Forensic Science International: Synergy 4 (2022) 100207

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy

The benefits of errors during training

Heidi Eldridge, PhD [a,*], Jon Stimac [b], John Vanderkolk [c]

[a] RTI International, 3040 E. Cornwallis Rd., Research Triangle Park, NC, 27709, USA
[b] Oregon State Police Forensic Services Division, 20355 Poe Sholes Dr. Ste 200, Bend, OR, 97703, USA
[c] Indiana State Police Laboratory, 5811 Ellison Rd., Fort Wayne, IN, 46804, USA

ARTICLE INFO

Keywords:
Human factors
Errors
Learning
Forensic science

ABSTRACT

Errors are generally not thought of as a positive thing – not in society at large, and especially not in forensic science. However, there is a large body of literature in the field of cognitive science (particularly from psychology and education research) that highlights the benefits that can be gained from using errors made in training to improve learning. Although none of these studies was done directly in the forensic science disciplines, there are nonetheless lessons to be learned about how errors may most effectively be used to maximize their benefits to learning. This article presents an overview of the literature on learning from errors and suggests principles that may be of benefit to forensic science today, as well as suggesting areas where specific research may be of benefit to forensic science in the future.

1. Introduction

No one likes making errors in their work. However, within the forensic sciences, an unusually strong emphasis is put upon the consequences of an error. This emphasis causes examiners to strive to avoid errors of any kind, and at any time throughout their career. Examiners wish to be able to answer the question, "have you ever made an error" with an emphatic, "no," even though the odds are high that they have made *some* kind of an error, whether a conclusion error during training, a clerical error in casework, or some other error.

Although errors may have bad consequences, there is a robust literature in cognitive psychology showing that errors made—and corrected—during training may also have positive consequences in that they may benefit the learning process and result in fewer errors during what is termed "the test that counts" [1]—which, in the case of forensic examiners, includes conclusions, reporting, and testimony in casework. Although this article focuses on the potential benefits of making errors, we are not suggesting that anyone should strive to commit errors, particularly in casework, but rather that trainers, supervisors, and quality managers should *use* errors made during training optimally, to help practitioners learn from them, and to improve the training phase to minimize errors in casework.

This article provides an overview of the psychological literature on learning from errors; it also presents commentary on ways this research may be applicable to the domain of forensic science followed by recommendations for practice and further research. Although there is no expectation that the reader will become an expert in the science of learning, the article does have two practical goals:

1) Provide forensic science managers with the understanding and literature needed to advocate with upper management for the time and training resources necessary to (a) encourage learning from errors during initial training and (b) implement a continuing course of challenging training exercises after competence has been established; and
2) Provide forensic science managers and practitioners with the understanding and literature needed to be able to explain in the courtroom why errors in training exercises should not be held against an examiner

1.1. *Challenging exercises that promote errors can be beneficial throughout an examiner's career*

Although the application of academic research on learning from errors to forensic science is in its infancy, previous research has noted a lack of skill development over an examiner's career. For instance, fingerprint comparison research by Langenburg [2] (pp 199–200) found that the false-positive rate across trainees was very high (9.2%), then dropped to nearly zero across examiners in the first two years of

[*] Corresponding author.
E-mail address: heldridge@rti.org (H. Eldridge).

https://doi.org/10.1016/j.fsisyn.2021.100207
Received 9 September 2020; Received in revised form 14 October 2021; Accepted 15 October 2021
Available online 20 December 2021

independent casework, only to rise over years of experience to a relatively high level (2.9%), where it plateaued.[1] In a black-box fingerprint comparison study (in which comparison trials were given and conclusions collected, without any data on how those conclusions were reached) conducted by the FBI and Noblis [3,4], no correlation was found between years of experience and false negative, true positive, or true negative rates (no meaningful statistical analysis could be done utilizing false positive rates, since so few false positive conclusions occurred). Perhaps continued exercises throughout the career that push the boundaries of an examiner's abilities might result in at least a gradual improvement in performance over time, rather than a stagnation or deterioration of accuracy.

Such suggestions are often met with resistance by laboratory management who fear that continued challenging exercises will expose their examiners to the risk of errors that will be used against them in court and will utilize unwarranted time resources for continued professional development. The National Commission on Forensic Science (NCFS) addressed the first issue in an unanimously approved Views Document [5] that emphasized both the need for performance testing that pushed the boundaries of an examiner's ability and for the courts to exercise caution in the ways they allowed the results of such testimony to be used.

As the NCFS document argues, in order to understand the conditions under which the forensic science laboratory system is likely to fail, performance testing must be challenging enough to induce errors. We learn nothing about the limits of a system if those limits are never exposed because the testing is too easy. NCFS goes so far as to state that "a low rate of failure indicates an inadequate study, and frequent failure is the hallmark of a successful, informative study" (p. 9). The errors thus induced, it is argued, will also serve a second purpose of identifying the areas in which examiners could benefit from additional training that would improve their abilities.

The document takes care to acknowledge the discomfort laboratories and examiners will feel about taking a challenging test in which they are anticipated to make errors and further states (p. 9–10),

> "the Commission urges state and federal judges to consider carefully circumstances under which information about errors in research studies should be admissible in the courtroom. Although the results of such research will be valuable and enlightening on a number of important issues, it would be misleading to equate the rate of error on research samples designed to be highly challenging with the rate of error for cases in general or with the probability of error in a specific case, particularly if the case involved relatively easy or straightforward analysis. Consequently, if the results of performance testing are admitted as evidence in the courtroom, it should only be under narrow circumstances, and with careful explanation of the limitations of such data for establishing the probability of error in a given case."

Although these potential failures need to occur in a protected training environment, there should be no negative ramifications to engaging in the exercise in the first place. This article endeavors to highlight the psychological research basis for the benefits of continued, challenging, errorful learning followed by appropriate and timely feedback as a mechanism for skill improvement. It should be noted that this article is focused on the potential benefits errors made in a training environment and promptly corrected can provide to the cognitive process of learning and does not address error rates or errors in forensic science generally. Additionally, since in casework the ground truth is not known, errors made in casework are a completely separate topic because they cannot by their nature be promptly and reliably corrected. Finally, the difficulty of dealing with the aftermath of a casework error is one more reason it is critical that forensic science as a field take care to distinguish between errors in training and errors in casework and strive to leverage the former to minimize the latter.

1.2. The United States educational philosophy emphasizes error avoidance

Given the high stakes attached to outcomes in the criminal justice system, it should be no surprise that errors are not well-tolerated within forensic casework; if an error is made, an innocent party may go to prison, or a guilty party may be left free to commit more crimes. Due to this strong aversion to injustice, remarkable levels of pressure are often put on examiners never to make any mistakes—a pressure that can result in examiners becoming so conservative that they are afraid to make any judgment at all and retreat to an inconclusive or ambiguous position as often as they can.

Somewhat more surprising, however, is the widespread phobia of errors during the training period, which in many forensic disciplines may be extensive. Anecdotally, many laboratory managers and even other forensic scientists seem to feel that if a trainee makes errors during the training process, this is a red flag—an indication that the trainee may just not be cut out for the demands of the job.

This attitude may have its roots in the prevailing philosophy of learning in the United States. Psychologists dating back to Skinner in 1953 [6] and Terrace in 1963 [7] have promoted the idea that making an error during learning will interfere with correct learning later. The foundation of this philosophy is that once an item is practiced incorrectly and learned, it is stored in the memory where it is difficult to overwrite with new information. Under this school of thought, errors are dangerous because if they are allowed to take hold, the student may never be successful in replacing them with correct information.

Interestingly, this outdated attitude appears to be particular to American education. In a study comparing American schoolchildren to those in Japan and China over multiple years, Stevenson and Stigler [8] found markedly different approaches to errors in learning between the American and Asian schools, and surprisingly, also found different learning outcomes for the students. As they state (p. 192), "For Americans, errors tend to be interpreted as an indication of failure in learning the lesson. For Chinese and Japanese, they are an index of what still needs to be learned."

Stevenson and Stigler go on to describe a typical Japanese classroom. When a new concept is introduced, children are first encouraged to propose their own answers, then prompted to talk about why they thought those answers were correct. Other students are prompted to declare which of their peers' answers they thought was correct, and why. Only after the students have struggled for some time to reach the correct solution through creative (and often incorrect) problem-solving does the teacher point out what makes the incorrect answers wrong, and guide the children to discovering the correct answer, and the reason for it. The goal of the Japanese approach is to provoke thoughtful discussion amongst the students that will deepen their understanding of the problem.

In an American classroom, in contrast, emphasis is placed on getting the desired answer quickly and moving on. In fact, any response that does not conform to the expected answer is frequently ignored, so as not to disrupt the learning of the class, while the expected answer is sought. In an example provided by Stevenson and Stigler, an American teacher, trying to teach the concept of "borrowing" during subtraction posed the question of subtracting 19 from 34. She asked the students whether 9 could be subtracted from 4. The answer she was seeking was that it could not; you need to borrow from the tens column. When one little girl answered "Yes, it is minus 5", she was summarily ignored by the teacher,

[1] It should be noted that this study was not constructed to be an error-rate study and the reported values were considered by Langenburg to be artificially high due to the difficulty of the test pairs chosen for comparison. The purpose of the research was to measure the effect of a tool Langenburg was testing on error rate; in order to observe and measure this effect, Langenburg required a design that would promote errors.

who asked the question of another child and received the response she wanted.

