledge estimates is
347 that knowledge about a given concept is similar to knowledge about other concepts that are nearby
348 in the embedding space. However, our analyses in Figure 3 and Supplementary Figure 2 show
349 that the embeddings of content from the two lectures (and of their associated quiz questions) are
350 largely distinct from each other. Therefore, any predictive power of these across-lecture knowledge
351 estimates must overcome large distances in the embedding space. To put this in concrete terms,
352 this test requires predicting participants' performance on individual, highly specific questions
353 about the formation of stars from their responses to just five multiple-choice questions about the
354 fundamental forces of the universe (and vice versa).

355 We found that, before viewing either lecture (i.e., on Quiz 1), participants' abilities to answer

356 Four Fundamental Forces-related questions could not be predicted from their responses to *Birth of*
357 *Stars*-related questions ($OR = 1.896$, 95% CI = [0.419, 9.088], $\lambda_{LR} = 0.712$, $p = 0.404$), nor could
358 their abilities to answer *Birth of Stars*-related questions be predicted from their responses to *Four*
359 *Fundamental Forces*-related questions ($OR = 1.522$, 95% CI = [0.332, 6.835], $\lambda_{LR} = 0.286$, $p = 0.611$).
360 Similarly, we found that participants' performance on questions about either lecture could not
361 be predicted given their responses to questions about the other lecture after viewing *Four Fundamental Forces*
362 but before viewing *Birth of Stars* (i.e., on Quiz 2; *Four Fundamental Forces* ques-
363 tions given *Birth of Stars* questions: $OR = 3.49$, 95% CI = [0.739, 12.849], $\lambda_{LR} = 3.266$, $p =$
364 0.083; *Birth of Stars* questions given *Four Fundamental Forces* questions: $OR = 2.199$, 95% CI =
365 [0.711, 5.623], $\lambda_{LR} = 2.304$, $p = 0.141$). Only after viewing *both* lectures (i.e., on Quiz 3) did
366 these across-lecture knowledge estimates reliably predict participants' success on individual quiz
367 questions (*Four Fundamental Forces* questions given *Birth of Stars* questions: $OR = 11.294$, 95% CI =
368 [1.375, 47.744], $\lambda_{LR} = 10.396$, $p < 0.001$; *Birth of Stars* questions given *Four Fundamental Forces* ques-
369 tions: $OR = 7.302$, 95% CI = [1.077, 44.879], $\lambda_{LR} = 4.708$, $p = 0.038$). Taken together, these results
370 suggest that our ability to form estimates solely across different content areas is more limited than
371 our ability to form estimates that incorporate responses to questions from both content areas (as in
372 Fig. 6, "All questions") or within a single content area (as in Fig. 6, "Within-lecture"). However, if
373 participants have recently received some training on both content areas, the knowledge estimates
374 appear to be informative even across content areas.

375 We speculate that these "Across-lecture" results might relate to some of our earlier work on
376 the nature of semantic representations [44]. In that work, we asked whether semantic similarities
377 could be captured through behavioral measures, even if participants' "true" internal representa-
378 tions differed from the embeddings used to *characterize* their behaviors. We found that mismatches
379 between an individual's internal representation of a set of concepts and the representation used to
380 characterize their behaviors can lead to underestimates of how semantically driven those behaviors
381 are. Along similar lines, we suspect that in our current study, participants' conceptual representa-
382 tions may initially differ from the representations learned by our topic model. (Although the topic
383 model's representations are still *related* to participants' initial internal representations; otherwise

384 we would have found that knowledge estimates derived from Quizzes 1 and 2 had no predictive
385 power in the other tests we conducted.) After watching both lectures, however, participants'
386 internal representations may become more aligned with the embeddings used to estimate their
387 knowledge (since those embeddings were trained on the lectures' transcripts). This could help
388 explain why the knowledge estimates derived from Quizzes 1 and 2 (before both lectures had been
389 watched) do not reliably predict performance across content areas, whereas estimates derived from
390 Quiz 3 do.

391 That the knowledge predictions derived from the text embedding space reliably distinguish
392 between correctly and incorrectly answered held-out questions (Fig. 6) suggests that spatial re-
393 lationships within this space can help explain what participants know. But how far does this
394 explanatory power extend? For example, suppose we know that a participant correctly answered
395 a question at embedding coordinate x . As we move farther away from x in the embedding space,
396 how does the likelihood that the participant knows about the content at a given location "fall off"
397 with distance? Conversely, suppose the participant instead answered that same question *incor-*
398 *rectly*. Again, as we move farther away from x in the embedding space, how does the likelihood
399 that the participant does *not* know about the content at a given coordinate change with distance?
400 We reasoned that, assuming our embedding space is capturing something about how individuals
401 actually organize their knowledge, a participant's ability to answer questions embedded very close
402 to x should tend to be similar to their ability to answer the question embedded *at* x . But once we
403 reach some sufficiently large distance from x , our ability to infer whether or not a participant
404 will correctly answer a question based on their ability to answer the question at x should be no
405 better than guessing based on their *overall* proportion of correctly answered questions. In other
406 words, beyond the maximum distance at which a participant's ability to answer the question at x
407 is informative of their ability to answer a second question at location y , guessing the outcome at
408 y based on the outcome at x should be no more successful than guessing based on a measure that
409 does not consider embedding-space distance.

410 With these ideas in mind, we asked: conditioned on a participant's ability to answer a given
411 question correctly, what proportion of all questions within some radius r of its embedding co-

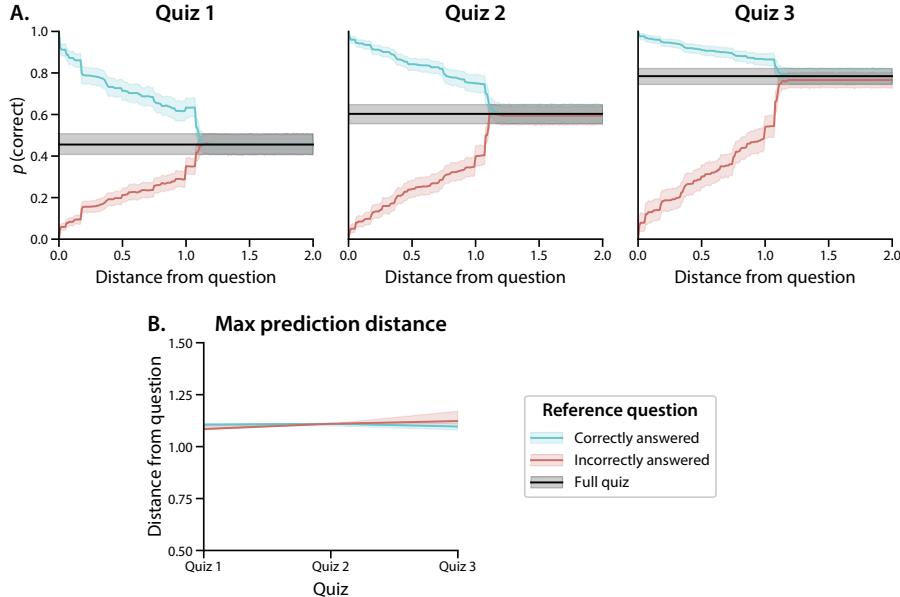


Figure 7: Knowledge falls off gradually in text embedding space. **A. Performance versus distance.** For each participant, for each correctly answered question (blue) or incorrectly answered question (red), we computed the proportion of correctly answered questions within a given distance of that question’s embedding coordinate. We used these proportions as a proxy for participants’ knowledge about the content within that region of the embedding space. We repeated this analysis for all questions and participants, and separately for each quiz (column). The black lines denote the average proportion correct across *all* questions included in the analysis at the given distance. **B. Maximum distance for which performance is reliably different from the average.** We used a bootstrap procedure (see *Estimating the “smoothness” of knowledge*) to estimate the point at which the blue and red lines in Panel A reliably diverged from the black line. We repeated this analysis separately for correctly and incorrectly answered questions from each quiz. **All panels.** Error ribbons denote bootstrap-estimated 95% confidence intervals.

