MEMORY
2018, VOL. 26, NO. 9, 1281–1290
https://doi.org/10.1080/09658211.2018.1464189

Routledge
Taylor & Francis Group

Learning from your mistakes: does it matter if you're out in left foot, I mean field?

Andrée-Ann Cyr[a] and Nicole D. Anderson[b]

[a]Department of Psychology, York University, Toronto, Canada; [b]Rotman Research Institute, Baycrest Health Sciences and Departments of Psychiatry and Psychology, University of Toronto, Toronto, Canada

ABSTRACT

Studies have shown that generating errors prior to studying information (*pencil*–?) can improve target retention relative to passive (i.e., errorless) study, provided that cues and targets are semantically related (*pencil–ink*) and not unrelated (*pencil-frog*). In two experiments, we manipulated semantic proximity of errors to targets during trial-and-error to examine whether it would modulate this error generation benefit. In Experiment 1, participants were shown a cue (*band–?*) and asked to generate a related word (e.g., *drum*). Critically, they were given a target that either matched the semantic meaning of their guess (*guitar*) or mismatched it (*rubber*). In Experiment 2, participants studied Spanish words where the English translation either matched their expectations (*pariente–relative*) or mismatched it (*carpeta–folder*). Both experiments show that errors benefit memory to the extent that they overlap semantically with targets. Results are discussed in terms of the retrieval benefits of activating related concepts during learning.

ARTICLE HISTORY
Received 14 November 2017
Accepted 8 April 2018

KEYWORDS
Learning; memory; retrieval; generation; errors

A consequence of deep and meaningful learning is that it triggers ideas related to the information that we are studying. For instance, when asked the question "*What is the capital of Canada?*" we not only activate the correct response, Ottawa, but also a wide array of concepts that are related to the answer either semantically (e.g., Vancouver) or to personal experience (e.g., *I once went ice skating on the Ottawa canal*). While these related memories may not be always directly relevant to the question, their activation can facilitate retrieval of sought after information by highlighting pathways for the search, and scaffolding retrieval (Anderson & Bower, 1974). This elaborative nature of retrieval is purported to be at the heart of retrieval-based learning strategies known to enhance memory (Carpenter, 2009; but also see the episodic context account of retrieval-based learning by Karpicke, Lehman, & Aue, 2014), such as the testing effect, i.e., the finding that memory for studied information is enhanced by a retrieval relative to restudy opportunity (see Karpicke & Roediger, 2008). For instance, when attempting to retrieve target information during word pair learning (*bread–?*), individuals generate mediators (*crust, butter*), and the ability to re-activate these related concepts on a later test is associated with better memory for the target (*basket*) (Carpenter, 2009; Pyc & Rawson, 2010). Thus, recalling the concepts that sprung to mind during a previous retrieval attempt can make target information more accessible on later retrieval attempts.

The benefits of semantic elaboration are also exemplified by studies on error generation effects in episodic memory. This literature examines the memorial benefits of retrieval prior to studying novel information, in contrast to the testing literature which considers the benefits of retrieval following initial study. Error generation is typically explored by contrasting memory performance following trial-and-error learning, where participants must guess what the target is in response to a cue (*bread–?*) prior to seeing the target (*bread–basket*), and errorless learning, where the correct cue-target pair is studied in full from the onset (*frog–pond*). Numerous studies have found improved memory for targets following trial-and-error relative to errorless learning for word pairs (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Kornell, Hays, & Bjork, 2009; Metcalfe, 2017; Vaughn & Rawson, 2012), and also for more pedagogical materials such as facts (Kang et al., 2011; Richland, Kornell & Kao, 2009; Kornell, Klein, & Rawson, 2015; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Pressley, Tanenbaum, McDaniel, & Wood, 1990). Our own work extends this advantage to healthy older adults (Cyr & Anderson, 2012, 2015), suggesting that active generation of conceptual errors can afford benefits akin to other deep encoding strategies known to minimise age-related declines in episodic memory (Luo, Hendriks, & Craik, 2007).

One theory of the benefits of trial-and-error learning advances that generating guesses enriches encoding by forging semantic connections and orienting learners to

CONTACT Andrée-Ann Cyr cyrandre@glendon.yorku.ca Department of Psychology, York University, Glendon Campus, 2275 Bayview Avenue, Toronto, ON, M4N 1J8, Canada

the relationship between the cue and the target (Cyr & Anderson, 2012; Huelser & Metcalfe, 2012; Kornell et al., 2009). This "elaborative retrieval hypothesis" of error generation effects is bolstered by the fact that wrong guesses are not found to boost memory when targets do not share a pre-existing semantic relationship with the cue (e.g., *bread–cloud*) (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012). Similarly, researchers have found no benefit of error generation when learners are forced to guess at fictional trivia questions (Question: *Who is the bouncy and egotistical friend of Kenny Peters?* Answer: *Albert*; Kornell et al., 2009, Experiment 2) and obscure factual questions wherein individuals have no clue as to the answer (Question: *Where is Disko Island?* Answer: *Greenland*; Kang et al., 2011). In sum, it appears that in order to be advantageous to episodic memory, errors must be semantically or conceptually informed by the cue.