The research results are striking. When comparing the ability of Japanese, Chinese, and American students on carefully-structured mathematics tests in kindergarten, first grade, and fifth grade, Stevenson and Stigler found that in kindergarten, Japanese students were already outperforming Chinese and American students, yet by first grade, the Chinese students had quickly caught up to their Japanese counterparts, and by fifth grade, the American students had been left far behind.[2]

The Japanese approach of allowing students to err, discussing the causes of the error, and then continuing the lesson with appropriate techniques has been implemented with some success in training courses in forensic domains. For example, in educational workshops on fracture examination given by Vanderkolk, students are encouraged the first day of class to simultaneously tear multiple pieces of paper along the same lines (creating highly similar, yet different, tear patterns). When instructed to re-assemble the paper fragments, errors are sometimes made in the process, which are then discussed and deconstructed prior to continuing with learning the appropriate technique for conducting, and documenting, a fracture examination. This exercise has resulted in increased awareness of both the need for increased documentation and the variety of features that must be considered in a fracture examination.

Although this article will summarize psychological research that supports the benefits of learning from errors, it is interesting to note at this juncture that even learners themselves are often unaware of the benefits of their errors, which may contribute to the cultural perception that errors during training are to be avoided, or may simply be a product of that ingrained belief. For example, Huelser and Metcalfe [9] had subjects study both related and unrelated cue-target word pairs that were presented in three different conditions. In the first two learning conditions, the cue and target words were presented together for either 5 or 10 s (e.g., shoe-ankle). The third was an error-generation condition in which the cue word was presented alone for 5 s (e.g., banana- ?) and the subject was encouraged to guess the target word, followed by both the cue word and the target word together for an additional 5 s. After a short distractor task, subjects were tested on their recall of each studied pair by being shown the first word of a pair and being asked to enter the second word of the pair into a text box. After completing the testing, subjects were asked which of the three conditions best helped them to learn the word pairs. Interestingly, the word pairs in the error-generation condition led to the highest proportion of correct answers on the test, however, the subjects reported that the 10-s learning condition had helped them to learn the information best. In other words, they gained a learning benefit from the error-generating condition that they did not recognize, even only minutes after having experienced it.

As we shall see, recent empirical evidence has demonstrated [10] that early assumptions that making errors impedes learning are incorrect. In fact, the generation of errors during learning actually promotes better learning and many theories are posited in the literature to explain this observation.

2. Corrective feedback creates enhances learning after errors

As described in the psychological literature on learning from errors, it is clear that not *every* error is beneficial. There are specific conditions under which errors can aid in better learning. The most obvious of these conditions is that the error must be corrected—an error that is never corrected will never be unlearned. However, there are a number of possible variables as to *how* the error will be corrected, *when* it will be corrected, and *what kinds* of errors can be successfully corrected.

2.1. Learners must generate the answers themselves

A body of literature [11] states that the very act of testing aids in learning. It is thought that the practice of retrieving stored information to answer a test question improves the learning of that information. Researchers have explored whether this phenomenon could be exploited to improve learning by forcing errors in a pre-testing situation, prior to the actual test.

To examine this hypothesis, Grimaldi [12] divided experimental subjects into two groups: a pre-test (learning through errors) condition and a no-pretest (studying) condition, and asked them to learn weakly related word pairs (e.g., tide-beach). In the pre-test group, subjects were shown the cue word and asked to guess the target word, which they usually guessed wrong. Then they were shown the correct pair to study. The no-pretest group simply studied the word pairs with no guessing and no feedback. When subjects were given a final cued recall test (i.e., shown the first word and asked to remember the second), the pre-test group that had made errors and received feedback during the learning phase successfully recalled more pairs than the no-pretest group.

In contrast, if, during the learning phase, subjects were constrained in their guesses by a hint providing the first 2 letters of a wrong word (e.g., tide-wa_ _), thus forcing them to make a specific error, performance during the final test was not enhanced, and in fact these subjects did worse than those who simply studied the word pairs without pre-testing.

Grimaldi proposes that the reason for this effect is "search set theory", which states that when a guess is made, the brain searches for the correct answer by producing a set of loosely-related possibilities from which it selects a response. If that response is wrong, and the correct response is provided, the association between the wrong response and the correct response (which both belong to the same related search set) helps to reinforce the learning of the correct response.

In Grimaldi's second situation, when a 2-letter hint prompted subjects to select a specific (but wrong) response, the correct answer was not part of the search set. Thus when the correct answer was presented, it had no association with the incorrect response and rather than reinforcing the correct response, the wrong guess interfered with learning it.

Similarly, under this theory, learning of unrelated words would not be helped by a pre-test that induced errors because the guessed (incorrect) word would not share a set with the correct target word, thus no association would be formed to help with learning of the correct word.

This research echoes the findings of Huelser and Metcalfe [9] who also report that errors must be related to the correct response to reap a learning benefit. Unrelated errors do not improve learning outcomes. Due to this effect, Huelser and Metcalfe suggest forcing an answer on every item in a test, positing that this procedure will create errors and allow their effect to be studied.

A similar approach is being tested by Busey and Vanderkolk in a series of recent workshops in which challenging fingerprint comparison tasks are presented to workshop attendees, who are fingerprint examiners. Rather than using the three traditional conclusions of exclusion, inconclusive, and identification [13], attendees had to select from an expanded range of six possible conclusions: exclusion, almost exclusion, tending towards exclusion, tending towards identification, almost identification, or identification. By using this scale of six conclusions without an option for a neutral inconclusive, the attendee was forced to take a stand in one direction or the other and risk being confronted with a ground truth answer from the other side of the divide. They will also have to justify their chosen conclusion, which forces them to examine their reasoning and their threshold levels and learn from their thought process, much in the way that was exemplified by the Japanese classroom described earlier in the article.

An additional benefit to using this sort of an exercise in a traditional

[2] Lest anyone conclude from these data that Asian students are simply more intelligent than Americans, the researchers also conducted culturally- and demographically-balanced IQ tests for the three groups and discovered that, while children in each culture exhibited different specific cognitive strengths and weaknesses, there was no significant difference among total IQ test scores for any of the three groups by the fifth grade.

training situation may be the insight it gives the trainer into the rationale behind the trainee's decision, which can help to tailor and direct future training and exercises.

2.2. Learners can be simultaneously aided and challenged by scaffolded feedback

Another means of forcing errors during training to benefit learning is the technique of "scaffolded feedback", a term that was coined by Finn and Metcalfe [14]. In scaffolded feedback, incremental hints to the target response of a general knowledge question (in the case of the study, one letter at a time of the target response) are presented until the correct answer is self-generated. This tactic has two advantages: first, it requires self-generation of a response, which we have just seen to be beneficial, and second, it ensures that learning occurs just on the border of what the subject knows, which continually challenges them.

In their study, Finn and Metcalfe tested four learning conditions: scaffolded feedback, standard feedback, minimal feedback, and answer-until-correct. Standard feedback simply provided the correct answer after the initial test, while minimal feedback indicated that the initial answer was wrong, but did not provide the correct answer. Final recall testing was delayed either 30 min or 1 day after the learning phase, and in all cases, scaffolded feedback produced the highest proportion of correctly learned responses.

There are several unknowns left from this study that the authors of the study recommended for future research. For instance, how would these results transfer to other domains of learning, such as problem-solving; what would be the effect of a longer delay before testing; how would different levels of motivation affect the learning; and would a more sophisticated system of semantic cues produce different results than simply revealing letters in the answer?

A technique similar to scaffolded learning has been used by Vanderkolk in the training of tool mark comparisons. In the exercise, a single, continuous striated tool mark was made by sliding a screwdriver blade approximately 10–15 inches across heavy duty aluminum foil. The application of the blade began at an oblique angle, then gradually changed as the screwdriver was dragged along the surface until it ended nearly perpendicular to the foil.

This continuous tool mark was then cut into 9 segments, which were lettered non-consecutively and presented to students for comparison. Students may begin by only associating a few pairs of the marks as coming from the same tool, but after repeated prompting from the instructor, managed to associate more and more pairs until they finally self-generate the realization that all nine marks were made by the same tool, whose appearance varied as the angle of application changed. Thus, while the consecutive sections of the mark shared very similar appearances, those at the extreme edges did not, and the gradual similarities of each segment in between were needed to bridge the gap between them (Fig. 1 - Fig. 3).

2.3. Optimal timing for corrective feedback depends on the task

The question of the optimal time delay for receiving feedback and between learning and testing is a salient one and has been explored in depth in the literature. Butler [15] found that a delay in feedback (as compared to immediate feedback or no feedback at all) improved learning outcomes in multiple choice testing on a memory task; however, the length of the delay was only until the end of the practice test and final testing was completed one week later.