412 ordinate were they able to answer correctly? We plotted this proportion as a function of r for
 413 questions that participants answered correctly, and for questions they answered incorrectly. As
 414 shown in Figure 7, we found that quiz performance falls off smoothly with distance, and the
 415 “rate” at which it falls off does not appear to differ across quizzes, as measured by the distance at
 416 which performance becomes statistically indistinguishable from a simple proportion-correct score
 417 (see *Estimating the “smoothness” of knowledge*). This suggests that, at least within the region of text
 418 embedding space spanned by the questions our study’s participants answered (and as charac-
 419 terized using our topic model), the rate at which knowledge changes with distance is relatively
 420 constant, even as participants’ overall level of knowledge varies across quizzes and regions of the

421 embedding space.

422 Knowledge estimates need not be limited to the contents of these particular lectures and quizzes.
423 As illustrated in Figure 8, our general approach to estimating knowledge from a small number
424 of quiz questions may be extended to *any* content, given its text embedding coordinate. To
425 visualize how knowledge “spreads” through text embedding space to content beyond the lectures
426 participants watched and the questions they answered, we first fit a new topic model to the lectures’
427 sliding windows with $k = 100$ topics. Conceptually, increasing the number of topics used by the
428 model functions to increase the “resolution” of the embedding space, providing a greater ability
429 to estimate knowledge for content that is highly similar to (but not precisely the same as) that
430 contained in the two lectures used to train the model. Aside from increasing the number of topics
431 from 15 to 100, all other procedures and model parameters were carried over from the preceding
432 analyses. As in our other analyses, we resampled each lecture’s topic trajectory to 1 Hz and
433 projected each question into a shared text embedding space.

434 We projected the resulting 100-dimensional topic vectors (for each second of the lectures and
435 each quiz question) onto a shared 2-dimensional plane (see *Creating knowledge and learning map*
436 *visualizations*). Next, we sampled points from a 100×100 grid of coordinates that evenly tiled a
437 rectangle enclosing the 2D projections of the lectures and questions. We then used Equation 4 to
438 estimate participants’ knowledge at each of these 10,000 sampled locations, and averaged these
439 estimates across participants to obtain an estimated average *knowledge map* (Fig. 8A). Intuitively,
440 the knowledge map constructed from a given quiz’s responses provides a visualization of “how
441 much” participants knew about any content expressible by the fitted text embedding model at
442 the point in time when they completed that quiz. We note that we used these 2D maps solely
443 for visualization; all relevant comparisons, distance computations, and statistical tests we report
444 above were carried out in the original 15-dimensional space, using the 15-topic model.

445 Several features of the resulting knowledge maps are worth noting. The average knowledge
446 map estimated from Quiz 1 responses (Fig. 8A, leftmost map) shows that participants tended to
447 have relatively little knowledge about any parts of the text embedding space (i.e., the shading is
448 relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked

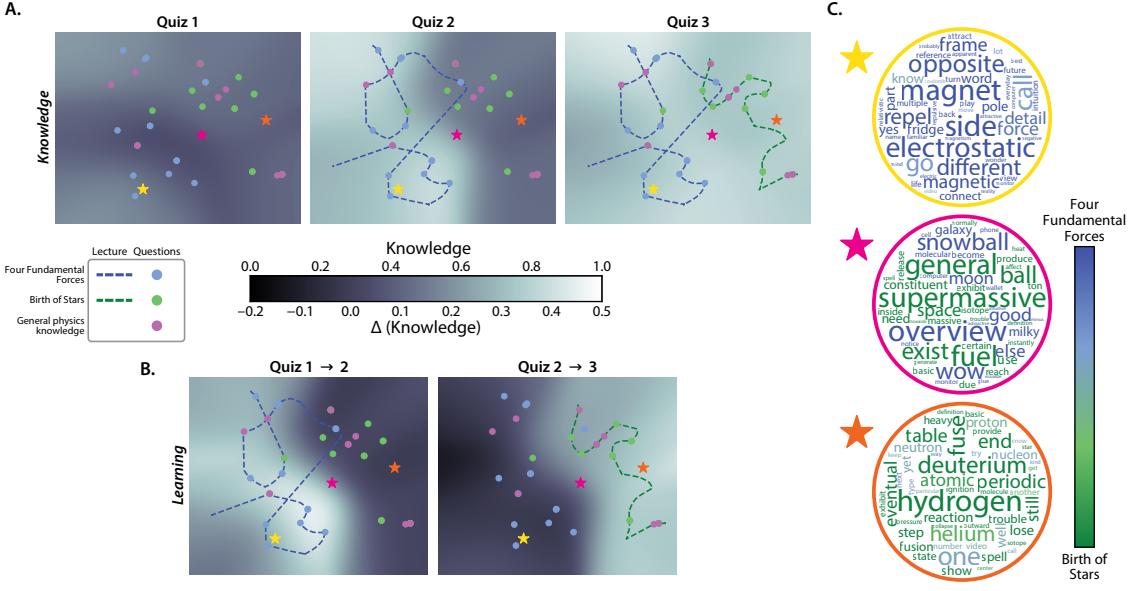


Figure 8: Mapping out the geometry of knowledge and learning. **A. Average “knowledge maps” estimated using each quiz.** Each map displays a 2D projection of participants’ estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*). The topic trajectories of the two lectures are indicated by dotted lines (blue: Lecture 1; green: Lecture 2), and the coordinates of each question are indicated by dots (light blue: Lecture 1-related; light green: Lecture 2-related; purple: general physics knowledge). Each map reflects an average across all participants. For individual participants’ maps, see Supplementary Figures 5, 6, and 7. **B. Average “learning maps” estimated between each successive pair of quizzes.** The learning maps follow the same general format as the knowledge maps in Panel A, but here the shading at each coordinate indicates the *difference* between the corresponding coordinates in the indicated *pair* of knowledge maps—i.e., how much the estimated knowledge “changed” between the two quizzes. Each map reflects an average across all participants. For individual participants’ maps, see Supplementary Figures 8 and 9. **C. Word clouds for sampled points in topic space.** Each word cloud displays the weighted blend of words underlying the topic proportions represented at the corresponding colored star’s location on the maps. In each word cloud, the words’ relative sizes correspond to their relative weights at the starred location, and their colors indicate their relative weights in the *Four Fundamental Forces* (blue) versus *Birth of Stars* (green) lectures, on average, across all timepoints’ topic vectors.

449 increase in knowledge on the left side of the map (around roughly the same range of coordinates
450 traversed by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words,
451 participants' estimated increase in knowledge is localized to conceptual content that is nearby (i.e.,
452 related to) the content from the lecture they watched prior to taking Quiz 2. This localization is
453 non-trivial: these knowledge estimates are informed only by the embedded coordinates of the
454 *quiz questions*, not by the embeddings of either lecture (see Eqn. 4). Finally, the knowledge map
455 estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region
456 surrounding the embedding of the *Birth of Stars* lecture participants watched immediately prior to
457 taking Quiz 3.

458 Another way of visualizing these content-specific increases in knowledge after participants
459 viewed each lecture is displayed in Figure 8B. Taking the point-by-point difference between the
460 knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map*
461 that describes the *change* in estimated knowledge from one quiz to the next. These learning maps
462 highlight that the estimated knowledge increases we observed across maps were specific to the
463 regions around the embeddings of each lecture, in turn.

464 Because the 2D projection we used to construct the knowledge and learning maps is invertible,
465 we may gain additional insights into these maps' meanings by reconstructing the original high-
466 dimensional topic vector for any location on the map we are interested in. For example, this could
467 serve as a useful tool for an instructor looking to better understand which content areas a student
468 (or a group of students) knows well (or poorly). As a demonstration, we show the top-weighted
469 words from the blends of topics reconstructed from three example locations on the maps (Fig. 8C):
470 one point near the *Four Fundamental Forces* embedding (yellow), a second point near the *Birth of*
471 *Stars* embedding (orange), and a third point between the two lectures' embeddings (pink). As
472 shown in the word clouds in Panel C, the top-weighted words at the example coordinate near the
473 *Four Fundamental Forces* embedding tended to be weighted more heavily by the topics expressed
474 in that lecture. Similarly, the top-weighted words at the example coordinate near the *Birth of Stars*
475 embedding tended to be weighted more heavily by the topics expressed in *that* lecture. The top-
476 weighted words at the example coordinate between the two lectures' embeddings show a roughly

477 even mix of words most strongly associated with each lecture.

