Content Matters: A Computational Investigation into the Effectiveness of Retrieval Practice and Worked Examples (Extended Abstract) *

Napol Rachatasumrit, **Paulo F. Carvalho**, **Sophie Li** and **Kenneth R. Koedinger**
Carnegie Mellon University
{napol, koedinger}@cmu.edu, pcarvalh@cs.cmu.edu, li.sophie80@gmail.com

Abstract

In this paper, we argue that computational models of learning can contribute precise theory to explain surprising student learning phenomena. In some past studies, practice produces better learning than studying examples, whereas other studies show the opposite result. We explain this contradiction by suggesting that retrieval practice and example study involve different learning cognitive processes, memorization and induction, and each process is optimal for different types of knowledge. We implement and test this theoretical explanation by extending an AI model of human cognition to include both memory and induction processes and comparing the behavior of the simulated learners to those of human participants. We show that the behavior of simulated learners with forgetting matches that of human participants better than simulated learners without forgetting. Simulated learners with forgetting learn best using retrieval practice in situations that emphasize memorization (such as learning facts), whereas studying examples improves learning when multiple pieces of information are available, so induction and generalization are necessary (such as learning skills).

1 Introduction

Retrieval practice — repeatedly trying to retrieve information by completing practice questions — has been shown to improve performance compared to re-studying [Roediger III *et al.*, 2011; Roediger III and Karpicke, 2006]. Interestingly, re-study trials in the form of worked examples have also been shown to improve performance compared to answering more practice questions [Van Gog *et al.*, 2006; Renkl, 2005]. This apparent contradiction poses both theoretical and practical issues. Theoretically, to which degree do we have a complete understanding of the learning process if opposite approaches can yield similar results? Practically, when making suggestions for the application of cognitive science findings to educational contexts, practitioners are left wondering which approach to use and when.

The Knowledge Learning Instruction framework (KLI) offers a potential explanation [Koedinger *et al.*, 2012]. A key premise of KLI is that optimal instructional design decisions depend on what learning process the student must engage in, which depends on the nature of the target knowledge component (KC). A KC is a stable unit of cognitive function that is acquired and modifiable. KLI identifies three types of Learning Events: Memory processes, induction processes, and sense-making processes, associated with the types of KCs. KLI also offers a taxonomy for KCs based on how they function across Learning Events. In this way, KCs can be classified based on their application and response conditions. Facts such as "the capital of France is Paris" are constant application and constant response KCs because there is only one single application of the KC and only one response. Conversely, skills such as equation solving are variable application and variable response KCs because multiple problems can elicit the same KC and there are multiple ways to apply this KC across different problems (e.g., solving an equation that one never saw, using a generalization from past examples).

Moreover, the KLI framework further suggests causal links between instructional principles (e.g., "retrieval practice", "worked-example study"), and changes in learner knowledge. For simple constant KCs such as facts, memory processes are more relevant. Conversely, for variable KCs such as skills, induction processes are more relevant. Thus, different types of KCs will interact with different types of Instructional Principles to create different learning. In the context of facts ("What is the capital of France?"), learners need to successfully encode all of the information presented and be able to retrieve it later. Learning facts only requires learning the specific pieces of single practice items but does not require any synthesis across practice items. Conversely, in the context of skills ("Calculate the area of a rectangle with the following measurements"), learners need to generalize their knowledge across a series of studied instances. In this sense, learning skills requires identifying which pieces of information are relevant for encoding and which are not.

Carvalho et al. [Carvalho *et al.*, 2022] proposed that retrieval practice improves memory processes and strengthens associations, whereas studying examples improves inference processes and information selection for encoding. This proposal is also consistent with previous work [Karpicke and Blunt, 2011; Salden *et al.*, 2010]. To test this hypothesis,

[*] The work was initially published at International Conference on Artificial Intelligence in Education (2023)

Carvalho et al. [Carvalho *et al.*, 2022] conducted an experiment using a basic mathematical domain (calculating the area of geometrical shapes). Human participants (N=95) were divided into 4 conditions: practice-only training of facts, study-practice training of facts, practice-only training of skills, and study-practice of skills. A significant interaction was found between the type of concept studied and the type of training.

