Title
Faulty Memories, Favored Outcomes: How Errors Impact Learning Processes

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Leonard, Bridget
hake, holly
Stocco, Andrea

Publication Date
2023

Peer reviewed

Faulty Memories, Favored Outcomes: How Errors Impact Learning Processes

Bridget Leonard (bll313@uw.edu)
Department of Psychology, University of Washington, Seattle, WA 98195 USA

Holly Sue Hake (hakehs@uw.edu)
Department of Psychology, University of Washington, Seattle, WA 98195 USA

Andrea Stocco (stocco@uw.edu)
Department of Psychology, University of Washington, Seattle, WA 98195 USA

Abstract

Recent studies suggest that errors facilitate learning in certain conditions. Despite this, reinforcement paradigms dominate learning methods, subscribing to the narrative that errorless learning is the foundation of an ideal learning environment. If we continue to view learning from this restrictive perspective, we may fail to capture and apply the benefits of errors. In this paper, we investigate two potential mechanisms of post-error learning. Participants (N = 61) learned word pairs in either a study or error trial before taking a final test. Supporting past error learning literature, errors before a study opportunity led to better performance on a final test. Differences in reaction times between conditions support the theory that errors increase learning by acting as a mediator, or secondary cue, to the correct answer on subsequent tests.

Keywords: learning from errors; memory; retrieval; elaboration; mediation; computational models; ACT-R

Introduction

Many people believe that ideal learning is errorless. Errors are often viewed as detrimental to learning with the expectation that they will interfere with the future retrieval of correct information (Ceraso, 1967). These concepts stem from studying more procedural behavior where learning is a constant reinforcement process; thus, errorless learning minimizes opportunities to reinforce errors. This understanding has biased how we optimize learning, backing a paradigm in which learners re-study concepts before even being given a chance to commit a memory error. However, modern-day experimental findings suggest that errors do not pollute the learning process.

Posing a question before revealing the answer, also known as *retrieval practice*, appears to enable stronger learning. Different experimental manipulations all seem to share a common thread: although the difficulty of retrieval practice impairs performance in the very first few trials, it ultimately results in better final test performance. This has been shown in multiple choice versus short-answer questions (Greving & Richter, 2018), problems based on a single lesson or spaced across multiple (Rohrer & Taylor, 2007), and retrieval schedules (wide spacing of retrieval attempts vs. short spacing; Roediger & Karpicke, 2010). This widespread finding suggests that errors made early in the learning process from retrieval difficulty can result in better learning.

Additionally, reinforcement paradigms would predict that errors made with higher confidence would reduce learning. Such phenomena would be consistent with interference theories of memory, positing that incorrect items with more substantial traces are likely to produce more interference when attempting to retrieve a correct item (Anderson & Reder, 1999). However, experimental results show that errors committed with high confidence are more likely to be successfully corrected (Butterfield & Metcalfe, 2001). This suggests that more processing between the onset of a question and the presentation of the answer (i.e., a retrieval attempt) leads to better answer processing, regardless of whether the attempt was successful (Kornell & Vaughn, 2016). Such findings may demonstrate that the behaviorist approach does not translate well to learning by memorization compared to more procedural learning. Achieving successful memorization requires the ability to retrieve information; errors may enhance this type of learning by promoting stronger encoding and thus, more successful recall. Thus, we may be able to enhance the most common form of learning for humans by critically examining how and when errors are beneficial to memorization and subsequent retrieval.

Researchers have introduced new paradigms to study post-error learning, or improvements in recall of subsequent fact presentations after an incorrect answer or memory errors, such as the *pretesting paradigm*. In this paradigm, participants generate an answer *before* studying it; with no previous exposure to the correct answer, participants are highly likely to generate an incorrect response (Mera, Rodriguez, & Marin-Garcia, 2021). Although this paradigm encourages errors, pretesting for information is still more beneficial than simply studying it. In a paired-associate task, retrieving a free associate of the cue word before learning the correct targets produces better results on a final test than studying the cue and target alone (Kornell, Hays, & Bjork, 2009). Many studies have confirmed this finding, extending the benefits of pretesting to real-world materials (trivia questions; Kornell, 2014), educational materials (math problems; Kapur & Bielaczyc, 2012), and older adults (Cyr & Anderson, 2015).