478 Discussion

479 We developed a computational framework that uses short multiple-choice quizzes to gain nuanced
480 insights into what learners know and how their knowledge changes with training. First, we show
481 that our approach can automatically match the conceptual knowledge probed by individual quiz
482 questions to the corresponding moments in lecture videos when those concepts were presented
483 (Fig. 4). Next, we demonstrate how we can estimate moment-by-moment “knowledge traces” that
484 reflect the degree of knowledge participants have about each lecture’s time-varying content, and
485 capture temporally specific increases in knowledge after viewing each lecture (Fig. 5). We then
486 show that these knowledge estimates can generalize to held-out questions and predict participants’
487 abilities to answer them correctly (Fig. 6). Finally, we use our framework to construct visual maps
488 that provide snapshot estimates of how much participants know about any concept within the
489 scope of our text embedding model, and how much their knowledge of those concepts changes
490 with training (Fig. 8).

491 Our work makes several contributions to the study of how people acquire conceptual knowl-
492 edge. First, from a methodological standpoint, our modeling framework provides a systematic
493 means of mapping out and characterizing knowledge in maps that have infinite (arbitrarily many)
494 numbers of coordinates, and of “filling out” those maps using relatively small numbers of multiple-
495 choice quiz questions. Our experimental finding that we can use these maps to predict success
496 on held-out questions has several psychological implications as well. For example, concepts that
497 are assigned to nearby coordinates by the text embedding model also appear to be “known to a
498 similar extent” (as reflected by participants’ responses to held-out questions; Fig. 6). This suggests
499 that participants also *conceptualize* similarly the content reflected by nearby embedding coordi-
500 nates. How participants’ knowledge “falls off” with spatial distance is captured by the knowledge
501 maps we infer from their quiz responses (e.g., Figs. 7, 8). In other words, our study shows that
502 knowledge about a given concept implies knowledge about related concepts, and how far this

503 implication extends in text embedding space.

504 In our study, we characterize the “coordinates” of participants’ knowledge using a relatively
505 simple “bag-of-words” text embedding model [LDA; 8]. More sophisticated text embedding
506 models, such as transformer-based models [18, 55, 68, 71], can leverage additional textual infor-
507 mation such as complex grammatical and semantic relationships between words, higher-order
508 syntactic structures, stylistic features, and more. We considered using transformer-based models
509 in our study, but we found that the text embeddings derived from these models were surprisingly
510 uninformative with respect to differentiating or otherwise characterizing the conceptual content
511 of the lectures and questions we used (see *Supplementary results*). We suspect that this reflects
512 a broader challenge in constructing models that are both high-resolution within a given domain
513 (e.g., the domain of physics lectures and questions) *and* sufficiently broad as to enable them to
514 cover a wide range of domains. Essentially, these “larger” language models learn these more
515 complex features of language through training on enormous and diverse text corpora. But as a
516 result, their embedding spaces also “span” an enormous and diverse range of conceptual content,
517 sacrificing a degree of specificity in their capacities to distinguish subtle conceptual differences
518 within a more narrow range of content. In comparing our LDA model (trained specifically on the
519 lectures used in our study) to a larger transformer-based model (BERT), we found that our LDA
520 model provides both coverage of the requisite material and specificity at the level of individual
521 questions, while BERT essentially relegates the contents of both lectures and all quiz questions
522 (which are all broadly about “physics”) to a tiny region of its embedding space, thereby blurring
523 out meaningful distinctions between different specific concepts covered by the lectures and ques-
524 tions (Supp. Fig. 11). We note that these are not criticisms of BERT, nor of other large language
525 models trained on large and diverse corpora. Rather, our point is that simpler models trained on
526 relatively small but specialized corpora can outperform much more complex models trained on
527 much larger corpora when we are specifically interested in capturing subtle conceptual differences
528 at the level of a single course lecture or quiz question. On the other hand, if our goal had been to
529 choose a model that generalized to many different content areas simultaneously, we would expect
530 our LDA model to perform comparatively poorly to BERT or other much larger general-purpose

531 models. We suggest that bridging this tradeoff between high resolution within a single content
532 area and the ability to generalize to many diverse content areas will be an important challenge for
533 future work.

534 At the opposite end of the spectrum from large language models, one could also imagine
535 using an even *simpler* “model” than LDA that relates the contents of course lectures and quiz
536 questions through explicit word-overlap metrics (rather than similarities in the latent topics they
537 exhibit). In a supplementary analysis (Supp. Fig. 10), we compared the LDA-based question-lecture
538 matches shown in Figure 4 with analogous matches based on the Jaccard similarity between each
539 question’s text and each sliding window from the corresponding lecture’s transcript. Similarly
540 to the embeddings derived from BERT, we found that this approach also blurred meaningful
541 distinctions between concepts presented in different parts of each lecture and tested by different
542 quiz questions. But rather than characterizing their contents at too *broad* a semantic scale, the sorts
543 of specificity in this approach arises from considering too *narrow* a semantic scale: the sorts
544 of concepts typically conveyed in course lectures and tested by quiz questions are not defined
545 (and meaningful similarities and distinctions between them do not tend to emerge) at the level of
546 individual words.

547 In other words, while the embedding spaces of more complex large language models afford
548 low resolution at the scale of individual course lectures and questions because they “zoom out”
549 too far, simpler word-matching measures afford low resolution because they “zoom *in*” too far. In
550 this way, we view our approach as occupying a sort of “sweet spot” between simpler and more
551 complex alternatives, in that it enables us to characterize the contents of course materials at the
552 appropriate semantic scale where relevant concepts “come into focus.” Our approach enables us to
553 accurately and consistently identify each question’s content in a way that matches it with specific
554 content from the lectures and distinguishes it from other questions about similar content. In turn,
555 this enables us to construct accurate predictions about participants’ knowledge of the conceptual
556 content tested by individual quiz questions (Fig. 6).

557 Another application for large language models that does *not* require explicitly modeling the
558 content of individual lectures or questions is to leverage these models’ abilities to generate text. For

559 example, generative text models like ChatGPT [55] and LLaMa [68] are already being used to build
560 a new generation of interactive tutoring systems [e.g., 45]. Unlike the approach we have taken here,
561 these generative text model-based systems do not explicitly model what learners know, or how
562 their knowledge changes over time with training. One could imagine building a hybrid system
563 that combines the best of both worlds: a large language model that can *generate* text, combined
564 with a smaller model that can *infer* what learners know and how their knowledge changes over
565 time. Such a hybrid system could potentially be used to build the next generation of interactive
566 tutoring systems that are able to adapt to learners' needs in real time, and provide more nuanced
567 feedback about what learners know and what they do not know.

568 One limitation of our approach is that topic models contain no explicit internal representations
569 of more complex aspects of "knowledge," like knowledge graphs, dependencies or associations
570 between concepts, causality, and so on. These representations might (in principle) be added
571 as extensions to our approach to more accurately and precisely capture, characterize, and track
572 learners' knowledge. However, modeling these aspects of knowledge will likely require substantial
573 additional research effort.

574 Within the past several years, a global pandemic forced many educators to suddenly adapt to
575 teaching remotely [35, 52, 64, 72]. This change in world circumstances is happening alongside (and
576 perhaps accelerating) geometric growth in the availability of high-quality online courses from plat-
577 forms such as Khan Academy [36], Coursera [73], EdX [38], and others [60]. Continued expansion
578 of the global internet backbone and improvements in computing hardware have also facilitated
579 improvements in video streaming, enabling videos to be easily shared and viewed by increasingly
580 large segments of the world's population. This exciting time for online course instruction provides
581 an opportunity to re-evaluate how we, as a global community, educate ourselves and each other.
582 For example, we can ask: what defines an effective course or training program? Which aspects of
583 teaching might be optimized and/or augmented by automated tools? How and why do learning
584 needs and goals vary across people? How might we lower barriers to receiving a high-quality
585 education?

586 Alongside these questions, there is a growing desire to extend existing theories beyond the

587 domain of lab testing rooms and into real classrooms [34]. In part, this has led to a recent
588 resurgence of “naturalistic” or “observational” experimental paradigms that attempt to better
589 reflect more ethologically valid phenomena that are more directly relevant to real-world situations
590 and behaviors [53]. In turn, this has brought new challenges in data analysis and interpretation. A
591 key step towards solving these challenges will be to build explicit models of real-world scenarios
592 and how people behave in them (e.g., models of how people learn conceptual content from real-
593 world courses, as in our current study). A second key step will be to understand which sorts
594 of signals derived from behaviors and/or other measurements [e.g., neurophysiological data; 4,
595 19, 50, 54, 57] might help to inform these models. A third major step will be to develop and
596 employ reliable ways of evaluating the complex models and data that are a hallmark of naturalistic
597 paradigms.