In this study, we use an AI model of human learning to demonstrate the likelihood of these mechanisms generating behavioral results similar to Carvalho et al.'s experiment and examine the extent to which the memory mechanism influences human learning. Furthermore, this study provides further evidence for the utility of computational models of human learning in the advancement of learning theory. As proposed by MacLellan et al.[Maclellan *et al.*, 2016], the use of such models enables a bridge between learning theory and educational data, allowing for the testing and refinement of fundamental theories of human learning. This study extends this concept by demonstrating the ability of these models to contribute to evaluating theories that can explain even surprising student learning phenomena, for which existing learning theories may offer inconsistent explanations.

2 A Computational Model of Human Learning

Simulated learners (SL) are AI systems that learn to perform tasks through an interactive process, such as human demonstrations and feedback, usually with mechanisms intended to model how humans learn. In this work, we used the Apprentice Learner framework (AL), a framework for creating SLs based on different mechanistic theories of learning. Details on AL and its operation can be found elsewhere [Maclellan *et al.*, 2016; Weitekamp *et al.*, 2021]; briefly, AL agents learn a set of production rules through an induction mechanism. The agents receive a set of states as input and search for the existing production rules that are applicable. If none are applicable, AL agents will request a demonstration of correct action and go through the induction process to construct a new rule for the current set of states. Later, when the agents encounter states that use the same production rule, the rule will get generalized or fine-tuned according to the examples they encounter. The learning process in AL is largely deterministic but some of the learning mechanisms have stochastic elements. For example, when multiple possible actions are possible, a stochastic probability matching process is used to select which one to execute.

The AL framework's production rules consist of two sets of conditions - the left-hand side (LHS) and right-hand side (RHS) - that include three essential components: where-part, when-part, and how-part. RHS is the action that AL thinks it should take in the form of a Selection-ActionType-Input (SAI) triple. The value of the SAI is calculated by a function composition called how-part, given the values extracted from the input states by the where-part. The where-part determines which state elements the production rule could be applied to. These sets of elements are called bindings and contain the elements to be used as arguments to the RHS of the skill and the selection (the element to be acted upon). The when-part of a skill, a binary function (often in the form of a set of conditions), determines whether or not a particular skill should be activated given the current state of the system. Initially, the when-part and where-part may be either over-specific or over-general but will be refined to the appropriate level as the agent receives additional demonstrations or feedback on the skill. For instance, if the agent is given an example of calculating the area of a rectangle (e.g. given l = 4, h = 3, then the area is 12) and another for a triangle (e.g. given l = 5, h = 4, then the area is 10), it will learn two distinct skills (i.e. rectangle-area and triangle-area skills). However, since each skill is demonstrated only once, the conditions in the LHS remain unrefined. Consequently, when the agent is presented with another rectangle-area problem, it may mistakenly activate the triangle-area skill due to over-general conditions. Upon receiving negative feedback, AL will refine the when-part of the triangle-area skill to be more specific, such as only activating it when the shape is a triangle.

In previous work, AL agents have been shown to demonstrate human-like behaviors in learning academic tasks, such as fractions arithmetic, and multi-column addition [Maclellan *et al.*, 2016]. Here, we used AL to test the mechanistic hypothesis that retrieval practice involves memory and retrieval processes, whereas studying examples involves induction processes. To do this, we developed a memory mechanism in AL and compared the performance of AL agents learning facts and skills in a setup similar to previous empirical results with humans (see also Simulation Studies below). We compare learning outcomes following training of facts and skills, using retrieval practice (practice-only) or worked examples (study-practice). In our study, we employed the same subject matter, but we altered the learning focus between fact acquisition (e.g, "What is the formula to calculate the area of a triangle?") and skill acquisition (e.g. "What is the area of the triangle below?").

3 Simulation Studies

3.1 Data

The current work replicates the findings of Carvalho et al.'s [Carvalho *et al.*, 2022] experiment on the effect of retrieval practice and worked examples on the different types of knowledge. In their studies, participants were divided into four groups: practice-only training of facts, study-practice training of facts, practice-only training of skills, and study-practice of skills. The participants learned how to calculate the area of four different geometrical shapes (rectangle, triangle, circle, and trapezoid) through a training phase consisting of studying examples and practicing memorizing formulas or solving problems. Multiple-choice tests were used as pre/posttests, divided into two types of knowledge: fact-based ("What is the formula to calculate the area of the square?") and skill-based ("What is the area of a square that is 9 ft wide?"). There was no feedback provided during practices.