In the context of everyday learning, however, our wrong guesses are typically more or less educated as opposed to total shots in the dark. While it is clear from the literature that "shots in the dark" are not helpful when learning information, it is unclear whether guesses that are educated but "out in left field" are as helpful as those that are "near misses". To illustrate, imagine a scenario where two students are asked the question "*Who is Justin Trudeau?*" and both respond incorrectly with the same level of confidence before being told the answer (*Prime Minister of Canada*): Student A has a vague sense that the name belongs to a political figure, and lands on "Prime Minister of France" based on the French sounding last name. Student B recognises that the name refers to someone famous and guesses him to be a former member of the American pop group NSYNC, confusing him with the singer Justin Timberlake. Both students are making educated guesses based on preexisting knowledge, but Student A is much closer conceptually to the correct answer than Student B. Which student is more likely to remember the correct answer on a later test? If errors are helpful to the extent that they overlap with targets in terms of cue-relevant features, as the elaborative retrieval hypothesis intimates, Student A should benefit most. However, a parallel literature suggests that Student B may attend to the feedback more than Student A as it would strike her as more surprising (e.g., Fazio & Marsh, 2009) or discrepant from her guess (Rescorla & Wagner, 1972) (e.g., "*Wow, I was on the wrong track*"). To date, no study has investigated error generation benefits as a function of the *degree* of conceptual similarity between errors and targets, an important question given that most learning occurs on this sliding scale.

In this study, we contrasted the effects of making mistakes that are "near misses" and "out in left field" on memory for target information. To do this, we manipulated the semantic distance between participant-generated guesses and targets during learning of cue-target pairs wherein the target was arbitrary and related to the cue

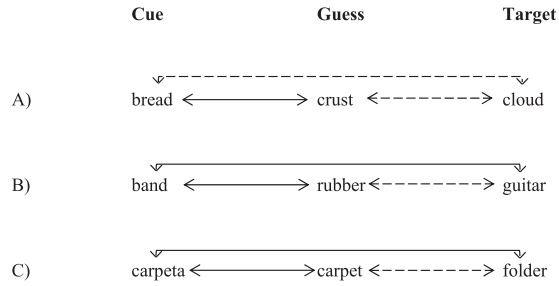

Figure 1. Schematic representation of the relationship among cues, errors and targets.

(Experiment 1) or represented a non-arbitrary answer (Experiment 2). Unlike past studies which have contrasted memory for related and unrelated cue-target pairs, we kept cue-target relatedness unchanged and instead varied the distance of guesses to targets (see Figure 1). We predicted that target memory would be better following trial-and-error relative to errorless learning, and that the benefits of trial-and-error learning would be enhanced when errors are semantically close rather than far from the target, consistent with the elaborative retrieval hypothesis.

Experiment 1

In this experiment, we used cues that are homographs, i.e., words associated with more than one meaning, e.g., *band*: a music group or a binding object. In the trial-and-error condition, participants generated a guess (*concert*) and were shown a target that was always related to the cue, but was either related (*guitar*) or unrelated (*elastic*) to the meaning of the generated error. Relative to errorless learning, we predicted that errors would afford greater increases in target memory when they matched the semantic meaning of the target (*concert* and *guitar*) on account of greater overlap and semantic integration relative to mismatched errors (*concert* and *elastic*). We also asked participants to recall their wrong guess along with the target at cued recall to examine the relationship between memory for errors and targets. Previous studies have found that remembering one's error is predictive of target memory (Knight et al., 2012; Yan, Yu, Garcia, & Bjork, 2014), provided that errors and targets share a conceptual relationship (Cyr & Anderson, 2015). Therefore, we predicted that memory for prior guesses would beneficially mediate target recall to a greater extent in the match relative to mismatch condition, reflecting the facilitative effects of semantic integration.

Method

Participants

Thirty two adults were recruited from the Baycrest research participant pool or responded to advertisements posted at Glendon Campus, and were paid for their participation.

One person's data were discarded due to a computer error so we report the data of 31 participants. To be eligible, individuals had to be free of any psychiatric, neurological, or medical condition known to affect cognition. Participants had a mean age of 23.66 ($SD = 3.33$) and had completed an average of 15.34 ($SD = 2.66$) years of formal education. All participants reported to have learned English before the age of 5 and the mean vocabulary score on the Shipley Institute of Living Scale (Shipley, 1940) was 29.06 ($SD = 4.38$). We measured levels of anxiety and depression by means of the The Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983): The mean anxiety score was at the clinical cut-off for possible anxiety[1] ($M = 8.00$; $SD = 4.61$) while the mean depression score was below the clinical cut-off ($M = 3.63$; $SD = 2.74$).

Materials

Sixty four homographs were selected from the Alberta homograph meaning frequency norms (Twilley, Dixon, Taylor, & Clark, 1994) based on the following criteria: a) the homograph was a noun; b) the homograph was predominantly associated with two meanings only; and c) the homograph's primary and secondary meaning were nouns. For each homograph (e.g., *port*), we then selected two associates as possible targets for each the primary (1: *boat*; 2: *dock*) and secondary (1: *wine*; 2: *brandy*) meaning. This pool of 64 homographs was then divided into four sets of 16 pairs. Word frequency and word length did not differ as a function of set or word position, Fs < 1. There were also no differences across sets in terms of the proportion of normed responses to the homograph for the primary and secondary meanings, $F(3,60) = 1.64$, $p = 0.190$, $\eta_p^2 = 0.08$. Next, a Latin-square was used to assign two sets to the errorless condition and two sets to the trial-and-error condition. For both sets, we counterbalanced whether homographs in the first and second half of each set were assigned a primary or secondary meaning target (e.g., *boat* or *wine*). We then randomised the order of the word pairs. Therefore, under both errorless and trial-and-error learning each participant was assigned 16 homographs paired with a primary meaning target, and 16 homographs paired with a secondary meaning target (i.e., 32 word pairs were studied in each of the errorless and trial-and-error conditions). For the trial-and-error sets, we counterbalanced whether homographs in the first and second half of each set were assigned to be Match or Mismatch trials, and randomised the order of the word pairs. As such, under trial-and-error learning participants studied 16 homographs paired with a target that matched their guess (Match condition) and 16 homographs paired with a target that did not match their guess (Mismatch condition). The assignment of which word (position one or two) within the selected meaning (primary or secondary) would be selected as an intended target was counterbalanced across participants. For example, if selecting a target for the homograph port from the primary meaning (i.e., 1: *boat* or 2: *dock*), the experimenter would provide *boat* if the assigned target word position was one, and *dock* if it was two.