598 Beyond specifically predicting what people *know*, the fundamental ideas we develop here also
599 relate to the notion of “theory of mind” of other individuals [26, 32, 49]. Considering others’ unique
600 perspectives, prior experiences, knowledge, goals, etc., can help us to more effectively interact and
601 communicate [58, 63, 67]. One could imagine future extensions of our work (e.g., analogous to
602 the knowledge and learning maps shown in Fig. 8), that attempt to characterize how well-aligned
603 different people’s knowledge bases or backgrounds are. In turn, this might be used to model how
604 knowledge (or other forms of communicable information) flows not just between teachers and
605 students, but between friends having a conversation, individuals on a first date, participants at
606 a business meeting, doctors and patients, experts and non-experts, political allies or adversaries,
607 and more. For example, the extent to which two people’s knowledge maps “match” or “align” in
608 a given region of text embedding space might serve as a predictor of how effectively they will be
609 able to communicate about the corresponding conceptual content.

610 Ultimately, our work suggests a rich new line of questions about the geometric “form” of
611 knowledge, how knowledge changes over time, and how we might map out the full space of
612 what an individual knows. Our finding that detailed estimates about knowledge may be obtained
613 from short quizzes shows one way that traditional approaches to evaluation in education may be
614 extended. We hope that these advances might help pave the way for new approaches to teaching

615 or delivering educational content that are tailored to individual students' learning needs and goals.

616 **Materials and methods**

617 **Participants**

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

623 Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09
624 years). A total of 15 participants reported their gender as male and 35 participants reported their
625 gender as female. A total of 49 participants reported their native language as "English" and 1
626 reported having another native language. A total of 47 participants reported their ethnicity as
627 "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants
628 reported their races as White (32 participants), Asian (14 participants), Black or African American
629 (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other
630 Pacific Islander (1 participant). (Note that some participants selected multiple racial categories.)

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

638 No participants reported that their focus was currently impaired (e.g., by drugs or alcohol).
639 Participants reported their current level of alertness, and we converted their responses to numerical

640 scores as follows: “very sluggish” (-2), “a little sluggish” (-1), “neutral” (0), “fairly alert” (1), and
641 “very alert” (2). Across all participants, a range of alertness levels were reported (range: -2–1;
642 mean: -0.10; standard deviation: 0.84).

643 Participants reported their undergraduate major(s) as “social sciences” (28 participants), “nat-
644 ural sciences” (16 participants), “professional” (e.g., pre-med or pre-law; 8 participants), “mathe-
645 matics and engineering” (7 participants), “humanities” (4 participants), or “undecided” (3 partici-
646 pants). Note that some participants selected multiple categories for their undergraduate major(s).
647 We also asked participants about the courses they had taken. In total, 45 participants reported hav-
648 ing taken at least one Khan Academy course in the past, and 5 reported not having taken any Khan
649 Academy courses. Of those who reported having watched at least one Khan Academy course,
650 7 participants reported having watched 1–2 courses, 11 reported having watched 3–5 courses, 8
651 reported having watched 5–10 courses, and 19 reported having watched 10 or more courses. We
652 also asked participants about the specific courses they had watched, categorized under different
653 subject areas. In the “Mathematics” area, participants reported having watched videos on AP
654 Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Cal-
655 culus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry
656 (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential
657 Equations (5 participants), Statistics and Probability (4 participants), AP Statistics (2 participants),
658 Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other
659 videos not listed in our survey (5 participants). In the “Science and engineering” area, participants
660 reported having watched videos on Chemistry, AP Chemistry, or Organic Chemistry (21 partic-
661 ipants); Physics, AP Physics I, or AP Physics II (18 participants); Biology, AP Biology, or High
662 school Biology (15 participants); Health and Medicine (1 participant); and other videos not listed
663 in our survey (5 participants). We also asked participants whether they had specifically seen the
664 videos used in our experiment. Of the 45 participants who reported having having taken at least
665 one Khan Academy course in the past, 44 participants reported that they had not watched the
666 *Four Fundamental Forces* video and 1 participant reported that they were not sure whether they had
667 watched it. All participants reported that they had not watched the *Birth of Stars* video. When

668 we asked participants about non-Khan Academy online courses, they reported having watched
669 or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test
670 preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 partic-
671 ipants), Computing (2 participants), and other categories not listed in our survey (17 participants).
672 Finally, we asked participants about in-person courses they had taken in different subject areas.
673 They reported taking courses in Mathematics (38 participants), Science and engineering (37 partic-
674 ipants), Arts and humanities (34 participants), Test preparation (27 participants), Economics and
675 finance (26 participants), Computing (14 participants), College and careers (7 participants), and
676 other courses not listed in our survey (6 participants).

677 **Experiment**

678 We hand-selected two course videos from the Khan Academy platform: *Four Fundamental Forces*
679 (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force;
680 duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed;
681 duration: 7 minutes and 57 seconds). All participants viewed the videos in the same order (i.e.,
682 *Four Fundamental Forces* followed by *Birth of Stars*).

683 We then hand-created 39 multiple-choice questions: 15 about the conceptual content of *Four*
684 *Fundamental Forces* (i.e., Lecture 1), 15 about the conceptual content of *Birth of Stars* (i.e., Lecture 2),
685 and 9 questions that tested for general conceptual knowledge about basic physics (covering material
686 that was not presented in either video). To help broaden the set of lecture-specific questions, our
687 team worked through each lecture in small segments to identify what each segment was “about”
688 conceptually, and then write a question about that concept. The general physics questions were
689 drawn from our team’s prior coursework and areas of interest, along with internet searches and
690 brainstorming with the project team and other members of J.R.M.’s lab. Although we attempted to
691 design the questions to test “conceptual knowledge,” we note that estimating the specific “amount”
692 of conceptual understanding that each question “requires” to answer is somewhat subjective, and
693 might even come down to the “strategy” a given participant used to answer the question at that
694 particular moment. The full set of questions and answer choices may be found in Supplementary

695 Table 1. The final set of questions (and response options) was reviewed and approved by J.R.M.
696 before we collected or analyzed the text or experimental data.

697 Over the course of the experiment, participants completed three 13-question multiple-choice
698 quizzes: the first before viewing Lecture 1, the second between Lectures 1 and 2, and the third
699 after viewing Lecture 2 (see Fig. 1). The questions appearing on each quiz, for each participant,
700 were randomly chosen from the full set of 39, with the constraints that (a) each quiz contained
701 exactly 5 questions about Lecture 1, 5 questions about Lecture 2, and 3 questions about general
702 physics knowledge, and (b) each question appear exactly once for each participant. The orders of
703 questions on each quiz, and the orders of answer options for each question, were also randomized.
704 We obtained informed consent from all participants, and our experimental protocol was approved
705 by the Committee for the Protection of Human Subjects at Dartmouth College. We used this
706 experiment to develop and test our computational framework for estimating knowledge and
707 learning.

708 **Analysis**

709 **Statistics**

710 All of the statistical tests performed in our study were two-sided. The 95% confidence intervals
711 we report for each correlation were estimated from bootstrap distributions of 10,000 correlation
712 coefficients obtained by sampling (with replacement) from the observed data.