To replicate the findings, our materials were adapted from the original study. Since the focus of our hypothesis is the interaction between types of training and types of knowledge, we simplify the encoding of the problems such that an agent can focus on picking the correct production rule and selecting the relevant information from the states presented. For example, instead of giving an agent a diagram, the input states include parsed information from the diagram, such as a shape type or lines' lengths. This simplification is also plausible as parsing shapes is likely to be prior knowledge that humans brought to the task. The solutions to the problems were given during study sessions, in contrast to only questions without solutions or feedback in pre/posttests. For fact-based materials, the solutions were in the form of a corresponding string, while relevant operations, foci-of-attention, and numerical answers were provided for the skill-based materials.

3.2 Method

To evaluate our hypothesis that memory and forgetting processes are necessary for a learning benefit of retrieval practice, we leveraged the AL framework to create two models of human learning: a model with forgetting (SLwF) and a model without forgetting (SLoF). Our memory mechanism implementation is based on Pavlik et al.'s memory model using ACT-R [Pavlik and Anderson, 2008]:

$$m_n(t_{1\ldots m}) = \beta + b_k + ln(\sum_{k=1}^{n} t_k^{-d_k}) \tag{1}$$

An activation strength (m_n) depends on the base activation (β), the strength of a practice type (b_k), ages of trials (τ_k), and decay rates (d_k). The decay rate for each trial depends on the decay scale parameter (c), the intercept of the decay function (α), and the activation strength of prior trials (m_{k-1}):

$$d_k(m_{k-1}) = ce^{m_{k-1}} + \alpha \tag{2}$$

The parameters were selected based on an initial parameter manual search. The activation strength of each production rule will be updated through the mathematical process described above, every time it is successfully retrieved both through demonstrations/examples or practice testing, but with different corresponding parameter values depending on the type of training. The probability of a successful recall for a production rule will be calculated using the recall equation when SLs attempt to retrieve the rule. In other words, the success of a production rule being activated for SLs depends on the model they are based on. In SLoF, the process is deterministic and the applicable rule will always be activated. However, in SLwF, the process is stochastic, with the probability of a successful being activated determined by the probability of a successful recall of the associated production rule.

There were 95 AL agents, each agent matching a human participant in [Carvalho *et al.*, 2022], assigned to one of four conditions: practice-only training of facts (N=27), study-practice training of facts (N=22), practice-only training of skills (N=18), and study-practice training of skills (N=28). Each agent went through the same procedure as human participants. It completed 16 pretest questions, 4 study sessions, and then completed 16 posttest question.

In the study session, The agents were divided into two groups: the practice-only group, where they were trained with one demonstration (worked example) followed by three practice tests, and the study-practice condition, where they alternated between both types of training. During the practice tests, the agents were only provided with binary corrective feedback without the correct answer. The objective of the learning process for facts was for the agents to effectively link the appropriate constant (i.e. a formula string) to the specific state of the problem, as specified by the constant-constant condition (e.g. shape == trapezoid corresponds to the formula "$A = \frac{1}{2}(a + b) * h$"). The objective of learning skills, on the other hand, was for the agents to not only identify the appropriate formula to apply, but also to select the relevant variables from the given state of the problem, as outlined by the variable-variable condition (e.g. shape == square, base-length == 5, and diagonal-length == $\sqrt{50}$ corresponds to 5^2).

To account for participants' prior knowledge, we pre-trained each SL to match human pretest performance [Weitekamp III *et al.*,] (*M* = 0.59 and 0.60, for facts and skills).

4 Results

4.1 Learning Gain

Similar to Carvalho et al. [Carvalho *et al.*, 2022], we analyzed posttest performance controlling for pretest performance, for each type of trained concept (skills vs. facts) and training type (practice-only, vs. study-practice). A two-way ANOVA was performed to analyze the effects of type of training and type of concept studied on learning gains, and the results showed that there was a statistically significant interaction between the effects of type of training and type of concept in SLwF (*F*(1, 471) = 9.448, *p* = .002), but none was found in the SLoF (*F*(1, 471) = -3.843, *p* = 1). Moreover, consistent with our prediction, simple main effects analysis showed that the type of training did have a statistically significant effect on learning gains in SLoF (*F*(1, 471) = 7.364, *p* = 0.007), but not in SLwF (*F*(1, 471) = 0.845, *p* = 0.359). On the other hand, the type of concept studied had a statistically significant effect on learning gains in both SLwF (*F*(1, 471) = 13.052, *p* ¡ 0.001) and SLoF (*F*(1, 471) = 29.055, *p* ¡ 0.001). The similar pattern can also be seen in Fig. 1, comparing the learning gains for each condition between human participants (a), SLs without forgetting - SLoF (b), and SLs with forgetting - SLwF (c). The results indicate that SLwF in a study-practice condition led to higher learning gains for skills than a practice-only condition (19.9% vs 15.8%), *t*(228) = -2.404, *p* = 0.009, but the opposite was true for facts (12.7% vs 15.6%), *t*(243) = 2.072, *p* = 0.020. However, SLoF led to higher learning gains for both skills (26.1% vs 24.4%), *t*(228) = 1.106, *p* = 0.135 and facts (21.6% vs 20.0%), *t*(243) = -1.713, *p* = 0.044, in the study-practice condition. These results suggest that SLwF better align with human learning patterns.