Design and procedure

This was a within-subjects study design with learning instructions and semantic proximity as independent variables. Participants were tested individually, and stimuli presentation and recording used E-prime software (Psychology Software Tools, Inc., Pittsburgh, Pennsylvania).

Each individual underwent 2 blocked study-test cycles, one errorless and one trial-and-error, each consisting of 32 different homograph word pairs. Order of learning instruction (errorless and trial-and-error) and counterbalanced across participants. In errorless learning, the experimenter presented the homograph on the screen (*calf*) and then immediately displayed the target (*ankle*). In trial-and-error learning, participants were shown a homograph (*pitcher*) and were prompted to guess the target word ("*Is it* baseball?"). Generation of guesses was self-paced. If it was a Match trial, the experimenter would provide a target from the same meaning as their guess ("*No, the target word is* catcher."). If it was a Mismatch trial, the experimenter would provide a target from the other meaning ("*No, the target word is* lemonade."). In other words, if they generated a word related to the primary meaning of the homograph, the target would be another word related to the primary meaning in a Match trial, and a word related to the secondary meaning in a Mismatch trial. Likewise, if they generated a word related to the secondary meaning, another word related to the secondary meaning was presented if it was a Match trial, and a word related to the primary meaning was presented if it was a Mismatch trial. For both errorless and trial-and-error learning, participants were instructed to commit to memory the correct target words for a later memory test. In trial-and-error learning, the experimenter always selected a target word that was not generated as a guess by the participant. This was done by assigning each word from the primary or secondary meaning to numbers one or two in E-Prime: If the participant guessed the word in position number one, and one was their assigned target word number (counterbalanced across participants), the experimenter pressed key number two to display a word that had not been generated. All targets were shown with the homograph (*calf–ankle*) for four seconds, followed by a one second inter-stimulus interval. Participants were asked to write down the targets as they were shown, making sure to cover their responses with another sheet of paper as they went along to discourage rehearsal. A 10-minute break followed the study phase during which the HADS and the Mini-Mental State Examination was administered. Once the break had expired, participants began the cued recall task. For the cued recall task participants were shown the studied homographs and asked to type in responses in succession: First, the wrong guess

they had generated at study (if it was a trial-and-error block), and second, the correct target. There were no new homographs, and the cued recall task was self-paced. Following both study- test cycles, participants were debriefed and compensated. The entire study lasted approximately 60 min, and was approved by the York University, University of Toronto, and Baycrest Research Ethics boards.

Results

An alpha level of 0.05 was used for all statistical tests.

We first wanted to determine whether target cued recall performance would vary as a function of whether it was associated with the normative primary or secondary meaning of the homograph cue. A within-subjects ANOVA on proportion correct recall in the Errorless condition showed that primary targets ($M = 0.61$; $SD = 0.20$) were better remembered than secondary targets ($M = 0.52$; $SD = 0.22$), $F(1, 30) = 8.19$, $p = 0.008$, $\eta_p^2 = 0.21$. We could not conduct this analysis for Trial-and-error targets given that a participant's guess is necessarily associated with their personal primary meaning, regardless of what the norms indicate. As such, Match and Mismatch targets are necessarily associated with an individual's primary and secondary meanings respectively, regardless of the norms. Rather, we examined whether the meaning of participant generated errors (primary or secondary) varied as a function of Trial-and-error condition. The mean probability of generating a guess associated with the primary meaning was 0.70 ($SD = 0.27$), but the proportion of generated guesses that were related to the primary meaning of the homograph did not vary as a function of Trial-and-error condition (Match or Mismatch), $F < 1$, $\eta_p^2 < 0.001$.

Next, we sought to examine whether we had replicated the error generation benefit, i.e., whether targets studied via trial-and-error learning would be better remembered than those studied via errorless learning. A repeated measures ANOVA with within-subjects factor of study condition (Errorless learning vs Trial-and-error learning collapsed across Match and Mismatch conditions) revealed an overall advantage in cued recall performance for Trial-and-error targets ($M = 0.71$; $SD = 0.18$) over Errorless targets ($M = 0.54$; $SD = 0.19$), $F(1, 30) = 45.89$, $p < 0.001$, $\eta_p^2 = 0.61$.