Butler's theory of this phenomenon was that during immediate feedback the correct answer and the wrong answer were competing to be committed to memory. In delayed feedback, the wrong answer had time to dissipate, allowing the correct answer to be recorded more easily. Once again, unknown in this study was whether there is a delay that is *too* long to be effective.

Metcalfe et al. [16] noted that in nearly all previous delayed feedback studies, the time between the learning phase and the final testing phase was held constant and the delayed feedback condition always produced better learning outcomes. In other words, the feedback was provided right before the test. This suggests that if feedback was provided immediately after learning, there was a long period between the time the feedback was received and the time the final test was given (lag-to-test), whereas if the feedback was delayed, there was a shorter lag-to-test (Fig. 4). They were concerned that the beneficial effect of delayed feedback was not a cognitive advantage of learning at all, but simply that the subjects did not have to remember the correct information as long.

To test this, they designed a layered study in which several learning

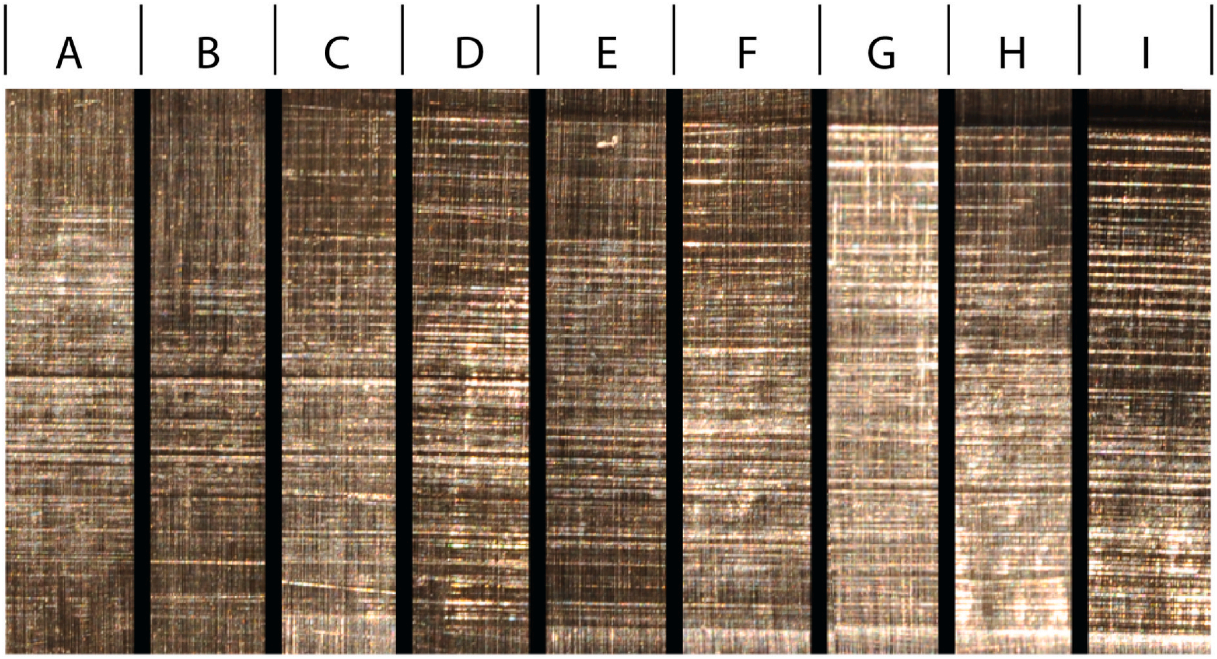

Fig. 1. Nine segments cut from a continuous toolmark, here designated as A-I for convenience. In a real training exercise, the labels would be non-continuous.

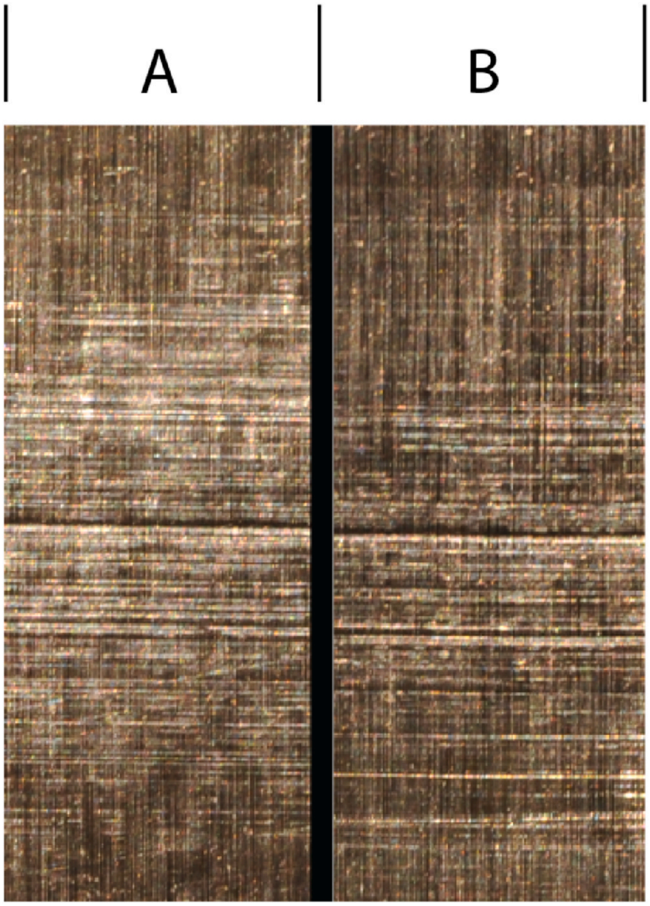

Fig. 2. Two consecutive segments of the tool mark. Note the similarity of the striations.

and testing sessions were scheduled on a rolling basis such that at each session, subjects were learning some new information, being tested immediately on some information, and being tested on a lag for some information. With this design in place, they were able to control for the effects of lag-to-test time. Their findings revealed that for sixth-grade students, the results from delayed feedback were always superior, even when lag-to-test was controlled, yet for university students, the benefit disappeared when the lag-to-test period was controlled (although both delayed feedback and immediate feedback were still superior to no feedback). These results do not offer a clear recommendation of when feedback is most effective and suggest that further research in this area may be needed, but do strongly support that it is critical that feedback *is* provided.

The timing of feedback may depend in part on the learning mechanism used by students. In more perceptual-based tasks that are similar to pattern comparison tasks found in forensic science, feedback is better if it is *immediate*. Maddox and colleagues [17,18] found that delaying feedback by as little as 5s resulted in a decline in accuracy, and attributed this decline to a failure to update the weightings of pathways in the visual system while they still had residual activation from viewing a stimulus. Delayed feedback (on the order of seconds) had little effect when participants were using a conscious, rule-based strategy (as opposed to a more intuitive, holistic strategy). This suggests that conscious maintenance of the strategy in memory can overcome a delay, while the fading visual image doesn't allow for learning with delayed feedback when an implicit strategy is used.

In a rare break from word-pair or general-knowledge testing studies often used in this domain, Kang [19] focused on studying the learning of unfamiliar scientific phenomena that were either explained upfront, or explained only after the subjects were required to think up a plausible explanation (which was often incorrect). In this study, the final test was administered either 10 min or 1 week after the learning phase, and under all experimental conditions, there was no negative effect found to having first guessed an incorrect explanation. This experiment seems to be closest to a real-world situation out of this series of studies because it required deep thinking about a scientific topic in order to generate the errors; still, it comes down to learning factual knowledge more than problem-solving or learning of a skill requiring decision-making. Future research applying these concepts to the forensic science domain directly would be of benefit.

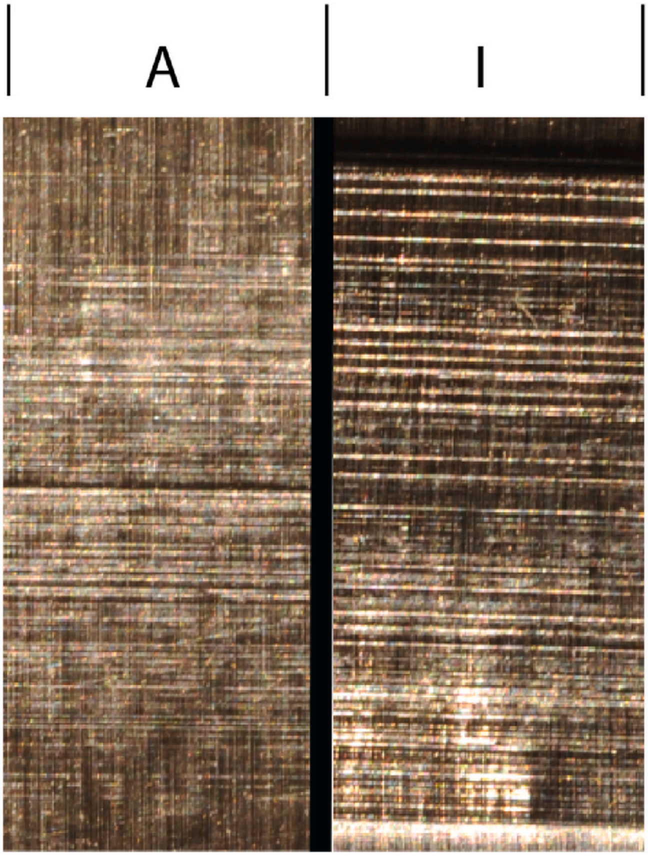

Fig. 3. Segments from the two extreme ends of the tool mark. Note how much the striations have changed in appearance.