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

714 We adapted an approach we developed in prior work [29] to embed each moment of the two
715 lectures and each question in our pool in a common representational space. Briefly, our approach
716 uses a topic model [Latent Dirichlet Allocation; 8] trained on a set of documents to discover a set
717 of k “topics” or “themes.” Formally, each topic is defined as a distribution of weights over words in
718 the model’s vocabulary (i.e., the union of all unique words across all documents, excluding “stop
719 words”). Conceptually, each topic is intended to give larger weights to words that are semantically

720 related (as inferred from their tendency to co-occur in the same document). After fitting a topic
721 model, each document in the training set, or any *new* document that contains at least some of
722 the words in the model’s vocabulary, may be represented as a k -dimensional vector describing
723 how much that document (most probably) reflects each topic. To select an appropriate k for our
724 model, as a starting point, we identified the minimum number of topics that yielded at least one
725 “unused” topic (i.e., in which all words in the vocabulary were assigned uniform weights) after
726 training. This indicated that the number of topics was sufficient to capture the set of latent themes
727 present in the two lectures (from which we constructed our document corpus, as described below).
728 We found this value to be $k = 15$ topics. We found that with a limited number of additional
729 adjustments following Boyd-Graber et al. [9], such as removing corpus-specific stop-words, the
730 model yielded (subjectively) sensible and coherent topics. The distribution of weights over words
731 in the vocabulary for each discovered topic is shown in Supplementary Figure 1, and each topic’s
732 top-weighted words may be found in Supplementary Table 2.

733 As illustrated in Figure 2A, we started by building up a corpus of documents using overlapping
734 sliding windows that spanned each lecture’s transcript. Khan Academy provides professionally
735 created, manual transcriptions of all lecture videos for closed captioning. However, such tran-
736 scripts would not be readily available in all contexts to which our framework could potentially be
737 applied. Khan Academy videos are hosted on the YouTube platform, which additionally provides
738 automated captions. We opted to use these automated transcripts [which, in prior work, we have
739 found to be of sufficiently near-human quality to yield reliable data in behavioral studies; 74]
740 when developing our framework in order to make it more directly extensible and adaptable by
741 others in the future.

742 We fetched these automated transcripts using the `youtube-transcript-api` Python pack-
743 age [17]. Each transcript consisted of one timestamped line of text for every few seconds (mean:
744 2.34 s; standard deviation: 0.83 s) of spoken content in the lecture (i.e., corresponding to each in-
745 dividual caption that would appear on-screen if viewing the lecture via YouTube, and when those
746 lines would appear). We defined a sliding window length of (up to) $w = 30$ transcript lines and
747 assigned each window a timestamp corresponding to the midpoint between the timestamps for its

748 first and last lines. This w parameter was chosen to match the same number of words per sliding
749 window (rounded to the nearest whole word, and before preprocessing) as the sliding windows
750 we defined in our prior work [29; i.e., 185 words per sliding window].

751 These sliding windows ramped up and down in length at the beginning and end of each
752 transcript, respectively. In other words, each transcript's first sliding window covered only its first
753 line, the second sliding window covered the first two lines, and so on. This ensured that each line
754 from the transcripts appeared in the same number (w) of sliding windows. We next performed a
755 series of standard text preprocessing steps: normalizing case, lemmatizing, removing punctuation
756 and removing stop-words. We constructed our corpus of stop words by augmenting the Natural
757 Language Toolkit [NLTK; 5] English stop word list with the following additional words, selected
758 using one of the approaches suggested by Boyd-Graber et al. [9]: "actual," "actually," "also," "bit,"
759 "could," "e," "even," "first," "follow," "following," "four," "let," "like," "mc," "really," "saw,"
760 "see," "seen," "thing," and "two." This yielded sliding windows containing an average of 73.8
761 remaining words, and spanning an average of 62.22 seconds. We treated the text from each sliding
762 window as a single "document" and combined these documents across the two lectures' windows
763 to create a single training corpus for the topic model.

764 After fitting the topic model to the two lectures' transcripts, we could use the trained model to
765 transform arbitrary (potentially new) documents into k -dimensional topic vectors. A convenient
766 property of these topic vectors is that documents that reflect similar blends of topics (i.e., documents
767 that reflect similar themes, according to the model) will yield similar coordinates (in terms of
768 correlation, cosine similarity, Kullback-Leibler divergence, Euclidean distance, or other geometric
769 measures). In general, the similarity between different documents' topic vectors may be used to
770 characterize the similarity in conceptual content between the documents.

771 We transformed each sliding window's text into a topic vector, and then used linear interpolation
772 (independently for each topic dimension) to resample the resulting time series to one vector
773 per second. We also used the fitted model to obtain topic vectors for each quiz question in our pool
774 (see Supp. Tab. 1). Taken together, we obtained a *trajectory* for each lecture video, describing its path
775 through topic space, and a single coordinate for each question (Fig. 2C). Embedding both lectures

776 and all of the questions using a common model enables us to compare the content from different
777 moments of the lectures, compare the content across lectures, and estimate potential associations
778 between specific questions and specific moments of lecture content.

779 **Estimating dynamic knowledge traces**

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

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

782 where

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

783 and where mincorr and maxcorr are the minimum and maximum correlations between the topic
784 vectors for any lecture timepoint and quiz question, taken over all timepoints in the given lecture
785 and all questions *about* that lecture appearing on the given quiz. We also define $f(s, \Omega)$ as the s^{th}
786 topic vector from the set of topic vectors Ω . Here t indexes the time series of lecture topic vectors
787 L , and i and j index the topic vectors of questions Q used to estimate the participant’s knowledge.
788 Note that “correct” denotes the set of indices of the questions the participant answered correctly
789 on the given quiz.

790 Intuitively, $\text{ncorr}(x, y)$ is the correlation between two topic vectors (e.g., the topic vector x
791 for one timepoint in a lecture and the topic vector y for one question on a quiz), normalized
792 by the minimum and maximum correlations (across all timepoints t and questions j) to range
793 between 0 and 1, inclusive. Equation 1 then computes the weighted average proportion of correctly
794 answered questions about the content presented at timepoint t , where the weights are given by the
795 normalized correlations between timepoint t ’s topic vector and the topic vectors for each question.
796 The normalization step (i.e., using ncorr instead of the raw correlations) ensures that every question
797 contributes some non-negative amount to the knowledge estimate.

798 **Generalized linear mixed models**

799 In the set of analyses reported in Figure 6, we assessed whether estimates of participants' knowl-
800 edge at the embedding coordinates of individual quiz questions could be used to reliably predict
801 their abilities to correctly answer those questions. In essence, we treated each question a given
802 participant answered on a given quiz as a "lecture" consisting of a single timepoint, and used
803 Equation 1 to estimate the participant's knowledge for its embedding coordinate based on their
804 performance on all *other* questions they answered on that same quiz ("All questions"; Fig. 6,
805 top row). Additionally, for each lecture-related question (i.e., excluding questions about general
806 physics knowledge), we computed analogous knowledge estimates based on two different subsets
807 of questions the participant answered on the same quiz: (1) all *other* questions about the same
808 lecture as the target question ("Within-lecture"; Fig. 6, middle rows), and (2) all questions about
809 the other of the two lectures ("Across-lecture"; Fig. 6, bottom rows).

810 In performing these analyses, our null hypothesis is that the knowledge estimates we compute
811 based on the quiz questions' embedding coordinates do *not* provide useful information about par-
812 ticipants' abilities to correctly answer those questions—in other words, that there is no meaningful
813 difference (on average) between the knowledge estimates we compute for questions participants
814 answered correctly versus incorrectly. Specifically, since we estimate knowledge for a given em-
815 bedding coordinate as a weighted proportion-correct score (where each question's weight reflects
816 its embedding-space distance from the target coordinate; see Eqn. 1), if these weights are un-
817 informative (e.g., randomly distributed), then our estimates of participants' knowledge should
818 be equivalent (on average) to the *unweighted* proportion of correctly answered questions used to
819 compute them. In general, for a given participant and quiz, this expected null value (i.e., that
820 participant's proportion-correct score on that quiz) is the same for any coordinate in the embed-
821 ding space (e.g., any lecture timepoint, quiz question, etc.). However, in the "All questions" and
822 "Within-lecture" versions of the analyses shown in Figure 6, we estimate each participant's knowl-
823 edge for each target question using all *other* questions (or all *other* questions about the same lecture)
824 they answered on the same quiz. This introduces a systematic dependency between a participant's

825 success on a target question and their proportion-correct score on the remaining questions available
826 to estimate their knowledge for it. For example, suppose a participant correctly answered n out
827 of q questions on a given quiz. If we hold out a single *correctly* answered question as the target,
828 the proportion of remaining questions answered correctly would be $\frac{n-1}{q-1}$, whereas if we hold out
829 a single *incorrectly* answered question, the proportion of remaining questions answered correctly
830 would be $\frac{n}{q-1}$. Thus, the proportion of correctly answered remaining questions (and therefore the
831 null-hypothesized value of a knowledge estimate computed from them) is always *lower* for target
832 questions a participant answered correctly than for those they answered incorrectly.