Although this phenomenon, known as *post-error learning*, is now well-documented, an investigation into its underlying mechanisms is sparse. Two prominent theories have arisen out of this research; the elaborative hypothesis and the mediator hypothesis. To go beyond speculation, both must

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

be examined empirically to successfully leverage post-error learning.

The Elaborative Hypothesis

The elaborative theory of post-error learning posits that unsuccessful retrieval attempts allow for a richer encoding of the correct answer. Also known as the Search Set theory, retrieval attempts (regardless of success) activate a variety of semantically related candidates. The search set contains many candidates, one of which is the correct answer, thus setting up a network among the cue and target words (Mera et al., 2021). In the pretesting paradigm and other semantic cue-target pairings, an error could help form a more meaningful relationship between the cue and target. For example, one may generate the word "swims" as a free associate in response to the cue word "whale" when the target word is actually "tail." Instead of simply encoding the pair "whale" and "tail," the individual may use the error to create a more robust network between the two words, perhaps thinking of a whale using its tail for swimming. The underlying idea is that prompted retrieval of the target word following the presentation of a cue word evokes several semantically related items. Merging these concepts forms an elaborative memory trace at the time of encoding, which is more likely to be retrieved later at subsequent cue presentations (Huelser & Metcalfe, 2012; Karpicke, 2017). One important finding supporting this theory is that weakly associated word pairs produce stronger retrieval learning than strong associates (Carpenter, 2009). Participants are more likely to retrieve the target word in response to a cue word if they are strongly related. In contrast, weakly related word pairs prompt participants to generate many related words to recall the target. Elaboration enhances future retrieval by assuming individuals encode additional semantically related items alongside the cue and target words.

A primary strength of this theory is its congruence with spreading activation theories of memory. In a spreading activation model, semantic relations link words together in the mental lexicon. The presentation of a cue word activates its corresponding node in the lexical web, with activation spreading to all the nodes connected or semantically related to it. Related words will be activated based on the strength of their association with the cue word (Collins & Loftus, 1975). Thus, the retrieval of a specific memory is not only a function of the frequency and recency of its presentation, or base-level activation, but its degree of relation to the current context or spreading activation (Anderson, 1983). Within the framework of this theory, it is clear how elaborative encoding could interact with an already established semantic lexical network. However, debate remains surrounding the actual mechanisms of such encoding and their effects on retrieval.

A standard critique regarding the elaborative hypothesis is its incompatibility with retrieval-induced forgetting, in which retrieval involves the suppression of non-target items, not the activation of several related items. Additionally, one can argue that the generation of related items may produce interference or cue overload with subsequent retrieval attempts. However, recent findings suggest that encoding processes may be more dynamic in a way that favors elaborative learning– reactivating older memories while encoding new ones significantly reduces memory interference at retrieval. fMRI data has revealed that this reactivation facilitates the integration of overlapping memories during encoding, thus reducing the interference of otherwise competing memories at retrieval (Chanales et al., 2019). Integration of memories during encoding could reduce interference of memories during retrieval in various contexts, including semantic learning.

The Mediator Hypothesis

The mediator hypothesis proposes that post-error learning occurs because the error acts as a secondary cue to retrieve the correct answer. In a paired-associate task, generating a non-target word related to the cue could mediate between the cue and target words (Huelser & Metcalfe, 2012; Mera et al., 2021). Instead of solely using a cue during retrieval, one can retrieve the error from the cue and the target from the error. Referring to the previous example, at subsequent presentations of the word "whale," one may recall their previous error, "swims," and from it, the correct target word, "tail."

This theory finds its strength in an episodic context account of memory retrieval. The episodic context account explains that people encode information about learning events and the episodic and temporal context in which they occur (Howard & Kahana, 2002). This episodic context may be restored during retrieval to facilitate correct recall (Lehman & Malmberg, 2013). In retrieval-based learning, retrieval increases recall because individuals think back to and then reinstate their prior learning contexts (Karpicke, 2017). Interestingly, retrieval practice in a list discrimination task led free-recall test performance to follow the temporal order of words more closely than the semantic relatedness of words (Whiffen & Karpicke, 2017). Such findings demonstrate that retrieval practice has clear implications for search strategies during future recall, specifically in a manner conducive to the temporal and episodic context of the learning event.