4.2 Error Type

To further investigate the extent to which memory plays a role in this mechanistic hypothesis, we analyzed the types of errors made by SLwF at posttest (since SLoF cannot commit a memory-based error, it would be unnecessary to conduct the analysis). We classified errors into two categories: memory-based and induction-based. Memory-based errors

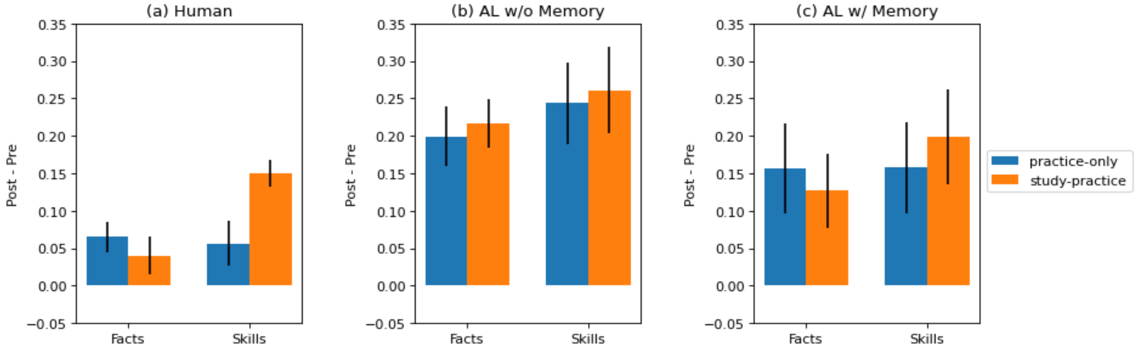

Figure 1: Learning Gains Comparison between type of training and type of concept.

occurred when an applicable production rule was learned but not retrieved in the final test, whereas induction-based errors occurred when incorrect production rules were found or none were found. Overall, SLwF committed more induction-based errors than memory-based errors (56.2% vs 43.8%). Additionally, SLwF in practice-only condition committed fewer memory-based errors compared to the ones from study-practice condition (41.8% vs 45.7%), but more induction-based errors (58.2% vs 54.2%), $t(466)$ = -1.467, p = 0.072; even though, both groups exhibited similar proportions of both categories (82.7% vs 83.9% for induction-based errors and 17.3% vs 16.1% for memory-based errors) at pretest.

5 General Discussion

Our results indicate SLwF align well with humans, with retrieval practice being more effective for facts and worked examples being more effective for skills. In contrast, for SLoF, worked examples are more beneficial for both facts and skills, as the lack of a memory mechanism does not allow for the benefits of retrieval practice to be realized. This supports our hypothesis that, according to the KLI framework, retrieval practice improves memory processes and strengthens associations, making it beneficial for learning facts where all presented information is important. Conversely, studying examples improves inference processes and information selection for encoding, making it beneficial for learning skills where only a subset of presented information is relevant.

Interestingly, the introduction of a memory mechanism slightly decreases learning gains (22.2% for SLoF vs 13.7% for SLwF), $t(948)$ = 9.409, p ¡ 0.0001, but does not negate the benefits of worked examples over retrieval practice for skills. Furthermore, the breakdown of error categories revealed more induction-based errors than memory-based errors (59.1% vs 40.9%). This supports our hypothesis that skills learning involves more selectivity and inference, which are better aided by worked examples than by increased memory activation through retrieval practice.