Trial-and-error cued recall performance was then analyzed using a repeated measures ANOVA with within-subjects factors of semantic proximity (Match vs Mismatch) and recall type (Target vs Guess). Refer to Figure 2 for means of cued recall accuracy. Results indicated that the main effect of semantic proximity was significant, $F(1, 30) = 10.67$, $p = 0.003$, $\eta_p^2 = 0.26$, indicating that cued recall accuracy was greater for the Match relative to Mismatch pairs. The main effect of recall type was also significant, $F(1, 30) = 47.99$, $p < 0.001$, $\eta_p^2 = 0.62$, revealing that guesses were better remembered overall than target items. Finally, the interaction was significant, $F(1, 30) = 10.61$, $p = 0.003$, $\eta_p^2 = 0.26$: Breaking down the interaction, we found that Match targets were better recalled than Mismatch targets, $F(1, 30) = 13.68$, $p = 0.001$, $\eta_p^2 = 0.31$, but that memory for past guesses did not vary as a function of semantic proximity, $F(1, 30) < 1$, $\eta_p^2 = 0.01$. In other words, participants' wrong guesses were equally accessible across Match and Mismatch conditions.

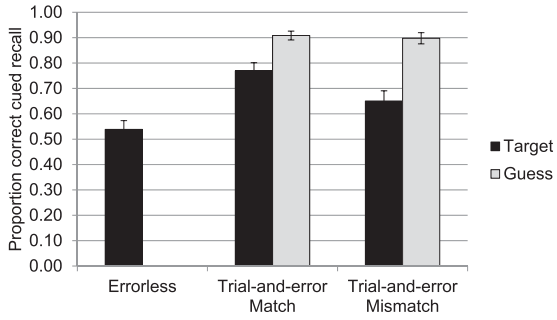

Figure 2. Cued recall performance for targets and guesses as a function of learning condition and semantic proximity in Experiment 1 (bars represent standard error of the mean).

We also sought to examine mediator decoding, i.e., whether memory for Trial-and-error targets (e.g., *pitcher–lemonade*) was mediated by memory for the associated prior guess, and whether its semantic proximity to the target would modulate this relationship (e.g., Match guess: *beer* vs Mismatch guess: *baseball*) (for a similar analysis, see Knight et al., 2012). For each participant, we selected the trials where the prior guess was correctly remembered and computed whether or not the target was subsequently retrieved. This would determine whether remembering one's error is more or less predictive of target recall as a function of semantic proximity. We conducted a repeated measures ANOVA to compare the effects of semantic proximity (Match, Mismatch) on the proportion of correct target recall, given that the guess was accurately recalled. This revealed a significant effect, $F(1, 30) = 15.78$, $p = 0.001$, $\eta_p^2 = 0.35$, showing that target memory was higher if it was preceded by recall of a Match ($M = 0.79$; $SD = 0.16$) relative to a Mismatch guess[2] ($M = 0.65$; $SD = 0.22$).

Discussion

The results of Experiment 1 show not only an error generation benefit, but a clear enhancement of this effect when errors and targets belong to similar relative to disparate semantic families. This is consistent with our hypothesis that proximal (e.g., *band*–(concert)–*guitar*) relative to distant (e.g., *pitcher*–(baseball)–*lemonade*) errors are better integrated in service of episodic memory. Moreover, this integration was reflected in an error mediation analysis: Successful recall of match guesses (*concert*) was more predictive of subsequent target memory than mismatch guesses (*baseball*). This is consistent with previous studies (Cyr & Anderson, 2015; Yan et al., 2014) which have

found that targets and errors are more likely to be retrieved together when the error generation benefit is present. A caveat is that we cannot say whether or not this error mediation was spontaneous because participants were explicitly prompted to retrieve their guess during the trial-and-error test. Due to the fact that errorless and trial-and-error performance was measured in separate study-test blocks, participants did not have to remember whether or not they had guessed; rather, on trial-and-error recall trials, they were told that they had generated a guess and were asked to retrieve it. However, mediation accounts of the error generation benefit have been supported in studies using a blocked design where trial-and-error recall trials were explicitly identified (Cyr & Anderson, 2015) and in intermixed designs where participants had to determine whether or not they had guessed (Knight et al., 2012; Yan et al., 2014).

We also found that errorless targets associated with the primary meaning were better remembered than those associated with the secondary meaning. This is reasonable given that during retrieval, both target memory and the semantic network surrounding the cue are activated: When both overlap, as is likely to be the case on primary meaning trials, recall is facilitated. Still, the facilitation afforded by primary meaning cannot explain why error generation would enhance target memory for matched but not mismatched targets.

The use of related word pairs in Experiment 1 has many methodological advantages: It ensures that participants generate the same number of errors and that preexisting knowledge minimally influences learning success. A setback of this paradigm, however, is that what constitutes an error is entirely arbitrary: A guess may be far from a target without necessarily violating the learner's expectations (*baseball* and *lemonade* are equally acceptable targets for the cue *pitcher*). In everyday learning contexts, there is typically a true answer to a question, and faraway guesses signify that you are off track in a meaningful sense. Another feature of everyday learning is that individuals are rarely required to retrieve their mistakes along with correct answers. In Experiment 2, we sought to replicate and extend the findings of Experiment 1 using a more realistic learning situation where the target is intrinsic to the cue, and where participants are not asked to retrieve their wrong guesses at recall.

Experiment 2

We had non-Spanish speaking participants study Spanish-English word pairs under trial-and-error and errorless learning conditions. To manipulate error-target proximity, we selected Spanish words that had been previously identified as false cognates, i.e., words that resemble an English word but mean something different. Some cognates are known as "false friends" (*carpeta*) because they strongly resemble an English word (*carpet*) but signify something very different (*folder*): Given that individuals are likely to generate a guess that is in line with the apparent meaning of the cue and not the target, these pairs served as our Mismatch condition. Other cognates are labelled "unreliable friends" (*carrera*) because while they also strongly resemble an English word (*career*), they differ less in terms of meaning (*degree*): Given that individuals are likely to generate a guess that aligns with the apparent meaning of the cue (and therefore semantically proximal to the target), these pairs served as our Match condition.