Prototype Models of Post-Error Learning

Both hypotheses point to distinct mechanisms of post-error learning. Examining these mechanisms could establish one as superior to the other or support their coexistence since, theoretically, they are not mutually exclusive. One way to do this is with the use of formal computational frameworks. Here, we use Anderson and Schooler's (1991) model, which is now a part of the Adaptive Control of Thought–Rational (ACT-R) cognitive architecture. In this model, memories in declarative memory are represented as chunks in a semantic network. Each chunk *i* has a corresponding base-level activation B_i based on the recency and frequency of its presentation as seen in Eq. (1):

$$B_i = \ln(\sum_{j=1}^{n} t_j^{-d}) \tag{1}$$

where n is the number of presentations of chunk i; t_j is the time since the jth presentation of i; d is the decay parameter reflecting how quickly chunks are forgotten. In addition to this base-level activation, the probability of retrieving i is also a function of spreading activation and noise. Altogether, chunks matching a retrieval request compete for successful retrieval following the formula, seen in Eq. (2)

$$A_i = B_i + \sum_k \sum_j W_{kj} S_{ji} + \varepsilon \tag{2}$$

where the sum of k sums spreading activation across all buffers set to provide it; the sum of j refers to the potential sources of activation that spread to chunks in buffer k; W of kj is the weight or amount of activation spread from source j to chunk i; S of ji is the strength of association from source j to chunk i. Lastly, ε reflects activation that makes the retrieval process non-deterministic. ACT-R accurately models forgetting and errors to predict real-world response times and accuracies using these formulas. In this way, ACT-R can comprehensively model a paired-associate task, producing results that closely fit human behavioral data (Anderson, 1981; Anderson & Reder, 1999; Pavlik & Anderson, 2005). Modifying components within ACT-R can produce models that test both hypothesized mechanisms of post-error learning.

Spreading activation in ACT-R can be used to properly model elaborative encoding via error commission in a paired-associate task. ACT-R's declarative memory module is formatted in such a way that includes the semantic relationships between chunks. Within declarative memory, single words link to various associates in a lexical semantic network. When an error is committed, and feedback is provided, chunks linking the cue and target words could be merged with chunks containing the cue and error words to form one elaborative chunk. In the previous paired-associate example, *whale-tail* would be merged with *whale-swims* to create a chunk: *whale-tail-whale-swims*. This chunk could represent the previously discussed meaningful links between cue and target words (i.e., the whale swims with its tail) or simply *whale-tail*, not *whale-swims*. Subsequent presentations of the cue spread more activation to this elaborative chunk; multiple references of the cue word within the chunk increase their strength of association (Figure 1). Overall, this elaborative encoding of the error alongside the cue and target increases its activation, specifically through spreading activation.

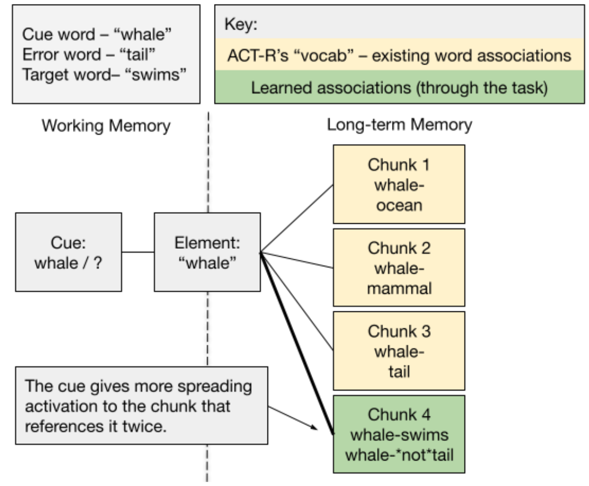

Figure 1: Using spreading activation to model elaborative error learning.