2.4. Correction is more effective on errors made with high confidence

Contrary to the predictions of traditional thought about learning, numerous studies have observed that correction is more effective on items about which the subject had high-confidence than those that were answered with low-confidence. This surprising phenomenon has been termed the hypercorrection effect and a great deal of study has been done into understanding why, and under what conditions, it occurs.

The best supported theory is that when a person is confronted with an error that they were highly confident they had right, they are surprised that they made the error and this surprise causes them to divert more attention to the correct answer than they would have if they received expected news about the rightness or wrongness of their answer. This increased attention assists in increased memory of the correct answer.

This theory has been supported by three particularly persuasive studies. In the first, Butterfield and Metcalfe [20] presented feedback from a general knowledge test while subjects were taking a tone recognition test (i.e. subjects listened to a tone being generated and

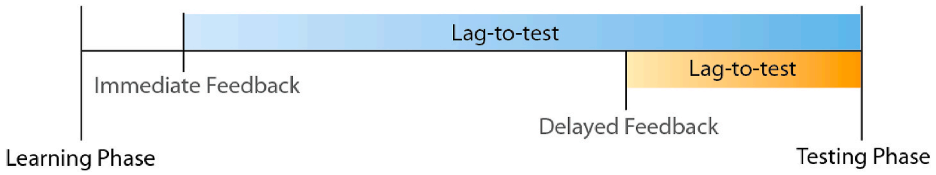

Fig. 4. The lag-to-test period. In the bar above the midline, immediate feedback is given and there is a long period between receiving feedback and final testing (the lag-to-test period). In the bar below the midline, delayed feedback is given, and the lag-to-test is shorter, reducing the amount of time a subject would need to remember the feedback.

indicated when they could hear it, much like the common hearing test). They found that subjects' performance suffered on the tone detection test while being notified of their high-confidence errors, in comparison to their low-confidence errors. This suggested that the subjects were distracted from the tone recognition task by the surprising news of their error.

In the second study, Fazio and Marsh [21] proposed that not only high-confidence errors but also low-confidence correct answers would be surprising to subjects and would thus demand higher attentional resources than their converses. In order to test this (because retesting would not detect a difference in items that were originally answered correctly), they measured attention to the feedback by measuring surrounding context when the feedback was provided. In their first experiment, the 'surprising' feedback was presented on-screen in a different font size and color than the unsurprising feedback and subjects were later asked to recall what color/size the feedback had been. In the second experiment, feedback was presented audibly by either a male or female voice. In both cases, subjects were better able to recall the manner of feedback for the surprising feedback than the unsurprising, indicating that they had attended to it more carefully.

In the final study on the surprise theory of hypercorrection summarized in this article, Metcalfe et al. [22] used fMRI imaging to show the regions of the brain that were activated while hearing surprising versus unsurprising feedback. Like Fazio and Marsh, they tested both reactions to high-confidence errors and low-confidence correct answers, both of which they termed together as "metacognitive mismatch". Their results supported that during the notification of an error, people (1) entertain both versions, and (2) try to suppress the incorrect one. It also supported the idea of surprise registering at the news of a metacognitive mismatch.

While the hypercorrection effect has been replicated in numerous studies and is under little doubt, questions still remained about its mechanics. For instance, how long does the effect last? A study by Butler, Fazio, and Marsh [23] found that high-confidence errors were indeed more likely to be corrected in the short term, but that after a one-week delay, they were also more likely to be repeated if the correct answer was forgotten. The researchers went a step further and investigated the effects of immediate practice or re-testing upon receiving feedback to reinforce the new information, and found that indeed, immediate practice solidified the learning and allowed the correct answers to be retained long term.

Sitzman et al. [24], however, hypothesized that more was going on. They noted that prior studies had not taken prior knowledge into account. They proposed that confidence is a proxy for prior knowledge, because you tend to be more confident in domains that you know more about. Testing for prior knowledge effects, they found that prior knowledge was a more reliable predictor of error correction than was confidence.

The hypercorrection effect has also been demonstrated on errors that were caused by erroneous inferences [25], which are often held with very high confidence. This result could be particularly relevant to forensic science because many forensic science conclusions are made by drawing inferences, and academic commentators have warned that contextual or confirmation biases can lead to erroneous inferences (see, e.g., [26,27]). Additionally, by comparing from known-to unknown--source samples, there is a risk of erroneously inferring features that are not truly present in the unknown, particularly in degraded samples (see, e.g., [28]).

In their study, Fazio and Marsh [25] induced erroneous inferences by providing subjects with sentences that heavily implied an action that was, in fact, not stated. For example, subjects might read the sentence: "The karate champion hit the cinder block". Subjects were later tested with the same sentence with the key action word missing: "The karate champion _____ the cinder block." They then recorded whether the blank was filled with the correct word, an erroneous inference (e.g. "broke"), or some other error (e.g. "kicked") and recorded the subjects' confidence in each response. They found that initially, the erroneous inference was more common than either the correct answer or other errors. However, after feedback was received, the hypercorrection effect was observed and the most high-confidence errors were more likely to be corrected.

Some research has examined the role of confidence in forensic science decisions. Such studies have used a forced-choice design in which participants must choose to identify or exclude without a neutal inconclusive option, followed by asking subjects how confident they were in their decisions [29–32]. However, because these studies did not follow up with presentations of feedback, the hypercorrection effect could not be measured. Future studies could be designed to test this research in a forensic science context.

2.5. Feedback should include a review of the reasoning that led to the error

Earlier in this article, we described how Japanese schools put an emphasis on understanding the reasoning behind errors before an explanation is given for why the right answer is correct. These principles are not entirely foreign to the United States. In fact, as described by Morrison and Meliza [33], the US armed forces utilize a training technique known as hotwashing, or After-Action Review, in which novice trainees participate in a battle scenario against an experienced team and are usually summarily thrashed. They then return to a debrief in which rank is disregarded and everyone is encouraged to have a voice in the proceedings. At this debrief, they review all the errors that led to the defeat and examine why each occurred and what should have been done instead. This technique has proven to be effective; the same mistake is rarely made in subsequent training exercises. When the stakes are high, such as in life-or-death battle scenarios, or when making forensic science decisions that impact the appropriate application of justice, it is easy to see why it is preferable to learn from mistakes in a training scenario, rather than making them in the field.

A similar method has been implemented at the Las Vegas Metropolitan Police Department Latent Print Unit. After an undesirable chain of erroneous exclusion decisions was not improved by occasional unit trainings on exclusions, they decided to try a new response method. After an erroneous exclusion was uncovered in casework, the original analyst would host a unit meeting in which they would show the error to their peers, describe how the error was made, and it would be discussed by the group in a non-judgmental environment so that all could learn from deconstructing the error.

Similarly, during crime scene training, collection exercises are commonly followed by a debriefing session during which trainees are encouraged to explain their thought process behind collecting certain items and leaving others behind. As, realistically, not every item present in a crime scene can be collected, it is vital that crime scene examiners be trained to walk the line between being “garbage collectors” and leaving behind potentially probative evidence. These exercises help them to fine-tune their logic and decision-making by critically evaluating the choices they made in a mock crime scene.

Finally, one of the best-known and most-embarrassing errors in forensic science occurred in the latent print discipline. In 2004, the Federal Bureau of Investigation erroneously identified a Portland, Oregon lawyer by the name of Brandon Mayfield as the source of a latent print on a bag that had held detonators used in the bombing of a commuter train in Madrid. This case is so well-known partly because it can be used as a touchstone for so many places where the process and the quality system broke down. However, here, it is of interest to us because of this notion of After-Action Review.

Two years after the Mayfield error came to light, the Office of the Inspector General (OIG) [28] released an in-depth report on the root causes of the error. This document, which was the result of many interviews and reconstructions, was 330 pages long and represented the most thorough deconstruction of a forensic science error to date. The Mayfield error is rich in lessons to be learned, from documentation to cognitive factors to application of the method to quality assurance. Laboratories the world over have read the OIG document and changed their policies, procedures, training programs, and quality assurance programs in response, to the betterment of the field.

3. Emotional consequences of errors

Although laboratory and forensic culture often put a high level of pressure on examiners not to make errors, and examiners often put the same pressure on themselves, the repercussions of actually making an error are often even worse. Examiners caught in an error may find themselves subject to re-training, ostracizing, public shaming, loss of credibility in the courtroom, or termination. Despite these potentially devastating outcomes, there is very little research in the area of trauma to forensic scientists from errors, and very few mental health resources available to them to get through these events [34]. The mental toll taken on doctors [35] and police officers [36] after errors has been studied, but research is needed to establish the extent of the toll taken on forensic scientists.