A trade-off in using learning materials which have intrinsic answers is that performance is subject to item selection effects: It is easy to select the trial-and-error study trials where targets were successfully retrieved, but such retrievable trials are impossible to identify under errorless learning (Pashler, Zarow, & Triplett, 2003), e.g., a participant studying word pairs in the errorless condition may have correctly guessed that *pariente* means *relative* if given the chance. It is also possible that some correct answers exist in participants' memories as marginal knowledge before the experiment (c.f., Berger, Hall, & Bahrick, 1999), and that they will be relearned easily when the answers are shown in either learning condition. Despite these pitfalls, we believe it is important to explore error generation effects using more complex materials that better approximate pedagogically motivated learning.

Method

Participants

As in Experiment 1, eligible participants had to be free of any psychiatric, neurological, or medical condition known to affect cognition. Thirty-two participants were recruited through the York Research Participant Pool and posters displayed throughout Glendon Campus. To be eligible for the study, participants could not have any knowledge of Spanish or have taken any Spanish courses in the past. The average age was 19.45 ($SD = 2.39$) and the average years of formal education was 17.38 ($SD = 3.17$) years. Participants scored an average of 28.90 ($SD = 4.09$) on the Shipley test (one of the participants' Shipley test was lost), and 8.68 ($SD = 3.37$) and 4.29 ($SD = 2.58$) on the anxiety and depression subscale of the HADS respectively.

Materials

Sixty four Spanish- English word pairs were sourced from various Internet websites. This pool of 64 word pairs was composed of 32 Spanish words identified by translators as false cognates (e.g., *carpeta*: confused with *carpet* but meaning *folder*). There are currently no norms available that quantify the strength of the relationship between the apparent and true meanings of false cognates, but it is widely acknowledged that some are more misleading (i.e., so-called false friends; *carpeta–folder*), while others are less misleading (i.e., so-called unreliable friends; *carrera–degree*). For the sake of clarity and consistency

with the terminology used in Experiment 1, we will be foregoing the parlance of translators and referring to unreliable friends as Match pairs, and false friends as Mismatch pairs. To create sets of Match and Mismatch pairs, we calculated the semantic proximity of apparent translations (*carrera*) to the true translation (*degree*) for each of the 64 Spanish cognates. We used an index of the semantic distance of each apparent meaning to the true meaning as derived by latent semantic analysis (LSA; Landauer, Foltz, & Laham, 1998; http://lsa.colorado.edu). LSA is a method for quantifying the similarity between words on the basis of statistical analyses of a large corpus of text. We used the topic space of "general reading up to first-year college (300 factors)" and "term-to-term comparison type". We calculated the LSA-derived semantic similarity value between apparent and true meaning for each Spanish cognate, where a higher score signifies greater semantic similarity. The 32 Spanish cognates with the highest LSA-derived values were assigned to the Match condition, and the 32 cognates with the lowest values were assigned to the Mismatch condition. The Match pairs had significantly higher LSA-derived values ($M = 0.22$; $SD = 0.06$) relative to the Mismatch pairs ($M = 0.06$; $SD = .07$), $t(62) = 6.11$, $p < 0.001$.

Each set of Match and Mismatch pairs was then divided into two sets of 16 pairs. Word frequency and word length did not differ as a function of set, $Fs < 1$. Next, a Latin-square was used to assign one set of both the unreliable and false friends to each the errorless and the trial-and-error condition. We then randomised the order of the word pairs. In sum, under both errorless and trial-and-error learning, each participant was assigned 16 Spanish-English word pairs from the Match set, and 16 Spanish-English word pairs from the Mismatch set.

Design and procedure

This was a within-subjects study design with learning instructions and semantic proximity as independent variables. Participants were tested individually on computers equipped with E-prime software.

Prior to beginning the experimental task, participants were administered the HADS. Each individual then underwent 2 study-test cycles of blocked errorless and trial-and-error learning, each consisting of 32 different Spanish-English word pairs. Order of learning instruction (errorless and trial-and-error) was counterbalanced across participants. In errorless learning, the full Spanish-English word pair was shown for 4 seconds. In trial-and-error learning, participants were shown the Spanish word only (*lectura*) and were prompted to guess what the one-word English translation was and type it in (generation was self-paced). Once they had submitted their response, they saw the correct Spanish-English word pair for 4 seconds (*lectura–reading*). For both errorless and trial-and-error learning, participants were instructed to commit to memory the correct translations for a later memory test. The study phase was followed by a 10-minute break during which they played Tetris (a visuospatial game). For the cued recall task, participants were shown the studied Spanish words and asked to type in the correct target. There were no new words, and it was self-paced. Following both study-test cycles, participants were debriefed and compensated. The entire study lasted approximately 60 min, and was approved by the York University Research Ethics board.

Results

An alpha level of 0.05 was used for all statistical tests.

We first examined whether our manipulation was successful, i.e., that Mismatch pairs produced guesses that were semantically farther from targets relative to Match pairs. We calculated the LSA-derived value of semantic similarity between each guess generated by participants in the trial-and-error condition and its corresponding target. We then conducted a within-subjects ANOVA, which revealed a significant main effect of Cue Type, $F(1,31) = 97.54$, $p < 0.001$, $\eta_p^2 = 0.76$, confirming that guesses generated in response to Match cues ($M = 0.24$; $SD = 0.06$) were significantly closer to targets relative to those invoked by Mismatch cues ($M = 0.12$; $SD = 0.05$).