In addition to the declarative module, ACT-R uses a procedural module to articulate cognitive steps (Anderson et al., 2004). The mediator hypothesis relies on remembering the error itself, suggesting that a cognitive process occurs when remembering and recalling an error. Thus, a production rule that checks for an error can model a mediator explanation of post-error learning (Figure 2). If a previous error commission is detected or remembered, an additional production can fire to retrieve the error and use it as a secondary cue.

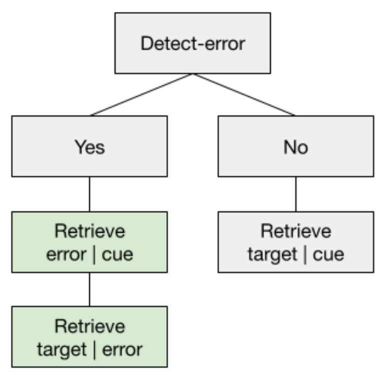

Figure 2: Using an additional production rule to model mediator error learning.

Theoretical Predictions Derived from the Models

It is possible to derive ordinal predictions from these models. Both models predict that error items would be retrieved better than normal study items. In the case of the elaborative model, this is due to the additional spreading activation and, in the case of the mediator model, to the existence of two retrieval routes. Thus, both models predict that error items would be associated with greater response accuracy.

The two models, however, make *opposite* predictions about the relative response times associated with study and error items, respectively. In ACT-R, response times depend on a number of factors, including non-retrieval times spent on perceptual and motor processes, indicated as T_{ER}, and the retrieval time associated with an item i, indicated as $R(i)$. Thus, in general, the response time for the i-th item is:

$$RT = T_{ER} + R(i) \tag{3}$$

In turn, $R(i)$ depends on the activation $A(i)$ of item i, which is the sum of its base-level and spreading activation. Specifically, retrieval times are related to activation by the equation (Anderson et al., 2004):

$$R(i) = k\, e^{-A(i)} = k / e^{A(i)} \tag{4}$$

Where k is another individual-specific parameter that scales the retrieval latency.

Both models assume that all study items have been encoded in the same session and practiced the same number of times, so they have comparable activations. The two models make different predictions for the times to retrieve an error item e.

In the *elaborative* model, the additional information encoded in the error item provides additional spreading activation, which sums up the global activation of the error item $A(e)$. We will indicate this additional activation as $S(e)$ so that $A(e) = A(i) + S(e)$. So, the retrieval time for an error item $R(e)$ is:

$$\begin{aligned} R(e) &= k\, e^{-[A(i)+S(e)]} \\ &= k / e^{[A(i)+S(e)} \\ &= [k / e^{A(i)}] / e^{S(e)} \\ &= R(i) / e^{S(e)} \end{aligned} \tag{5}$$

Note that, because $S(e) > 0$, $e^{S(e)} > 1$, and thus $R(e) < R(i)$.

According to the *mediator* hypothesis, error items do not differ in terms of activation but in terms of retrieval attempts. That is, on a fraction f of trials involving error items, participants would first retrieve an incorrect target, then they would detect the error, and finally retrieve the correct item. Both the correct and incorrect items would have comparable activation levels and thus take approximately the same retrieval time as a study item, $R(i)$. Thus, if we indicate the fraction of trials f in which an error is retrieved, we obtain

$$R(e) = (1\text{-}f)R(i) + f[R(i) + R(i)] = (1+f)R(i) \tag{6}$$

Because $0 < f < 1$, response time will be *longer* for error items, with the specific amount depending on f.

Thus, although both models leave much room for individual differences across participants (due to differences in the T_{ER}, $S(e)$, k, and f parameters), the models make clearly opposite predictions about the relative time to respond to study and error items.

Experimental Predictions

Based on the previous theoretical analysis, we can make the following predictions. Firstly, we expect to confirm the results of previous pretesting research (Huelser & Metcalfe, 2012; Kornell et al., 2009). That is, participants should perform better on error generation items compared to study items on the final test of our first experiment.