In the general psychological literature, some studies have examined the emotional impact of errors on subjects and how they affect decision-making and motivation.

Crowe and Higgins [37] explore how the regulatory focus of individuals can influence how they are affected by errors. There are two types of regulatory focus: Promotion focus and Prevention focus. The characteristics of each are summarized in Table 1.

A person may tend naturally toward one regulatory focus or the other as a function of their personality; however, regulatory focus can also be manipulated by framing of the consequences. By priming a person’s regulatory focus through promotion or prevention feedback, motivation and decision-making may be influenced.

Table 1
The characteristics of the two types of regulatory focus. Regulatory focus type can influence how a person is affected by errors and may occur naturally or be manipulated by a framing of consequences.

Promotion Focus	Prevention Focus
Concerned with advancement, growth, and accomplishment	Concerned with security, safety, and responsibility
Tries to match a desired end state	Tries to avoid a mismatch of the desired end state
Risky bias	Conservative bias
Eagerness	Vigilance
Wants to accomplish “hits” and avoid errors of omission	Wants to attain correct rejections and avoid errors of commission

In their experiments, Crowe and Higgins determined the natural regulatory state for their participants to control for predispositions, then prior to assigning them to several experimental tasks within a particular induced promotion or prevention focus framing condition, asked them to rank several activities using a 7-point Likert scale ranging from “Dislike Very Much” to “Like Very Much”. These activities were things like playing a video game or proofreading a document. For each participant, one activity was chosen that they clearly liked, and one that they clearly did not like. There were four framing conditions in which participants were told that they would be doing either the liked activity or the disliked activity after completing a test (the test comprised five types of questions: Characteristic listing of objects, Counting backwards by set increments, Sorting into subgroups according to a single criterion, Embedded figures within a larger complex figure, and Anagrams) and that which activity they did would be contingent on their performance on the test (in a fifth, non-contingent, condition used as the control, they were told the activity they would perform would be random).

The instructions provided for each of the four focus framing conditions were as follows (reproduced from Crowe and Higgins [37]:

a) Promotion Working – “If you do well on the exercises I’m about to give you, you will get to do the [participant’s liked task] instead of the other task.”
b) Promotion Not Working – “If you don’t do well on the exercises I’m about to give you, you won’t get to do the [participant’s liked task] but will do the other task instead.”
c) Prevention Working – “As long as you don’t do poorly on the exercises I’m about to give you, you won’t have to do the [participant’s disliked task] but will do the other task instead.”
d) Prevention Not Working – “If you do poorly on the exercises I’m about to give you, you will have to do the [participant’s disliked task] instead of the other task.”

Mood was also assessed throughout the study to see how the induced states were affecting the subjects. As hypothesized, Crowe and Higgins found that they could induce a promotion or prevention focus in their subjects and that these in turn influenced strategies or decision patterns in problem-solving. Individuals in promotion-focused conditions were more risk-taking, focused on trying to generate “hits” while individuals in prevention-focused conditions were more cautious, focused on avoiding errors.

Furthermore, they found an effect of difficulty in which when individuals were working on a difficult task, or had just experienced failure, those who were in a prevention focus were more likely (54%–35%) to quit the task and thus avoid an error than those in a promotion focus.

Although this research was done on non-forensic science tasks and the stakes were relatively low, nonetheless, the research has obvious implications for forensic practice. If an examiner’s regulatory state can affect their motivation, decision-making, and propensity to quit a difficult task, it could very well affect the number of missed identifications or exclusions made, and the number of inconclusive decisions reported. Interestingly, one of the tasks performed in the study, called “embedded figures” involved locating a previously seen simple figure inside a larger, complex figure. This task could be a direct analog to the pattern-searching tasks in many pattern evidence disciplines (e.g., searching for a small target group within a large, somewhat distorted, palm exemplar). Research should be done to see whether this effect is reproduced in forensic science examples and how higher-stakes stress affects participants’ susceptibility to regulatory focus states.

Crowe and Higgins suggest that more research is needed, but that employer-employee contingencies (i.e. a reward or punishment system) could be communicated through incentives and feedback to influence motivation and performance in the desired direction. Clearly, more knowledge in this area could have direct implications for policies that could help to motivate examiners who are too cautious and to reduce errors of commission in examiners who are too risk tolerant. This knowledge may also find application in the area of personnel selection and assessment, where hiring managers may wish to consider applicants' natural regulatory focus.

The intentional framing of consequences may help to make more explicit different examiners' risk tolerances and help inform their utility functions. Utility functions in forensic science decision-making have been introduced by Biedermann et al. [38]. The idea behind the utility function is that for every decision that one makes in forensic science, there is a potential benefit if you are correct, and a potential cost if you are wrong. These costs and benefits may be set by agency preference, or by the hopes and fears of the individual examiner and can include things such as being fired or facing a departmental lawsuit in the case of a bad identification, or gaining praise or aiding the criminal justice system in the case of a true identification. Other outcomes, such as false exclusions or true exclusions, may have costs and benefits that are less extreme. Biedermann et al. propose that, in addition to the evidence that is examined, these potential costs and benefits factor into an examiner's decision (e.g., "I see some information in agreement between these two images, but I'm not quite sure it's enough—should I risk being publicly humiliated if I call it an ID and I'm wrong?"). This weighing of costs and benefits may not happen consciously, or it may be quite explicit. The intentional influencing of regulatory focus through agency policies and incentives could help to make that cost and benefit tradeoff more transparent.

In a second study exploring how emotions generated by errors affect learning, Zhao [39] focused on two main variables: how subjects perceived management tolerance of errors, and the subjects' emotional stability. Errors were defined as a difference between the result and the expected result that had a negative consequence for the business or customer (i.e. not just a sub-optimal result). Learning from errors was defined as a purposeful process of reflecting on the error, determining its root cause, and learning how to avoid it in the future. This made it more similar to the way accredited forensic laboratories deal with errors than most of the psychological literature, which deals with memorizing facts or word pairs.

Zhao used a business-model simulation to test the following relevant hypotheses:

H1. Perceived manager intolerance to error increases negative emotions

H2. People with high emotional stability are better able to regulate negative emotions

H3. High negative emotion negatively impacts motivation to learn (because attention is diverted from learning to the negative emotions)

H4. Motivation is positively related to learning from errors

Hypotheses 1, 2, and 4 were supported by the results and Hypothesis 3 was not supported by the results. However, the experimental design had a critical flaw in that the business simulation that was provided had no real emotional stakes. While subjects were told that the management had a high or low tolerance for errors, since they knew it was a simulation and there were no consequences for poor performance, there was no true motivation to perform well and no true emotional backlash if things didn't go well. Thus, it was impossible to accurately measure anyone's negative emotions in relation to the exercise. The author recognized this and acknowledged that the participants never really felt negative emotions.

The results of this study, had it been properly tied to emotional stakes, could have been enlightening both in managing laboratory culture (i.e. management tolerance for errors) and in making employee selection recommendations (i.e. hiring candidates with high emotional stability). Unfortunately, this research needs to be re-done in a more realistic situation with meaningful stakes before policy recommendations could be made.

4. Errors in the context of operational laboratories and the criminal justice system

4.1. Broader consequences of errors

In addition to the emotional consequences of an error faced by the case examiner, there are broader implications to errors that necessarily influence a laboratory's tolerance for errors from its employees. Although clerical errors or disagreements over sufficiency of data to support a conclusion are fairly commonplace and easily dealt with, more significant errors such as erroneous conclusions can have disastrous and lasting consequences.

For the laboratory, a serious error will engender a large expenditure of resources. If the examiner has already completed training and is authorized to perform independent casework, an erroneous conclusion can lead to extensive reviews, retraining, or termination. These tax agency resources through labor-intensive reviews from the legal, administrative, forensic science, and investigative parts of the agency. Most likely, the examiner would be removed from casework during this review, which could result in larger backlogs and additional pressure for other examiners in the unit. There may be a review of other cases that were worked by the examiner in question. The examiner or their laboratory could face costly lawsuits for their part in a wrongful conviction. Plus, the agency for whom the mistaken examiner works would be scrutinized and could be subject to losing their accreditation status, if they were an accredited laboratory to begin with.

The situation is somewhat better, but still intensive, when errors cause a trainee to fail the training program. Training periods may have to be extended, workplans will have to be generated and approved by Human Resources. The Legal department may have to get involved. The problem could be with the trainer and a new trainer would need to be provided. The problem could be with the training program, which would need to be re-designed. The problem could be with the trainee who may be terminated, thus representing a loss of a large investment in their hiring, background check, and time spent training them thus far, which will all have to be repeated for a new trainee, further tying up resources, increasing backlog, and decreasing morale.