Participants infrequently guessed the correct answer during trial-and-error learning ($M = 1.22$; $SD = 1.34$). In examining cued recall performance as a function of learning style, we ran two analyses: One where we included all trial-and-error trials, regardless of whether participants guessed the target correctly or not during the study phase, and one where we included only trials where participants guessed wrong during study. Our reasoning was that participants would have been equally likely to correctly estimate the meaning of the Spanish words in the Errorless as in the Trial-and-Error condition, and that showing both would speak to potential selection effects in our results (for similar reasoning, see Potts & Shanks, 2014).

Cued recall performance: all trials

We conducted a 2(Learning Condition: Errorless; Trial-and-error) × 2(Cue Type: Match; Mismatch) within-subjects ANOVA, where the dependent variable was proportion correct cued recall. This revealed a non-significant main effect of Learning Condition, $F(1,31) = 1.35$, $p = 0.254$, $\eta_p^2 = 0.04$, and a significant effect of Cue Type, $F(1,31) = 27.32$, $p < 0.001$, $\eta_p^2 = 0.47$, revealing that participants were more likely to recall the correct English word on Match ($M = 0.41$; $SD = 0.19$) relative to Mismatch trials ($M = 0.30$; $SD = 0.21$). The Learning Condition × Cue Type interaction was significant, $F(1,31) = 4.20$, $p = 0.049$, $\eta_p^2 = 0.12$, showing that memory for Match pairs benefited more than Mismatch pairs from Trial-and-error learning (see Figure 3). Indeed, the benefit of trial-and-error over errorless learning reached significance for the Match pairs, $F(1,31) = 4.41$, $p = 0.044$, $\eta_p^2 = 0.13$, but not the Mismatch pairs, $F(1,31) < 1$, $\eta_p^2 = 0.01$.

Cued recall performance: only wrong trials

For the following analyses, we excluded all trial-and-error trials where participants guessed the target correctly during the study phase. We conducted a 2 (Learning Condition: Errorless; Trial-and-error) × 2(Cue Type: Match; Mismatch) mixed ANOVA, where the dependent variable was proportion correct at cued recall. There was no main effect of Learning Condition, $F < 1$, $p = 0.859$, $\eta_p^2 < 0.001$, but the Cue Type main effect was significant, $F(1,31) = 24.77$, $p < 0.001$, $\eta_p^2 = 0.44$: Match pairs were better remembered that Mismatch pairs ($M = 0.39$; $SD = 0.18$ and $M = 0.29$; $SD = 0.21$, respectively). The Learning Condition × Cue Type interaction was marginally non-significant, $F(1,31) = 3.58$, $p = 0.068$, $\eta_p^2 = 0.10$, but similar to what we saw when all trials were included.

Finally, we wanted to examine the relationship between accuracy for trial-and-error targets (match and mismatch) on the cued recall test and the semantic proximity of their corresponding errors. We computed the LSA-derived value of semantic similarity between each target and the participant's wrong guess, and binned targets according to whether or not they were recalled during the cued recall test. We then ran a within-subjects ANOVA to compare the values of targets that were correctly recalled and those that were forgotten (Correct; Incorrect). This revealed that recalled targets shared a closer semantic relationship to their associated errors than did targets that were not recalled, $F(1,31) = 29.90$, $p < 0.001$, $\eta_p^2 = 0.49$, ($M = 0.18$; $SD = 0.05$ and $M = 0.13$; $SD = 0.03$, respectively).

Discussion

The aim of Experiment 2 was to examine the effects of semantic proximity between errors and targets on the error generation benefit using a paradigm where there was a non-arbitrary correct answer. The findings partially supported our hypotheses: Memory performance was better overall when there was a strong semantic overlap between errors and targets, and this benefit was particularly apparent in trial-and-error relative to errorless learning (see Figure 3). However, the error generation benefit appeared only for matched targets when correctly guessed trials were included in the analyses, although it neared significance when only wrong trials were considered ($p = 0.068$). As previously mentioned, it is important to consider both analyses given that participants were just as likely to happen upon the correct answer on errorless learning trials. The lack of an error generation benefit for the mismatched pairs (*éxito–success*) suggests that these were more difficult to integrate semantically, and in this sense they may have acted more like unrelated word pairs which are known not to elicit error generation benefits (e.g., Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012).

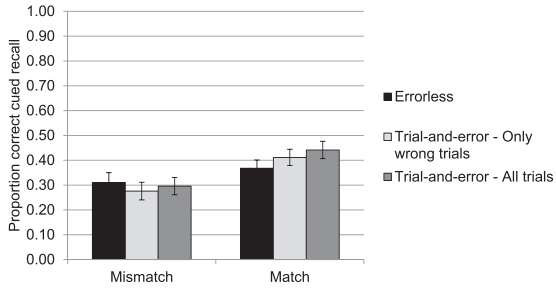

Figure 3. Cued recall performance for targets as a function of learning condition and cue type in Experiment 2 (bars represent standard error of the mean).