Additionally, we expect to find a difference in response times on the final test between conditions. However, we are unsure about the directionality of this difference. Longer reaction times in the error condition suggest the majority of participants are learning from errors in a mediator method. Shorter reaction times in the error condition suggest the majority of participants are learning from errors in an elaborative method. Unclear results may reflect a combination of the two methods. People may learn from errors by combining both mechanisms or there could be individual differences in how people learn from errors (i.e., mediator learners vs elaborative learners).

Materials and Methods

Participants

University of Washington undergraduate students ($N = 61$) were recruited on a rolling basis over the course of a quarter for the pretesting task and provided with course credit for their participation.

Pretesting Task

To replicate Huelser and Metcalfe (2012), 60 weakly related word pairs were selected from Nelson, McEvoy, and Schreiber's (1998) norms. This experiment had three phases: learning, distractor, and final test. In the learning phase, the task randomly interleaved study trials and test trials. In study trials, the cue word (e.g., "whale") and its corresponding target ("tail") were presented simultaneously on the screen for 10 seconds. On test trials, only the cue word was presented on the screen (e.g., "whale"). Participants were asked to respond by typing what they thought the target word was in a textbox (e.g., "swims"). They were given 5 seconds to respond before they were shown the cue word and correct target word simultaneously for 5 seconds. After learning all pairs once, participants played a visuospatial game for 5 minutes as a distractor to prevent rehearsal of word pairs. Finally, participants took a self-paced final test containing all 60 word pairs.

Measures

The primary measure of post-error learning is accuracy. Analysis for the pretesting task uses only results from the final test. Accuracy is calculated as the number of items to which the participant responded correctly over the total items and will be split by condition (error or study) for analysis. Higher cued-recall accuracy reflects better learning in this paradigm.

The second measure of interest is response time. Response times were calculated as the time it takes participants to respond (first key press) to each test item since the onset of the presentation.

Results

Replicating the Pretesting Effect

On average, participants had higher final cued-recall accuracy on error items (M = 0.70 +/- 0.16) than study items (M = 0.60 +/- 0.18), as seen in Figure 3. To analyze accuracy values, linear mixed models were used to account for variability and individual differences. Specifically, we fitted a mixed model to all of the experimental trials, including the particular trial condition (Study vs. Error) as a fixed effect and the participant-level intercept as a random effect; the latter accounts for individual differences in response accuracies. Because accuracy is a binary variable, the model used a binomial distribution to capture the predicted variable. The model uncovered a large main effect of condition (β = -0.47, SE = 0.08, t = -6.13, $p <$ 0.0001). The complete results of the model are shown in Table 1. These findings confirm the results of previous studies (Huelser & Metcalfe, 2012; Kornell et al., 2009).

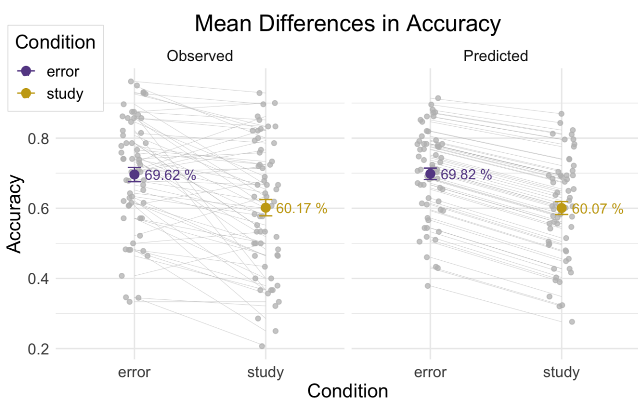

Figure 3: Differences in average final cued-recall accuracy split by condition. Gray dots and lines represent data for individual participants; colored dots and error bars represent means +/- SE for the Error (blue) and Study (red) conditions

Table 1: Results of the Mixed-Level Model for Accuracy

Statistical Test	β estimate	*SE*	*t*	*p*
(Intercept)	0.925***	0.107	8.651	5e-18
Condition	- 0.468***	0.076	- 6.130	8e-10

**p < 0.05 **p < 0.01 *** p < 0.001*

Reaction Times

To remove extreme values from our data, we used a maximum cutoff point of 15000ms and a minimum cutoff point of 200ms. Only correct trials were included in the analysis.