The judicial system also faces consequences when erroneous conclusions are reported in casework. An erroneous exclusion or an inappropriate inconclusive could result in the actual criminal not being arrested, charged, tried, convicted, or sentenced. An erroneous inclusion could result in an innocent person being investigated, arrested, charged, tried, convicted, and sentenced. These situations do not serve justice and can be costly for the judicial system in re-trials, wrongful conviction lawsuits, or suits from family members of subsequent victims. Additionally, the examiner often loses credibility with the prosecutor, who may no longer feel they can put them on the stand.

Finally, the discipline in which the examiner worked can often feel a ripple effect from an erroneous conclusion. These cases will be re-tried in the popular media and may set court precedent as well. Both of these actions can hurt the credibility of the entire discipline, who may be answering questions about a high-profile error for years.

With the stakes this high, it is critical that errors in "the test that counts" be minimized by any means at laboratories' disposal. Because neither science nor humans are error-free, it is not realistic to expect that an error-free environment is possible. In addition to human error, method error, system error, and random error all have a role to play. However, the goal is to minimize or mitigate errors wherever possible.

4.2. Generating errors in training

While the psychological literature seems to support the idea that generating errors during training is highly beneficial to learning, many questions will need to be resolved to successfully transfer this idea to forensic science practice, such as: what is the best way to build exercises that incorporate beneficial error experiences, or what is considered an error in the particular discipline?

In most forensic disciplines, there are two obvious and agreed-upon errors: the erroneous identification and the erroneous exclusion. However, once we move in from these two conclusions at the extreme ends of the spectrum, things get a little more nebulous. Different forensic disciplines have different ranges of conclusions and some allow for many shades of gray along the continuum of conclusions. What is an error? In highly interpretive disciplines such as crime scene reconstruction or bloodstain pattern analysis, the definition of an error can be even more difficult to pin down and may often come down to a matter of how well the conclusion was supported, or whether sufficient alternative conclusions were considered or tested.

Even when considering the 'inconclusive' decision, there is ambiguity about when it is an error. In a test where ground truth is known, is 'inconclusive' *always* an error? After all, the ground truth was either same source or different source. However, the correctness of the conclusion will never be known in casework and one must rely on the sufficiency of the data to support the offered conclusion. Therefore, when we are relying on the notion of sufficiency, 'inconclusive' could very well be the more appropriate response (for more in-depth discussion and examples, see [40,41]). For example, imagine a comparison between two images of low quantity and quality of information. Perhaps the ground truth is that these two images were made by the same source. Yet there is so little reliable information in common visible between the two images that it would be irresponsible to claim an identification and impossible to conclude that those few murky, indistinct features could not *also* appear to occur in an unrelated source. In this case, the identification conclusion (though true) would be inappropriate and should be counted as an error, while the inconclusive decision would be correct.

But who gets to decide? When we are discussing sufficiency thresholds, it is very difficult to define an objective threshold for when inconclusive is appropriate and when it is not [40,42]. Thus, how does a trainer determine that the trainee has made an 'error' simply because their own sufficiency threshold does not match that of the trainer? In fact, making a decision near the threshold of an inconclusive rather than a conclusive decision is one of the great challenges within forensic science. Ulery et al. [4] found that most disagreements between fingerprint examiners were a question of sufficiency as opposed to being diametrically opposed opinions of identity versus exclusion. The same argument can be applied to the suitability decision—whether or not the unknown image has enough reliable detail to proceed to comparison—which could be considered an error if the trainee does not assign suitability to images the trainer would have.

In an attempt to provide guidance to fingerprint examiners on when an identification is warranted, when it is not, and where the 'gray area' lies, SWGFAST [13] offered a sufficiency graph (Fig. 5), which gave zones according to the quantity of minutiae observed and the quality of the unknown impression. This graph is a very useful tool, but still has limitations. First, it includes only two dimensions (quantity and quality), without representing the selectivity of the arrangements or other features that an examiner may use to make a decision. Second, it does not indicate what constitutes an 'error'; only where the danger zone lies. Third, it is not the result of a validated study, but represents the consensus view of a group of experienced examiners.

Thus, while it is easy to identify an error when ground truth is known and the conclusion is diametrically opposed to it (e.g., an identification decision made on different source images), the notion of errors becomes trickier when dealing with the range of inconclusives. This is one reason the tactic being tried by Busey and Vanderkolk of removing the inconclusive option during a training exercise to force a choice may have advantages. As one of the goals of training is to reduce the 'gray area' of indecision to the extent possible (keeping in mind that sometimes 'inconclusive' really is the most appropriate response), any training that emphasizes making a choice, documenting and defending the reasons for that choice, and then receiving feedback on the ground truth state of the comparison should help to fine-tune the ability to make these decisions when the stimuli are less clear.

Ideally, training programs should be designed so that they increase in difficulty as the trainee progresses. However, as many forensic science disciplines lack objective measures for the difficulty of a sample or a comparison, this too is often a subjective determination. One way to measure that the training program is in fact becoming increasingly difficult is that the trainee should still be making errors as they progress through it. If the trainee is making no errors, their inconclusive 'gray area' is probably too wide, the exclusion and identification determinations are probably too easy, and little learning is likely taking place. Unfortunately, when an examiner relies too heavily on the inconclusive decision, they are losing sensitivity (i.e., allowing identifications to be passed by that could have been made) [43]. Some agencies have taken the fear of error to such an extreme, they have set very conservative thresholds in their training program that have resulted in a high degree of conformity between examiners, but also a decrease in cases that have been resolved [2]. Thus, it is desirable that a training program in any discipline should generate challenging exercises near the inconclusive threshold that will force examiners both to make errors, and to gain awareness of where this threshold should lie.

At this point, the reader may be concerned that this article is advocating that trainees should be encouraged to just guess without putting any thought behind their conclusions. Nothing could be further from the truth. Outright guessing is not encouraged, but errors made under uncertainty will provide teachable moments that can lead to better decisions under uncertainty in casework. Furthermore, even if the trainee *does* resort to guessing, the psychological research supports that guessing while in a problem-solving state of mind, and being corrected with thoughtful feedback, may actually lead to a deeper understanding.

Finally, a thorough training program in any forensic science discipline should incorporate challenging mock court exercises. Trainees typically do not enjoy these exercises, which can be a source of significant stress. However, they represent a safe place where the trainee can make errors on the stand, be corrected on them, and be less likely to repeat those errors in a real courtroom situation. Once again, the benefits of learning from error in a training environment can be applied to improve real-world outcomes.

4.3. Training exercises for caseworking forensic scientists

Not only trainees can benefit from learning from errors in a training environment. As was mentioned earlier in this article, many examiners' skills will stagnate or deteriorate during their careers and older adults are also capable of reaping the benefits of errorful learning.

Although it was argued earlier in this article that the very act of taking a test can aid in learning, the only testing that most forensic examiners receive after being signed off for independent casework are annual proficiency tests. As many commercially-available proficiency tests are not challenging [44–46] and also can affect examiners in accredited agencies' ability to continue doing casework, these are not an ideal instrument for learning.

However, consideration should be given to designing an ongoing testing program outside of annual proficiency testing that is intended for measurement and growth of examiners' skills. These performance tests, advocated for by the NCFS [5], could incorporate difficult tasks that would challenge even experienced examiners. This would allow managers to identify areas where the examiners would benefit from additional training to refresh or improve their skills. In this way, the training would never really end, and examiners could continue to hone their skills throughout their career, from within a safe training environment that would not be used against them in court.

This approach is being used in the DNA section of the Indiana State Police. In a deliberate attempt to improve the accuracy of experienced examiners near the threshold of sufficiency, they have created an

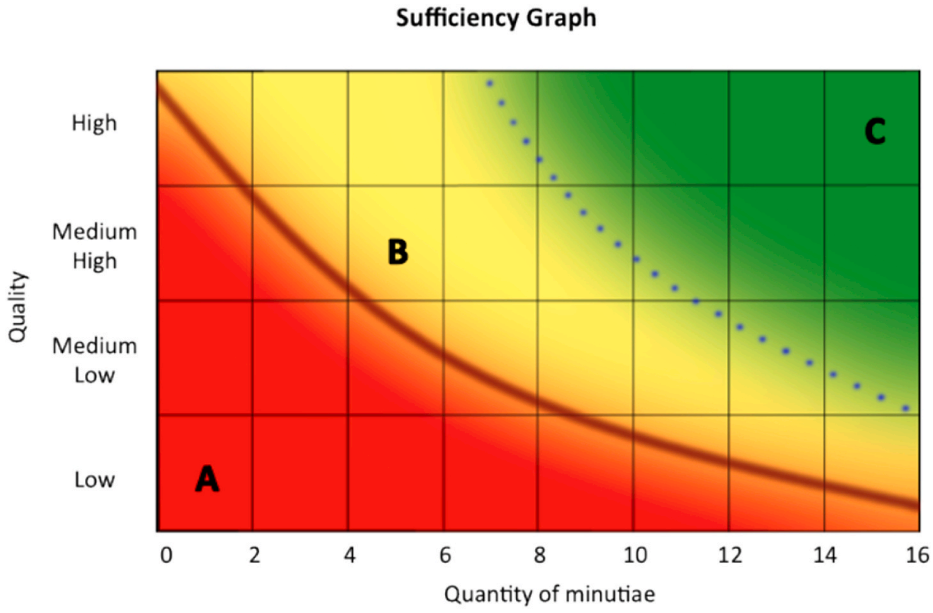

Fig. 5. The SWGFAST sufficiency chart. This graph provides guidance on when an identification may, or may not, be warranted, based upon the number of minutiae present and the quality of the image. In the area marked "A" an individualization decision is not warranted, whereas the dotted line separating areas "B" and "C" delineates those examinations that are considered complex (B) versus non-complex (C).

internal continuing training program in which a DNA mixture sample is diluted to the point at which signal begins to drop out.[3] This program has been separated from the annual Proficiency Testing program and is given to all examiners who are authorized to perform independent casework. While proficiency testing is subject to accreditation review and discovery, internal training exercises are intended to be a safe space in which to challenge examiners, allow errors, and learn to reach appropriate conclusions in difficult cases.