833 To correct for this baseline difference under our null hypothesis, we used a rebalancing pro-
834 cedure that ensured our knowledge estimates for questions each participant answered correctly
835 and incorrectly were computed from the *same* proportion of correctly answered questions. For
836 each target question on a given participant’s quiz, we first identified all remaining questions with
837 the opposite “correctness” label (i.e., if the target question was answered correctly, we identified
838 all remaining incorrectly answered questions, and vice versa). We then held out each of these
839 opposite-label questions, in turn, along with the target question, and estimated the participant’s
840 knowledge for the target question using all *other* remaining questions. Since each of these subsets
841 of remaining questions was constructed by holding out one correctly answered question and one
842 incorrectly answered question from the participant’s quiz responses, if the participant correctly
843 answered n out of q questions total, then their proportion-correct score on each subset of questions
844 used to estimate their knowledge would be $\frac{n-1}{q-2}$, regardless of whether they answered the target
845 question correctly or incorrectly. Finally, we averaged over these per-subset knowledge estimates
846 to obtain a rebalanced estimate of the participant’s knowledge for the target question that lever-
847 aged information from all remaining questions’ embedding coordinates, but whose expected value
848 under our null hypothesis was the same as that of each individual subset ($\frac{n-1}{q-2}$). By equalizing the
849 null-hypothesized values of knowledge estimates for correctly and incorrectly answered ques-
850 tions, this procedure ensures that any meaningful relationships we observe between participants’
851 estimated knowledge for individual quiz questions and their abilities to correctly answer them
852 reflect the predictive power of the embedding-space distances we use to weight questions’ con-

tributions to the knowledge estimates, rather than an artifact of our testing procedure. Note that if a participant answered all or no questions on a given quiz correctly, their responses contained no opposite-label questions with which to perform this rebalancing, and we therefore excluded their data from our analyses for that quiz. We used this rebalancing procedure when constructing knowledge estimates for the “All questions” and “Within-lecture” versions of the analyses shown in Figure 6, but not for the “Across-lecture” analyses as, in this case, the target questions and the questions used to estimate participants’ knowledge for them were drawn from different subsets of quiz questions (those about one lecture, and those about the other), and were therefore independent.

In each version of this analysis (i.e., row in Fig. 6), and separately for each of the three quizzes (i.e., column in Fig. 6), we then fit a generalized linear mixed model (GLMM) with a logistic link function to the set of knowledge estimates for all questions (or all questions about a particular lecture) that participants answered on the given quiz. We implemented these models in R using the `lme4` package [3] and fit them following guidance from Bates et al. [2] and Matuschek et al. [46]. Specifically, we initially fit each model with the maximal random effects structure afforded by our design, which we identified as:

$$\text{accuracy} \sim \text{knowledge} + (\text{knowledge} | \text{participant}) + (\text{knowledge} | \text{question})$$

where “accuracy” is a binary value indicating whether each target question was answered correctly or incorrectly, “knowledge” is estimated knowledge at each target question’s embedding coordinate, “participant” is a unique identifier assigned to each participant, and “question” is a unique identifier assigned to each quiz question. For models we fit using knowledge estimates for target questions about multiple content areas (i.e., in the “All questions” version of the analysis), we also included an additional random effect term, $(\text{knowledge} | \text{lecture})$, where “lecture” is a categorical value denoting whether the target question was about *Four Fundamental Forces*, *Birth of Stars*, or general physics knowledge. Note that with our coding scheme, identifiers for each question are implicitly nested within levels of lecture and so do not require explicit nesting in

878 our model formula. We then iteratively removed random effects from the maximal model until it
879 successfully converged with a full-rank random effects variance-covariance matrix. We obtained
880 the odds ratios reported in Figure 6 by exponentiating the estimated coefficient for “knowledge”
881 from each fitted model. Conceptually, these odds ratios represent how many times greater the odds
882 are that a given participant will answer a given question correctly if their estimated knowledge
883 for its embedding coordinate is 1, compared to if it is 0. We estimated 95% confidence intervals
884 for each odds ratio by generating 10,000 random subsamples (of full size, with replacement) from
885 the data used to fit each model, and refitting the models to each subsample to obtain bootstrap
886 distributions of 10,000 odds ratios.

887 To assess the predictive value of our knowledge estimates, we compared each GLMM’s ability
888 to explain participants’ success on individual quiz questions to that of an analogous model which
889 assumed (as we assume under our null hypothesis) that knowledge estimates for correctly and
890 incorrectly answered questions did *not* systematically differ, on average. Specifically, we used the
891 same sets of observations to which we fit each “full” model to fit a second “null” model with
892 the same random effects structure, but with the coefficient for the fixed effect of “knowledge” con-
893 strained to zero (i.e., we removed this term from the null model). We then compared each full model
894 to its reduced (null) equivalent using a likelihood-ratio test (LRT). Because the standard asymptotic
895 χ_d^2 approximation of the null distribution for the LRT statistic (λ_{LR}) can be anti-conservative for
896 finite sample sizes [25, 61, 66], we computed p -values for these tests using a parametric bootstrap
897 procedure [14, 27]. For each of 10,000 bootstraps, we used the fitted null model to simulate a
898 sample of observations of equal size to our original sample. We then re-fit both the null and full
899 models to this simulated sample and compared them via an LRT. This yielded a distribution of λ_{LR}
900 statistics we may expect to observe given data that conforms to our null hypothesis. We computed
901 a corrected p -value for our observed λ_{LR} as $\frac{r+1}{n+1}$, where r is the number of simulated model com-
902 parisons that yielded a λ_{LR} greater than our observed value and n is the number of simulations we
903 ran (10,000).

904 **Estimating the “smoothness” of knowledge**

905 In the analysis reported in Figure 7A, we show how participants’ ability to correctly answer
906 quiz questions changes as a function of distance from a given correctly or incorrectly answered
907 reference question. We used a bootstrap-based approach to estimate the maximum distances over
908 which these proportions of correctly answered questions could be reliably distinguished from
909 participants’ overall average proportion of correctly answered questions.

910 For each of 10,000 iterations, we drew a random subsample (with replacement) of 50 partic-
911 ipants from our dataset. Within each iteration, we first computed the 95% confidence interval
912 (CI) of the across-subsample-participants mean proportion correct on each of the three quizzes,
913 separately. To compute this interval for each quiz, we repeatedly (1,000 times) subsampled par-
914 ticipants (with replacement, from the outer subsample for the current iteration) and computed
915 the mean proportion correct of each of these inner subsamples. We then identified the 2.5th and
916 97.5th percentiles of the resulting distributions of 1,000 means. These three intervals (one for each
917 quiz) served as our thresholds for confidence that the proportion correct within a given distance
918 from a reference question was reliably different (at the $p < 0.05$ significance level) from the average
919 proportion correct across all questions on the given quiz.

920 Next, for each participant in the current subsample, and for each of the three quizzes they
921 completed (separately), we iteratively treated each of the 15 questions appearing on the given
922 quiz as the “reference” question. We constructed a series of concentric 15-dimensional “spheres”
923 centered on the reference question’s embedding-space coordinate, where each successive sphere’s
924 radius increased by 0.01 (correlation distance) between 0 and 2, inclusive (i.e., tiling the range
925 of possible correlation distances with 201 spheres in total). We then computed the proportion
926 of questions enclosed within each sphere that the participant answered correctly, and averaged
927 these per-radius proportion-correct scores across reference questions that were answered correctly,
928 and those that were answered incorrectly. This resulted in two number-of-spheres sequences of
929 proportion-correct scores for each subsample participant and quiz: one derived from correctly
930 answered reference questions, and one derived from incorrectly answered reference questions.

931 We computed the across-subsample-participants mean proportion correct for each radius value
932 (i.e., sphere) and “correctness” of reference question. This yielded two sequences of proportion-
933 correct scores for each quiz, analogous to the blue and red lines displayed in Figure 7A, but for
934 the present subsample. For each quiz, we then found the minimum distance from the reference
935 question (i.e., sphere radius) at which each of these two sequences of per-radius proportion-correct
936 scores intersected the 95% confidence interval for the overall proportion correct (i.e., analogous to
937 the black error bands in Fig. 7A).