On average, participants had longer response times on error items (M = 4104 +/- 779ms) than study items (M = 3920 +/- 936ms), as seen in Figure 4. The difference between condition response times was compared with a linear mixed model. As in the previous case, the model includes each trial condition (Study vs. Error) as a fixed effect and the participant-level intercept as a random effect; the latter accounts for individual differences in response latencies. Unlike the previous case, the model used a Gaussian distribution to model the dependent variable. Additional random factors, such as random slopes to account for different effects for each participant, did not improve the fit of the model.

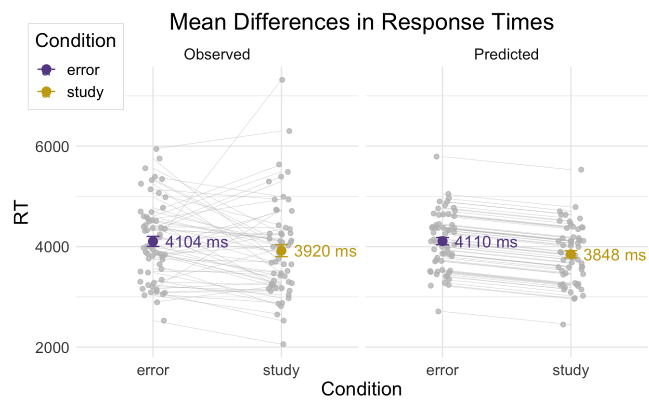

Figure 4: Differences in final cued-recall response times split by condition. Gray dots and lines represent data for individual participants; colored dots and error bars represent means +/ SE for the Error (blue) and Study (red) conditions

The model confirmed a large and significant main effect of condition (β = -255.99, SE = 89.18, t = -2.87, p < 0.005). The complete results of the model are shown in Table 2.

To examine the possibility that different individuals might use different strategies, a second linear mixed model was created, which included the participant-level slope as a random effect. This model allows for different individuals to have either shorter or longer RTs in the error conditions, thus allowing the possibility that some individuals might use an elaborative strategy. This second model replicated the results of the first, finding a significant main effect of the condition (β = -255.99, SE = 89.18, t = -2.87, p < 0.005). An ANOVA test confirmed that the second model does not provide a greater fit than the first ($\chi(3)$ = 0.38, p > 0.94); furthermore, all the fitted slopes in the ensuing model were negative, suggesting that the apparent upward slopes in Figure 4 are due to outlier responses, rather than systematic use of the elaborative model.

Table 2: Results of the Mixed-Level Model for Response Times.

Statistical Test	**□ estimate**	***SE***	***t***	***p***
(Intercept)	4109.70***	103.06	39.877	2e-16
Condition	-261.40**	88.79	- 2.944	0.003

**p < 0.05 **p < 0.01 *** p < 0.001*

Discussion

In this paper, we examined if errors facilitate learning, including an investigation to probe potential underlying mechanisms. The pretesting paradigm reflects the benefit of a retrieval process where participants guess an answer in response to a cue before studying the correct cue-answer pair. This paradigm is commonly used in error learning research. As such, its success in improving learning compared to just studying is well documented (Huelser & Metcalfe, 2012; Kornell et al., 2009). However, research into the underlying cognitive processes facilitating this phenomenon remains speculative. Without understanding how post-error learning works, applications remain restrictive. Revealing underlying mechanisms may allow us to discriminate when errors are beneficial. In this way, current learning and memory models could increase their efficacy by harnessing the power of errors. Our study aims to provide a baseline to research post-error learning mechanisms with the use of cognitive models.

Importantly, we were able to replicate the pretesting results of existing literature; final cued-recall accuracy was higher in the error-generation condition than in the study-only condition. This not only helps to confirm the benefit of retrieval attempts before study opportunities but advocates for further research into the mechanisms of this process.