General discussion

In recent years, there have been a high number of studies bent on uncovering the benefits of error generation on episodic memory, as well as the underlying mechanisms. These studies have converged in finding that error generation is generally an ally of learning in the laboratory and in the classroom (for a review, see Metcalfe, 2017), which is reassuring to educators as making mistakes is an often unavoidable part of learning. There are boundaries to the error generation effect, however, and exploring these boundaries is critical for generalising its benefits to more complex study materials and different populations. In two studies, we asked whether the benefits of conceptual error generation could be modulated by the semantic proximity between errors and targets. Overall, we found support for our hypothesis that the more semantically similar errors are to correct answers, the better they support episodic memory. However, making such errors only benefited memory over errorless learning in Experiment 1. These findings have implications for current mechanistic theories of error generation.

Before we discuss these implications, we raise, and reject, one possible explanation of the current results. One may attribute the error advantage in memory to differences in processing time at encoding between trial-and-error and errorless learning: Learners are likely to spend more time on trial-and-error relative to errorless study trials given that they need to come up with a guess in addition to processing the feedback. However, researchers have equated the duration of study trials across learning conditions and found that timing does not impact the memorial effects of errors (Guild & Anderson, 2012; Kornell et al., 2009). Recently, Vaughn, Hausman, and Kornell (2017) varied the time given to participants to retrieve an initial response and found that it had no effect on later memory performance. They conclude that it is the initial intensity of the retrieval effort–not the duration–that predicts learning success. In our case in particular, participants spent comparably longer time in the match and mismatch trial-and-error conditions than in the errorless conditions, yet the benefit to later memory was greater in the matched conditions. Thus, we find it highly

unlikely that time on task can explain the error generation benefit.

Implications for theories of the error generation benefit

Retrieval is thought to be beneficial because it increases the activation of related information in memory, and this information can mediate target retrieval later on (Carpenter, 2009, 2011). Applied to error generation effects, this elaborative retrieval hypothesis suggests that mistakes can serve as stepping stones to the right answer. Paradigms used in past studies have contrasted learning of related (*frog–pond*) and unrelated pairs (*bread–cloud*) under errorless and trial-and-error instructions to show that unrelated pairs do not elicit the error generation benefit (Grimaldi & Karpicke, 2102; Huelser & Metcalfe, 2012; Knight et al., 2012). Similarly, in a previous study we had participants study cue-target pairs that were either conceptually (*flower–rose*) or lexically (*ro___–rose*) related under trial-and-error and errorless learning (Cyr & Anderson, 2015). On a subsequent cued recall test, participants were required to write down the correct target associated with each cue, along with the wrong guess that they had generated during the study phase. Trial-and-error learning supported memory, but only in the conceptual condition. We furthermore found that participants were more likely to recall the target if they also remembered their corresponding error – but only in the conceptual condition – suggesting that these mistakes were useful in guiding retrieval (see also Knight et al., 2012). However, these paradigms confound the detrimental effects of producing unrelated guesses with the difficulty of integrating semantically disparate or altogether unrelated cues and targets. The results of Experiment 1 disentangle these confounds, suggesting that when answers make sense in context of the cue (*band–guitar*), errors that are both related (*drum*) and unrelated (*rubber*) to the target enhance retrieval. Moreover, our data suggest that this enhancement is greater for related relative to unrelated errors. The results of Experiment 1 replicate these findings and go further to show that within conceptual error generation, the greater the overlap between the error and the target, the better for memory. This is a novel finding because cues and targets were always meaningfully associated in our conditions, allowing us to isolate the effects of varying error-target similarity on memory.

The results of Experiment 2 mapped on to those of Experiment 1: Greater semantic similarity between guesses and targets did predict retrieval success, extending the supportive effects of semantic proximity to vocabulary learning. However, the error generation main effect did not reach significance. One possibility is that unlike Experiment 1, integrating the target with the cue was more challenging given that the latter were unfamiliar Spanish words. The fact that participants showed a benefit of error generation for the match pairs (*lectura–reading*) but not the mismatch pairs (*bomba–pump*) supports this proposition. Another explanation may be that the errors produced by participants in this paradigm were very constrained: Participants do not have a large pool of candidates to choose from when presented with the *lectura*, and indeed they may have felt that *lecture* was the only possible response. In line with this idea, Grimaldi and Karpicke (2012) found that restricting what participants could generate as an error harmed as opposed to benefited their memory.

Our findings in both studies are best explained by an elaborative retrieval account of error generation benefits. Theories of prediction error (Rescorla & Wagner, 1972; Rumelhart & McClelland, 1986) would have forecast opposite results, i.e., greater memory for targets that are discrepant from our initial guesses. From this perspective, greater learning is required when outcomes differ widely from expectation, leading to greater attentional deployment to feedback. Indeed, individuals are more likely to correct their own general knowledge mistakes if they were held with high relative to low confidence (Butterfield & Metcalfe, 2001): For example, people are more likely to remember that Canberra is the capital of Australia if they initially felt certain relative to uncertain that it was Sydney. This hypercorrection effect of high-confidence errors may be due to greater processing of feedback that violates rather than confirms expectations (Fazio & Marsh, 2009). Our paradigms may not have triggered the affective component required to elicit expectation violation: In both studies, being wrong was unlikely to be surprising (the fact that participants in Experiment 2 correctly translated one word on average supports this idea). Nonetheless, our results are not necessarily at odds with hypercorrection: Using latent semantic analysis, Finn and Metcalfe (2010) found that high-confidence errors were more similar to correct answers than were low-confidence errors. In this sense their results map onto ours, such that answers that were better remembered were more likely to overlap semantically with participants' wrong guesses. Disentangling the memorial effects of learner expectations from those of error-target similarity would be an important goal for future research.