938 This resulted in two “intersection” distances for each quiz (for correctly answered and incor-
939 rectly answered reference questions). Repeating this full process for each of the 10,000 bootstrap
940 iterations output two distributions of intersection distances for each of the three quizzes. The
941 means and 95% confidence intervals for these distributions are plotted in Figure 7B.

942 **Creating knowledge and learning map visualizations**

943 An important feature of our approach is that, given a trained text embedding model and partic-
944 ipants’ performance on each quiz question, we can estimate their knowledge about *any* content
945 expressible by the embedding model—not solely the content explicitly probed by the quiz ques-
946 tions, or even appearing in the lectures. To visualize these estimates (Fig. 8, Supp. Figs. 5, 6, 7, 8,
947 and 9), we used Uniform Manifold Approximation and Projection [UMAP; 47, 48] to construct a
948 2D projection of the text embedding space. Whereas our main analyses used a 15-topic embedding
949 space, we used a 100-topic embedding space for these visualizations. This change in the number
950 of topics overcame an undesirable behavior in the UMAP embedding procedure, whereby embed-
951 ding coordinates for the 15-topic model tended to be “clumped” into separated clusters, rather
952 than forming a smooth trajectory through the 2D space. When we increased the number of topics
953 to 100, the embedding coordinates in the 2D space formed a smooth trajectory through the space,
954 with substantially less clumping (Fig. 8). Creating a “map” by sampling this 100-dimensional
955 space at high resolution to obtain an adequate set of topic vectors spanning the embedding space
956 would be computationally intractable. However, sampling a 2D grid is trivial.

957 At a high level, the UMAP algorithm obtains low-dimensional embeddings by minimizing

958 the cross-entropy between the pairwise (clustered) distances between the observations in their
959 original (e.g., 100-dimensional) space and the pairwise (clustered) distances in the low-dimensional
960 embedding space (in our approach, the embedding space is 2D). In our implementation, pairwise
961 distances in the original high-dimensional space were defined as 1 minus the correlation between
962 each pair of coordinates, and pairwise distances in the low-dimensional embedding space were
963 defined as the Euclidean distance between each pair of coordinates.

964 In our application, all of the coordinates we embedded were topic vectors, whose elements
965 are always non-negative and sum to one. Although UMAP is an invertible transformation at
966 the embedding locations of the original data, other locations in the embedding space will not
967 necessarily follow the same implicit “rules” as the original high-dimensional data. For example,
968 inverting an arbitrary coordinate in the embedding space might result in negative-valued vectors,
969 which are incompatible with the topic modeling framework. To protect against this issue, we
970 log-transformed the topic vectors prior to embedding them in the 2D space. When we inverted
971 the embedded vectors (e.g., to estimate topic vectors for word clouds, as in Fig. 8C), we passed
972 the inverted (log-transformed) values through the exponential function to obtain a vector of non-
973 negative values, and normalized them to sum to one.

974 After embedding both lectures’ topic trajectories and the topic vectors of every question, we
975 defined a rectangle enclosing the 2D projections of the lectures’ and quizzes’ embeddings. We then
976 sampled points from a regular 100×100 grid of coordinates that evenly tiled this enclosing rectangle.
977 We sought to estimate participants’ knowledge (and learning, i.e., changes in knowledge) at each
978 of the resulting 10,000 coordinates.

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

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

983 The λ term in the RBF equation controls the “smoothness” of the function, where larger values

984 of λ result in smoother maps. In our implementation we used $\lambda = 50$. Next, we estimated the
985 “knowledge” at each coordinate, x , using:

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

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

990 Author contributions

991 Conceptualization: P.C.F., A.C.H., and J.R.M. Methodology: P.C.F., A.C.H., and J.R.M. Software:
992 P.C.F. Validation: P.C.F. Formal analysis: P.C.F. Resources: P.C.F., A.C.H., and J.R.M. Data curation:
993 P.C.F. Writing (original draft): J.R.M. Writing (review and editing): P.C.F., A.C.H., and J.R.M. Visu-
994 alization: P.C.F. and J.R.M. Supervision: J.R.M. Project administration: P.C.F. Funding acquisition:
995 J.R.M.

996 Data availability

997 All of the data analyzed in this manuscript may be found at <https://github.com/ContextLab/efficient-learning-khan>.

999 Code availability

1000 All of the code for running our experiment and carrying out the analyses may be found at
1001 <https://github.com/ContextLab/efficient-learning-khan>.

1002 **Acknowledgements**

1003 We acknowledge useful discussions, assistance in setting up an earlier (unpublished) version of
1004 this study, and assistance with data collection efforts from Will Baxley, Max Bluestone, Daniel
1005 Carstensen, Kunal Jha, Caroline Lee, Lucy Owen, Xinming Xu, and Kirsten Ziman. Our work was
1006 supported in part by NSF CAREER Award Number 2145172 to J.R.M. The content is solely the
1007 responsibility of the authors and does not necessarily represent the official views of our supporting
1008 organizations. The funders had no role in study design, data collection and analysis, decision to
1009 publish, or preparation of the manuscript.

1010 **References**

- 1011 [1] Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*,
1012 56:149–178.
- 1013 [2] Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015a). Parsimonious mixed models. *arXiv*,
1014 1506.04967.
- 1015 [3] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015b). Fitting linear mixed-effects models
1016 using lme4. *Journal of Statistical Software*, 67(1):1–48.
- 1017 [4] Bevilacqua, D., Davidesco, I., Wan, L., and Chaloner, K. (2019). Brain-to-brain synchrony and
1018 learning outcomes vary by student-teacher dynamics: evidence from a real-world classroom
1019 electroencephalography study. *Journal of Cognitive Neuroscience*, 31(3):401–411.
- 1020 [5] Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text
1021 with the natural language toolkit*. Reilly Media, Inc.
- 1022 [6] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in chil-
1023 dren: distinguishing response flexibility from conceptual flexibility; the protracted development
1024 of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.

- 1025 [7] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International*
1026 *Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing
1027 Machinery.
- 1028 [8] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
1029 *Learning Research*, 3:993–1022.
- 1030 [9] Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models:
1031 problems, diagnostics, and improvements. In Airolidi, E. M., Blei, D. M., Erosheva, E. A., and
1032 Fienberg, S. E., editors, *Handbook of Mixed Membership Models and Their Applications*. CRC Press.
- 1033 [10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,
1034 Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,
1035 R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M.,
1036 Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei,
1037 D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.
- 1038 [11] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the
1039 evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.
- 1040 [12] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-
1041 Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal
1042 sentence encoder. *arXiv*, 1803.11175.
- 1043 [13] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual
1044 knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.
- 1045 [14] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge
1046 Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- 1047 [15] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004).
1048 Evidence for a new conceptualization of semantic representation in the left and right cerebral
1049 hemispheres. *Cortex*, 40(3):467–478.

- 1050 [16] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
1051 Indexing by latent semantic analysis. *Journal of the American Society for Information Science*,
1052 41(6):391–407.
- 1053 [17] Depoix, J. (2018). YouTube transcript API. <https://github.com/jdepoix/youtube-transcript-api>.
- 1055 [18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep
1056 bidirectional transformers for language understanding. *arXiv*, 1810.04805.
- 1057 [19] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J.,
1058 Michalareas, G., van Bavel, J. J., Ding, M., and Poeppel, D. (2017). Brain-to-brain synchrony
1059 tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27(9):1375–1380.
- 1060 [20] Estes, W. K. (1986a). Array models for category learning. *Cognitive Psychology*, 18(4):500–549.
- 1061 [21] Estes, W. K. (1986b). Memory storage and retrieval processes in category learning. *Journal of*
1062 *Experimental Psychology: General*, 115:155–174.
- 1063 [22] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical*
1064 *Transactions of the Royal Society A*, 222(602):309–368.
- 1065 [23] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge.
1066 *School Science and Mathematics*, 100(6):310–318.
- 1067 [24] Gluck, M. A., Shohamy, D., and Myers, C. E. (2002). How do people solve the “weather
1068 prediction” task? individual variability in strategies for probabilistic category learning. *Learning*
1069 *and Memory*, 9:408–418.
- 1070 [25] Goldman, N. and Whelan, S. (2000). Statistical Tests of Gamma-Distributed Rate Heterogeneity
1071 in Models of Sequence Evolution in Phylogenetics. *Molecular Biology and Evolution*, 17(6):975–978.
- 1072 [26] Goldstein, T. R. and Winner, E. (2012). Enhancing empathy and theory of mind. *Journal of*
1073 *Cognition and Development*, 13(1):19–37.