Different reaction time hypotheses arose from our two post-error learning models based on the ACT-R architecture. First, an elaborative model predicts that error learning results in quicker response times on subsequent tests. This is because elaboration works through spreading activation, adding activation to the correct answer which speeds up retrieval and response times. Alternatively, a mediator model predicts that error learning results in slower response times on subsequent tests. Mediation uses an extra step to retrieve an error as a secondary cue to get the correct answer. Although this procedure increases accuracy, it also costs extra time, resulting in longer response times. Results from the current study demonstrate that average response times are longer on error items than on study items. This supports the mediator hypothesis of post-error learning. However, it is important to note that at the individual participant level, there are cases in which average reaction times on study items are longer than on error items. This could mean that error learning mechanisms differ on an individual level. Future research should attempt to distinguish between mechanisms by focusing on individual differences. Additionally, the ability to learn from errors itself may be highly individualized. Focusing on individual differences to look at (1) who is learning from errors, and (2) how they are doing it can better specify post-error learning and its applications to learning environments.

The most notable limitation of this study is the generalizability of the error-learning paradigm. Errors as we think of them in the real world are often committed after a study opportunity. Looking not only at retrieval errors, but encoding errors, in error learning would broaden the applications to more settings. Overall, it's important to extend error learning research to paradigms that involve real-world errors. Another limitation of this paradigm is it does not look at memory over longer periods of time. Testing memory over days, weeks, and months, may reveal stronger benefits of errors. Deeper encoding processes resulting from error commission may lead to facts that are more resistant to decay and forgetting. Alternative paradigms must be designed and used to further establish post-error learning beyond retrieval practice. An example of such paradigm is the adaptive fact-learning system developed by Sense and van Rijn (2022; Sense et al., 2016), in which new paired associations are presented at a pace that is individualized to each participant to optimize retention. Importantly, their paradigm internally makes use of ACT to model each individual's memory, and yields highly reliable estimates of each individual and each item's decay rate (Sense et al., 2016). A modification of this paradigm that includes an error-generating phase provides important information as to whether, for example, error items are forgotten at lower speeds, rather than (or in addition to) having additional retrieval routes.

Models are unique in their ability to reconceptualize behavioral results. By decoding human behavior, models begin to reveal cognition by stabilizing the messiness of data. As such, the proposed cognitive models in this paper can help identify mechanisms of post-error learning. These models could distinguish different learners from one another and propose ways to manipulate post-error learning by targeting the relevant cognitive processes. Moreover, these findings could extend to fields outside of cognitive psychology, advocating for the benefit of making mistakes in various educational settings and assisting in developing AI and machine learning advancements that update comprehensive feedback histories with each new learning experience.

References

Anderson, J. R. (1981). Interference: The relationship between response latency and response accuracy. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 326–343.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261–295.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. Psychological science, 2(6), 396-408.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review, 111*(4), 1036.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, *128*(2), 186–197.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491–1494.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569.

Ceraso, J. (1967). The interference theory of forgetting. *Scientific American*, *217*(4), 117–127.

Chanales, A. J. H., Dudukovic, N. M., Richter, F. R., & Kuhl, B. A. (2019). Interference between overlapping memories is predicted by neural states during learning. *Nature Communications*, *10*(1), 5363.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Cyr, A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 841–850.

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9*.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, *40*(4), 514–527.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, *21*(1), 45–83.

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In *Learning and Memory: A Comprehensive Reference* (pp. 487–514). Elsevier.

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 106–114.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998.

Kornell, N., & Vaughn, K. E. (2016). Chapter five - How retrieval attempts affect learning: A review and synthesis. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 65, pp. 183–215). Academic Press.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*, *120*(1), 155–189.

Mera, Y., Rodríguez, G., & Marin-Garcia, E. (2022). Unraveling the benefits of experiencing errors during learning: Definition, modulating factors, and explanatory theories. *Psychonomic bulletin & review*, *29*(3), 753–765.

Nelson, D., McEvoy, C., & Schreiber, T. (1998). *University of South Florida Free Association Norms*.

Pavlik Jr, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation based model of the spacing effect. *Cognitive science*, *29*(4), 559-586.

Roediger, H. L. I., & Karpicke, J. D. (2010). Intricacies of spaced retrieval: A resolution. In *Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork* (pp. 23–47). Taylor & Francis Group.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498.

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science, 8*(1), 305-321.

Sense, F., & van Rijn, H. (2022, January 27). Optimizing Fact-Learning with a Response-Latency-Based Adaptive System.

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1036–1046.