In fact, the gap between retrieval practice and worked examples for skill learning increases even more in SLwF (1.7% vs 3.1%). A closer examination suggested that this is because they can ”forget” incorrect production rules, so the correct rules are used more effectively. SLwF were found to be more likely to select the correct production rules, due to stronger memory activation (because correct production rules are usually learned after incorrect ones, and not vice versa, allowing the correct production rules to have a stronger activation in memory). This offers a cursory insight into the importance of ”forgetting” misconceptions for successful learning, but further research is required to fully understand this mechanism.

Here, we have presented evidence that computational models of human learning can be a bridge between learning theory and data. This approach allows for an examination of learning theory in a variety of scenarios. In particular, Carvalho et al. [Carvalho *et al.*, 2022] have proposed a plausible mechanism to explain the inconsistencies between the effects of retrieval practices and worked examples on learning focusing on the selectivity of encoding of the tasks. Leveraging computational models of learning, we employed SLs as a means of validating the proposed theoretical framework. These SLs served as a valuable tool for investigating the mechanism of learning in greater depth, as we were able to analyze the types of errors made during the learning process. Additionally, by comparing the proposed theory to other existing theories, we were able to determine which theory best aligns with human data. Furthermore, an in-depth examination of SLs revealed interesting insights, such as the potential benefits of forgetting in skill acquisition, which can serve as a guide for future research directions. Therefore, we have emphasized the possibilities that can be achieved through the use of computational models in education research.

6 Conclusions

This study has highlighted the utility of computational models of human learning in bridging the gap between learning theory and data, as demonstrated through examination of unexpected learning phenomena. We started with an unexpected learning phenomena (inconsistencies in the effects of retrieval practices and worked examples on learning), and a proposed plausible mechanism (a mechanism focusing on the selectivity of encoding of the tasks). Then, with computational models, we were able to not only confirm but also examine this proposed learning theory in more depth, which highlights the potential of computational models in the field of education research. Our findings demonstrate the potential for these models to inform the development of more effective teaching strategies and guide future research in this area.

References

[Carvalho *et al.*, 2022] Paulo F. Carvalho, Napol Rachatasumrit, and Kenneth R Koedinger. Learning depends on knowledge: The benefits of retrieval practice vary for facts and skills. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2022.

[Karpicke and Blunt, 2011] Jeffrey D Karpicke and Janell R Blunt. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018):772–775, 2011.

[Koedinger *et al.*, 2012] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.

[Maclellan *et al.*, 2016] Christopher J Maclellan, Erik Harpstead, Rony Patel, and Kenneth R Koedinger. The apprentice learner architecture: Closing the loop between learning theory and educational data. *International Educational Data Mining Society*, 2016.

[Pavlik and Anderson, 2008] Philip I Pavlik and John R Anderson. Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14(2):101, 2008.

[Rachatasumrit *et al.*, 2023] Napol Rachatasumrit, Paulo F Carvalho, Sophie Li, and Kenneth R Koedinger. Content matters: A computational investigation into the effectiveness of retrieval practice and worked examples. In *International Conference on Artificial Intelligence in Education*, pages 54–65. Springer, 2023.

[Renkl, 2005] Alexander Renkl. The worked-out-example principle in multimedia learning. *The Cambridge handbook of multimedia learning*, pages 229–245, 2005.

[Roediger III and Karpicke, 2006] Henry L Roediger III and Jeffrey D Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3):249–255, 2006.

[Roediger III *et al.*, 2011] Henry L Roediger III, Pooja K Agarwal, Mark A McDaniel, and Kathleen B McDermott. Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, 17(4):382, 2011.

[Salden *et al.*, 2010] Ron JCM Salden, Kenneth R Koedinger, Alexander Renkl, Vincent Aleven, and Bruce M McLaren. Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review*, 22(4):379–392, 2010.

[Van Gog *et al.*, 2006] Tamara Van Gog, Fred Paas, and Jeroen JG Van Merriënboer. Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction*, 16(2):154–164, 2006.

[Weitekamp *et al.*, 2021] Daniel Weitekamp, Christopher MacLellan, Erik Harpstead, and Kenneth Koedinger. Decomposed inductive procedure learning, 2021.

[Weitekamp III *et al.*,] Daniel Weitekamp III, Erik Harpstead, Christopher J MacLellan, Napol Rachatasumrit, and Kenneth R Koedinger. Toward near zero-parameter prediction using a computational model of student learning. *Ann Arbor*, 1001:48105.