- 1074 [27] Halekoh, U. and Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric
1075 Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *Journal of*
1076 *Statistical Software*, 59(9):1–32.
- 1077 [28] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual
1078 learning, pages 212–221. Sage Publications.
- 1079 [29] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behav-
1080 ioral and neural signatures of transforming experiences into memories. *Nature Human Behaviour*,
1081 5:905–919.
- 1082 [30] Huebner, P. A. and Willits, J. A. (2018). Structured semantic knowledge can emerge au-
1083 tomatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*,
1084 9:doi.org/10.3389/fpsyg.2018.00133.
- 1085 [31] Hulbert, J. C. and Norman, K. A. (2015). Neural differentiation tracks improved recall of com-
1086 peting memories following interleaved study and retrieval practice. *Cerebral Cortex*, 25(10):3994–
1087 4008.
- 1088 [32] Kanske, P., Böckler, A., and Singer, T. (2015). Models, mechanisms and moderators dissociating
1089 empathy and theory of mind. In *Social Behavior From Rodents to Humans*, pages 193–206. Springer.
- 1090 [33] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*.
1091 Columbia University Press.
- 1092 [34] Kaufman, D. M. (2003). Applying educational theory in practice. *British Medical Journal*,
1093 326(7382):213–216.
- 1094 [35] Kawasaki, H., Yamasaki, S., Masuoka, Y., Iwasa, M., Fukita, S., and Matsuyama, R. (2021).
1095 Remote teaching due to COVID-19: an exploration of its effectiveness and issues. *International*
1096 *Journal of Environmental Research and Public Health*, 18(5):2672.
- 1097 [36] Khan, S. (2004). *The Khan Academy*. Salman Khan.

- 1098 [37] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.
- 1099 [38] Kolowich, S. (2013). How EdX plans to earn, and share, revenue from its free online courses.
1100 *The Chronicle of Higher Education*, 21:1–5.
- 1101 [39] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic
1102 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
1103 104:211–240.
- 1104 [40] Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic
1105 events. *Nature Communications*, 13(4235):doi.org/10.1038/s41467-022-31965-2.
- 1106 [41] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of
1107 Educational Studies*, 53(2):129–147.
- 1108 [42] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
1109 function? *Psychological Review*, 128(4):711–725.
- 1110 [43] Manning, J. R. (2023). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors,
1111 *Handbook of Human Memory*. Oxford University Press.
- 1112 [44] Manning, J. R. and Kahana, M. J. (2012). Interpreting semantic clustering effects in free recall.
1113 *Memory*, 20(5):511–517.
- 1114 [45] Manning, J. R., Menjunatha, H., and Kording, K. (2023). Chatify: A Jupyter extension
1115 for adding LLM-driven chatbots to interactive notebooks. [https://github.com/ContextLab/
1116 chatify](https://github.com/ContextLab/chatify).
- 1117 [46] Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type i error
1118 and power in linear mixed models. *Journal of Memory and Language*, 94:305–315.
- 1119 [47] McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: Uniform manifold approximation and
1120 projection for dimension reduction. *arXiv*, 1802(03426).

- 1121 [48] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). UMAP: Uniform Manifold
1122 Approximation and Projection. *Journal of Open Source Software*, 3(29):861.
- 1123 [49] Meltzoff, A. N. (2011). Social cognition and the origins of imitation, empathy, and theory of
1124 mind. In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Wiley-Blackwell.
- 1125 [50] Meshulam, M., Hasenfratz, L., Hillman, H., Liu, Y. F., Nguyen, M., Norman, K. A., and Hasson,
1126 U. (2020). Neural alignment predicts learning outcomes in students taking an introduction to
1127 computer science course. *Nature Communications*, 12(1922):doi.org/10.1038/s41467-021-22202-3.
- 1128 [51] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word represen-
1129 tations in vector space. *arXiv*, 1301.3781.
- 1130 [52] Moser, K. M., Wei, T., and Brenner, D. (2021). Remote teaching during COVID-19: implications
1131 from a national survey of language educators. *System*, 97:102431.
- 1132 [53] Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of
1133 experimental control in cognitive neuroscience. *NeuroImage*, 15(222):117254–117261.
- 1134 [54] Nguyen, M., Chang, A., Micciche, E., Meshulam, M., Nastase, S. A., and Hasson, U. (2022).
1135 Teacher-student neural coupling during teaching and learning. *Social Cognitive and Affective*
1136 *Neuroscience*, 17(4):367–376.
- 1137 [55] OpenAI (2023). ChatGPT. <https://chat.openai.com>.
- 1138 [56] Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models.
1139 *arXiv*, 2208.02957.
- 1140 [57] Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., and Hansen, L. K. (2017). EEG
1141 in the classroom: synchronised neural recordings during video presentation. *Scientific Reports*,
1142 7:43916.
- 1143 [58] Ratka, A. (2018). Empathy and the development of affective skills. *American Journal of*
1144 *Pharmaceutical Education*, 82(10):doi.org/10.5688/ajpe7192.

- 1145 [59] Reilly, D. L., Cooper, L. N., and Elbaum, C. (1982). A neural model for category learning.
1146 *Biological Cybernetics*, 45(1):35–41.
- 1147 [60] Rhoads, R. A., Berdan, J., and Toven-Lindsey, B. (2013). The open courseware movement in
1148 higher education: unmasking power and raising questions about the movement’s democratic
1149 potential. *Educational Theory*, 63(1):87–110.
- 1150 [61] Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for a zero random
1151 effect variance or polynomial regression in additive and linear mixed models. *Computational
1152 Statistics & Data Analysis*, 52(7):3283–3299.
- 1153 [62] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter
1154 Student conceptions and conceptual learning in science. Routledge.
- 1155 [63] Shao, Y. N., Sun, H. M., Huang, J. W., Li, M. L., Huang, R. R., and Li, N. (2018). Simulation-
1156 based empathy training improves the communication skills of neonatal nurses. *Clinical Simula-
1157 tion in Nursing*, 22:32–42.
- 1158 [64] Shim, T. E. and Lee, S. Y. (2020). College students’ experience of emergency remote teaching
1159 during COVID-19. *Children and Youth Services Review*, 119:105578.
- 1160 [65] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for
1161 conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in
1162 Mathematics Education*, 35(5):305–329.
- 1163 [66] Snijders, T. A. B. and Bosker, R. (2011). More powerful tests for variance parameters. In
1164 *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, chapter 6, pages
1165 94–108. Sage Publications, 2nd edition.
- 1166 [67] Stepien, K. A. and Baernstein, A. (2006). Education for empathy. *Journal of General Internal
1167 Medicine*, 21:524–530.
- 1168 [68] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B.,

- 1169 Goyal, N., Hambro, E., Azhar, F., Rodriguz, A., Joulin, A., Grave, E., and Lample, G. (2023).
1170 LLaMA: open and efficient foundation language models. *arXiv*, 2302.13971.
- 1171 [69] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Pio-
1172 ntkovskaya, I., Nikolenko, S., and Burnaev, E. (2023). Intrinsic dimension estimation for robust
1173 detection of AI-generated texts. *arXiv*, 2306.04723.
- 1174 [70] van Paridon, J., Liu, Q., and Lupyan, G. (2021). How do blind people know that blue is cold?
1175 distributional semantics encode color-adjective associations. *Proceedings of the Annual Meeting of*
1176 *the Cognitive Science Society*, 43(43).
- 1177 [71] Viswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and
1178 Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing*
1179 *Systems*.
- 1180 [72] Whalen, J. (2020). Should teachers be trained in emergency remote teaching? Lessons learned
1181 from the COVID-19 pandemic. *Journal of Technology and Teacher Education*, 28(2):189–199.
- 1182 [73] Young, J. R. (2012). Inside the Coursera contract: how an upstart company might profit from
1183 free courses. *The Chronicle of Higher Education*, 19(7):1–4.
- 1184 [74] Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is
1185 automatic speech-to-text transcription ready for use in psychological experiments? *Behavior*
1186 *Research Methods*, 50:2597–2605.