4.4. *Defending training errors in court*

One common fear shared by many examiners is having to answer the question, "Have you ever made an error?" during court testimony. They should not hesitate to answer, "Yes, while in training." This answer should be followed by a description of the training program. The examiner should be able to explain that their training program was progressive, moving through tiered difficulty levels that were structured to provide a challenge at every level. They could describe that many of the exercises they completed were forced-choice exercises in which no inconclusive response was allowed that were designed to provoke errors that could be discussed to improve their understanding and performance. Finally, they could reference some of the psychological literature covered in this article that describes the benefits to learning provided by a training program that encourages learning from errors.

Even errors made during continuing education training after being signed off for independent casework should be comfortable to discuss in this framework. The examiner can describe the continuing education program, emphasizing that its purpose it to stimulate continual improvement in examiners by challenging them in a safe environment where ground truth is known and using errors made in the training to identify areas for improvement and to improve skills in casework.

When put in the proper context, examiners should be able to demonstrate in court that errorful learning followed by correction is highly desirable in producing competent examiners who will improve their skills throughout their career, rather than allowing them to stagnate.

4.5. *Shifting attitudes toward error*

The forensic science community has begun making strides toward publicly acknowledging and addressing errors. The National Institute of Standards and Technology (NIST) hosted International Symposia on Forensic Science Error Management in 2015 and 2017. These symposia gave forensic science practitioners and researchers from around the world the opportunity to openly discuss ways to detect, measure, and mitigate forensic science errors in a non-judgmental environment.[4]

Some agencies have begun to publicly publish their Standard Operating Procedures, Training Manuals, and/or Quality Assurance documents.[5] Houston Forensic Science Center (https://records.hfscdiscovery.org/) has gone so far as to maintain a website that contains a record of all errors identified in their laboratory each year, and how each was resolved. These shifts toward transparency allow forensic scientists and forensic scientist supervisors within the laboratory system, and from outside agencies, to review and borrow methods to detect, measure, and mitigate forensic science errors in casework. They also allow attorneys and opposing experts the opportunity to examine the Quality System in place and understand the steps that are taken to minimize and mitigate errors.

These steps and others toward transparency in the process of

[3] The expectation is that examiners will receive a random aliquot containing different concentrations of the sample and due to stochastic effects such as drop-out, heterozygous imbalance, stutter, and possible contamination, they will arrive at different conclusions. The analysts are then expected to explain and defend their conclusions within the framework of their pre-established analytical thresholds and procedures.

[4] NIST, 2015 https://www.nist.gov/news-events/events/2015/07/international-symposium-forensic-science-error-management and NIST, 2017 https://www.nist.gov/news-events/events/2017/07/2017-international-forensic-science-error-management-symposium.

[5] For example, operations, quality, or training manuals; protocols; or CVs are available from the Indiana State Police (https://www.in.gov/isp/labs/2332.htm), Washington State Patrol (https://www.wsp.wa.gov/forensics/crimelab_docs.php), Idaho State Police (https://isp.idaho.gov/forensics/), Houston Forensic Science Center (http://www.houstonforensicscience.org/), and the Federal Bureau of Investigation (https://fbilabqsd.com/).

identifying and remediating errors will go a long way toward destigmatizing them and allowing errors to be seen as an opportunity for improvement – not only for individuals but also for laboratory policies and procedures (see Busey et al., this issue).

4.6. The role of a quality assurance program

Quality Assurance Programs are a critical tool for the identification and mitigation of errors in a forensic laboratory environment. A robust Quality Assurance Program will contain specific policies and procedures to be followed should an error be found and a corrective action become necessary. Corrective action plans can encompass root cause analysis, selection and implementation of corrective actions, monitoring of corrective actions, completion of additional audits of laboratory documents and casework, and having a strategy for preventive actions where potential future nonconformities are anticipated.

Although corrective action plans are most often implemented in casework, they could be extended into the training environment so expectations are clear of what the consequences will be if the trainee makes insignificant or significant errors during various phases of training. These 'consequences' need not all be negative; they should be scaled according to the stage of the training program and the severity of the errors and should be designed with learning theory in mind to best exploit the nature of the error to improve training outcomes. These could also be structured with a prevention or promotion regulatory focus in mind, as desired, to influence the motivation of the trainee [37]. Better awareness of consequences of training errors could serve to alleviate some of the 'error phobia' within the trainee.

Trainers may sometimes feel personally responsible for the errors of the trainee or may actually be held accountable by their agency for the performance of their trainee, and this could inhibit the trainer from the presenting challenging exercises needed to boost the trainee's learning. Providing a Quality Assurance Program with explicit expectations around error generation and mitigation for use during a training program could reduce the fear and stigma of making errors during training and allow for the potential benefits of error-generation to be maximized.

Of course, none of this is to suggest that all training errors should be given a pass and that all trainees will be suitable to successfully complete the training program and begin independent casework. There must be an appropriate balance between a trainee making acceptable and correctable errors and a trainee not grasping significant aspects of forensic science.

Additionally, Quality Assurance Programs will ideally contain policies and procedures to guide the conflict resolution process when differing opinions are found during any type of review.

5. Recommendations

5.1. Recommendations for further research

A review of the psychological literature in the area of Learning from Errors has revealed findings that may be of value to the forensic science community. However, it has also revealed that much research remains to be done. Most of the existing research in the literature has not been conducted in a forensic science context, or even a problem-solving or decision-making context, and thus its applicability to forensic science work is unknown. Following are some recommendations of areas where further research could provide information that is more directly applicable to real-world forensic science challenges.

- Learning from Errors concepts should be studied using a relevant forensic matching or decision-making task to test their applicability to problems that are more complex than memorizing word pairs. For example, research could feature close non-matches to promote errors, then a review of how the error(s) happened. Or, research could feature inconclusive decisions and discussion of where the inconclusive threshold is appropriate. Finally, further research into decision-making tasks where a forced choice model is employed may be illuminating, particularly in studying the hypercorrection effect
- The effects of an extended delay of feedback in a forensic domain should be studied to substantiate the optimal window for providing feedback to forensic trainees to maximize learning
- The effects of documentation for re-creating the mindset in which an initial error occurred should be examined to explore whether documentation requirements in training and casework may provide a learning benefit
- Research should be done to explore whether witnessing feedback to the errors of others helps, or whether the learning benefit is only observed for the person who made the error. This will help to inform the effectiveness of "hotwashing" and other group error review techniques for forensic science practice
- The impact of fear of errors on decision-making in forensic science should be examined. Specifically, research should consider both fear of management consequences and fear of emotional consequences for the examiner
- The use of prevention focus in training should be explored. Can an awareness of riskier decision-making environments (such as when the examiner is fatigued) be trained to induce a prevention focus that might reduce errors? In general, the assessment and manipulation of regulatory focus in forensic science training and practice should be explored

5.2. Recommendations supported by the current literature

While additional research in this area is certainly needed, there are recommendations that can be gleaned from the current psychological literature that should be immediately implementable. These suggestions should improve training outcomes in forensic science and lead to a reduction in errors in casework.

- Incorporate progressively challenging exercises designed to induce errors into training programs
- Deconstruct errors as they occur (both in training and casework) to understand how they happened and learn from them
- Have trainers sit down and engage with trainees when giving results, rather than just returning the results in a written format without discussion
- Present challenging training exercises to examiners for continuing education, even after initial competence has been established

6. Conclusion

Contrary to a popularly-held belief in American education, errors are essential for improvement, particularly when appropriate feedback on those errors is provided in a timely manner. This article provides the reader with an overview of the literature in cognitive psychology that supports the need for errorful learning during training, as well as suggests mechanisms for optimizing the benefits of the errors.

We hope that after reading this article, the reader will understand that errors are not a shameful failure of an individual examiner and should not be treated as such. Rather, errors made during training should be viewed as opportunities to learn and improve. Errors made during casework should be viewed similarly, along with being viewed as a way to illuminate possible systemic problems within the laboratory that should be examined and rectified while recognizing that a certain number of errors will always occur and the best we can do is try to minimize those errors and engage in appropriate risk management behaviors to minimize their impact.