Conclusion

In summary, this research adds to the growing literature on error generation effects by elucidating the factors that amplify its effectiveness. This work has clear implications for educational practice given that questions can vary enormously in terms of how they orient retrieval. For example, it may be better to ask questions that guide learners to guess in the right ballpark (e.g., Question: *What kind of living thing is an earwig?*; Answers: *reptile, bird, insect*) as opposed to left field (e.g., Question: *What is an earwig?*; Answers: *earring, hairpiece, insect*). Successful learning, these results suggest, occurs better when guesses are near misses, rather than out in left field.

Notes

1. There is evidence that the HADS anxiety scale overestimates the extent of clinical anxiety in student populations like the one studied here (Andrews, Hejdenberg, & Wilding, 2006).
2. Twelve participants had perfect recall for either Match or Mismatch guesses; as such, we could not conduct the same set of analyses restricted to unsuccessful recall trials for all individuals. Running the analyses with the remaining participants revealed no significant difference in target memory as a function of whether a Match ($M = 0.54$; $SD = 0.39$) or Mismatch ($M = 0.53$; $SD = 0.40$) guess was unsuccessfully recalled, $F < 1$, $p = 0.91$, $\eta_p^2 = 0.001$.

Acknowledgments

The authors would like to thank Natalia Ladyka-Wojcik for her help with testing and data coding.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by a grant from the Canadian Institutes of Health Research (MOP 123484) awarded to both authors and start-up funds given to Andrée-Ann Cyr by York University, Glendon Campus.

References

Anderson, J. R., & Bower, G. H. (1974). A propositional theory of recognition memory. *Memory & Cognition*, *2*(3), 406–412.

Andrews, B., Hejdenberg, J., & Wilding, J. (2006). Student anxiety and depression: Comparison of questionnaire and interview assessments. *Journal of Affective Disorders*, *95*(1-3), 29–34. doi:10.1016/j.jad.2006.05.003

Berger, S. A., Hall, L. K., & Bahrick, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. *Journal of Experimental Psychology: Applied*, *5*, 438–447. doi:10.1037/1076-898X.5.4.438

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491–1494. doi:10.1037/0278-7393.27.6.1491

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. doi:10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547–1552. doi:10.1037/a0024140

Cyr, A.-A., & Anderson, N. D. (2012). Trial-and-error learning improves source memory among young and older adults. *Psychology and Aging*, *27*(2), 429–439. doi:10.1037/a0025115

Cyr, A.-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. doi:10.1037/xlm0000073

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, *16*(1), 88–92.

Finn, Bridgid, & Metcalfe, Janet. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition*, *38*(7), 951–961. http://dx.doi.org/10.3758/MC.38.7.951

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*, 505–513. doi:10.3758/s13421-011-0174-0

Guild, E. B., & Anderson, N. D. (2012). Self-generation amplifies the errorless learning effect in healthy older adults when transfer appropriate processing conditions are met. *Aging, Neuropsychology, and Cognition*, *19*(5), 592–607. doi:10.1080/13825585.2011.639869

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, *40*, 514–527. doi:10.3758/s13421-011-0167-z

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, *103*(1), 48–59. doi:10.1037/a0021977

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787–1794.

Karpicke, J. D., & Roediger, H. L. I. I. I. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. doi:10.1126/science.1152408

Knight, J. B., Ball, H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*(4), 731–746.

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. doi:10.1037/a0015729

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294. doi:10.1037/a0037850

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284. doi:10.1080/01638539809545028

Luo, L., Hendriks, T., & Craik, F. I. M. (2007). Age differences in recollection: Three patterns of enhanced encoding. *Psychology and Aging*, *22*, 269–280. doi:10.1037/0882- 7974.22.2.269

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489. doi:10.1146/annurev-psych-010416-044022

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting : Choices and consequences. *Psychonomic Bulletin & Review*, *14*(2), 187–193.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is Temporal Spacing of Tests Helpful Even When It Inflates Error Rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1051–1057. doi:10.1037/0278-7393.29.6.1051

Potts, R., & Shanks, D. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*(2), 644–667. doi:10.1037/a0033194

Pressley, M., Tanenbaum, R., McDaniel, M. A., & Wood, E. (1990). What happens when university students try to answer prequestions that accompany textbook material? *Contemporary Educational Psychology*, *15*(1), 27–35. doi:10.1016/0361-476X(90)90003-J

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335. doi:10.1126/science.1191465

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavolvian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York: Appleton-Century-Crofts.

Richland, L. E., Kornell, N. & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. doi:10.1037/a0016496

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Shipley, W. C. (1940). A Self-Administering Scale for Measuring Intellectual Impairment and Deterioration. *The Journal of Psychology*, *9*(2), 371–377. doi:10.1080/00223980.1940.9917704

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of homograph meaning frequency. *Memory & Cognition*, *22*, 111–126.

Vaughn, K. E., Hausman, H., & Kornell, N. (2017). Retrieval attempts enhance learning regardless of time spent trying to retrieve. *Memory*, *25*(3), 298–316. doi:10.1080/09658211.2016.1170152

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, *19*, 899–905. doi:10.3758/s13423-012-0276-0

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition*, *42*(8), 1373–1383. doi:10.3758/s13421-014-0454-6

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, *67*(6), 361–370. doi:10.1111/j.1600-0447.1983.tb09716.x