In order to induce errors in a controlled environment and promote continued learning and improvement, challenging training exercises should be provided to examiners throughout their careers and the results of these exercises should be discussed and used for improvement.

However, the results of these exercises should not be used against examiners in a court setting as evidence of incompetence; rather, the examiner should be able to explain that the purpose of the exercise was to induce errors in a safe environment in order to improve skills and reduce errors in casework.

Naturally, not every employee is ideally suited to every task. There will be employees who, whether willfully, through laziness, or through sheer inability, will display a persistent pattern of errors that are resistant to correction and training. These employees were not the focus of this article. If a persistent pattern of errors that do not improve is observed, it is not the message of this article that those errors are useful (other than as an indicator of an unsuccessful employee) nor that the employee should be retained. However, it is the hope of the authors that appropriate errors during training become viewed as opportunities for growth and that they are encouraged and promoted as a way to ensure the training materials are sufficiently challenging and to test the limits of the examiner's ability. We similarly hope that laboratory management will take from this article an inspiration to include challenging continuing training exercises for their examiners throughout their careers.

Funding statement

This work was supported, in part, by the Forensic Technology Center of Excellence (Award 2016-MU-BX-K110), awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

Acknowledgements

The authors wish to thank Dr. Janet Metcalfe for her valuable contributions to the early phases of this project, including providing a starting literature list, insightful presentation, and fascinating conversations that all helped materially in shaping this article.

References

[1] J. Metcalfe, Learning from errors, Annu. Rev. Psychol. 68 (2017) 465–489.
[2] G. Langenburg, A Critical Analysis and Study of the ACE-V Process, PhD, School of Criminal Science, University of Lausanne, Lausanne, Switzerland, 2012.
[3] B.T. Ulery, et al., Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. U. S. A. 108 (19) (2011) 7733–7738.
[4] B.T. Ulery, et al., Repeatabiilty and reproducibility of decisions by latent fingerprint examiners, PLoS One 7 (3) (2012), e32800.
[5] National Commission on Forensic Science (NCFS), Views of the Commission: Facilitating Research on Laboratory Performance, (adopted Unanimously September 13, 2016), 2016.
[6] B. Skinner, Science and Human Behavior, MacMillan, New York, 1953.
[7] H. Terrace, Discrimination learning with and without errors, J. Exp. Anal. Behav. 6 (1963) 1–27.
[8] H. Stevenson, J. Stigler, The Learning Gap: Why Our Schools Are Failing and what We Can Learn from Japanese and Chinese Education, Simon & Schuster, New York, 1992.
[9] B.J. Huelser, J. Metcalfe, Making related errors facilitates learning, but learners do not know it, Mem. Cognit. 40 (4) (2012) 514–527.
[10] N. Kornell, J. Metcalfe, in: V. Benassi, C. Overson, C. Hakala (Eds.), The Effects of Memory Retrieval, Errors, and Feedback on Learning, in Applying Science of Learning in Education: Infusing Psychological Science into the Curriculum, Am. Psychol. Assoc. Soc. Teach. Psychol., Washington, DC, 2013.
[11] H.L. Roediger, J.D. Karpicke, Test-enhanced learning:taking memory tests improves long-term retention, Psychol. Sci. 17 (3) (2006) 249–255.
[12] P.J. Grimaldi, J.D. Karpicke, When and why do retrieval attempts enhance subsequent encoding? Mem. Cognit. 40 (4) (2012) 505–513.
[13] Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST). Standards for Examining Friction Ridge Impressions and Resulting Conclusions, Ver. 2.0. 2013 8.23.19]; Available from:: http://clpex.com/swgfast/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf.
[14] B. Finn, J. Metcalfe, Scaffolding feedback to maximize long-term error correction, Mem. Cognit. 38 (7) (2010) 951–961.
[15] A.C. Butler, H.L. Roediger, Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing, Mem. Cognit. 36 (3) (2008) 604–616.
[16] J. Metcalfe, N. Kornell, B. Finn, Delayed versus immediate feedback in children's and adults' vocabulary learning, Mem. Cognit. 37 (8) (2009) 1077–1087.
[17] W.T. Maddox, F.G. Ashby, C.J. Bohil, Delayed feedback effects on rule-based and information-integration category learning, J. Exp. Psychol. Learn. Mem. Cognit. 29 (4) (2003) 650–662.
[18] W.T. Maddox, A.D. Ing, Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning, J. Exp. Psychol. Learn. Mem. Cogn. 31 (1) (2005) 100–107.
[19] S.H.K. Kang, et al., Does incorrect guessing impair fact learning? J. Educ. Psychol. 103 (1) (2011) 48–59.
[20] B. Butterfield, J. Metcalfe, The correction of errors committed with high confidence, Metacognition and Learning 1 (1) (2006) 69–84.
[21] L.K. Fazio, E.J. Marsh, Surprising feedback improves later memory, Psychon. Bull. Rev. 16 (1) (2009) 88–92.
[22] J. Metcalfe, et al., Neural correlates of people's hypercorrection of their false beliefs, J. Cognit. Neurosci. 24 (7) (2012) 1571–1583.
[23] A.C. Butler, L.K. Fazio, E.J. Marsh, The hypercorrection effect persists over a week, but high-confidence errors return, Psychon. Bull. Rev. 18 (6) (2011) 1238–1244.
[24] D.M. Sitzman, et al., The role of prior knowledge in error correction for younger and older adults, Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 22 (4) (2015) 502–516.
[25] L.K. Fazio, E.J. Marsh, Correcting false memories, Psychol. Sci. 21 (6) (2010) 801–803.
[26] I.E. Dror, Human Expert Performance in Forensic Decision Making: Seven Different Sources of Bias, Australian Journal of Forensic Sciences, 2017, pp. 1–7.
[27] G. Langenburg, C. Champod, P. Wertheim, Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons, J. Forensic Sci. 54 (3) (2009) 571–582.
[28] Office of the Inspector General, *A Review Of the FBI's Handling Of the Brandon Mayfield Case*, U.S. Department of Justice, Washington, DC, 2006.
[29] P.J. Kellman, et al., Forensic comparison and matching of fingerprints: using quantitative image measures for estimating error rates through understanding and predicting difficulty, PLoS One 9 (5) (2014) e94617.
[30] T.A. Busey, et al., Accounts of the confidence-accuracy relation in recognition memory, Psychon. Bull. Rev. 7 (1) (2000) 26–48.
[31] N. Brewer, A. Keast, A. Rishworth, The confidence-accuracy relationship in eyewitness identification: the effects of reflection and disconfirmation on correlation and calibration, J. Exp. Psychol. Appl. 8 (1) (2002) 44–56.
[32] N. Brewer, G.L. Wells, The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates, J. Exp. Psychol. Appl. 12 (1) (2006) 11–30.
[33] J.E. Morrison, L.L. Meliza, Foundations of the after Action Review Process, in: United States Army Research Institute for the Behavioral and Social Sciences, Institute for Defense Analyses, Alexandria, VA, 1999.
[34] A.M. Jeanguenat, I.E. Dror, Human factors effecting forensic decision making: workplace stress and well-being, J. Forensic Sci. 63 (1) (2018) 258–261.
[35] A. Wu, Medical error: the second victim, BMJ 320 (2000) 726–727.
[36] L.N. Blum, J.M. Polisar, Why things go wrong in police work, Police Chief Magazine 71 (2004) 49–52.
[37] E. Crowe, T. Higgins, Regulatory focus and strategic inclinations: promotion and prevention in decision-making, Organ. Behav. Hum. Decis. Process. 69 (2) (1997) 117–132.
[38] A. Biedermann, S. Bozza, F. Taroni, Decision theoretic properties of forensic identification: underlying logic and argumentative implications, Forensic Sci. Int. 177 (2–3) (2008) 120–132.
[39] B. Zhao, Learning from errors: the role of context, emotion, and personality, J. Organ. Behav. 32 (3) (2011) 435–463.
[40] H. Eldridge, M. De Donno, C. Champod, Testing the accuracy and reliability of palmar friction ridge comparisons – A black box study, Forensic Sci. Int. (2020) 318, https://doi.org/10.1016/j.forsciint.2020.110457.
[41] H. Eldridge, M. De Donno, C. Champod, Mind-set – how bias leads to errors in friction ridge comparisons, Forensic Sci. Int. (2020) 318, https://doi.org/10.1016/j.forsciint.2020.110545.
[42] I.E. Dror, G. Langenburg, "Cannot decide": the fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide, J. Forensic Sci. 64 (1) (2019) 10–15.
[43] G. Langenburg, A performance study of the ace-V process: a pilot study to measure the accuracy, precision, reproducibility, repeatability, and biasability of conclusions resulting from the ace-V process, J. Forensic Ident. 59 (2) (2009) 219–257.
[44] President's Council of Advisors on Science & Technology, Report to the President, Forensic Science, in: Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Executive Office of the President of the United States, 2016. Washington, D.C.
[45] B. Max, J. Cavise, R. Gutierrez, Assessing latent print proficiency tests: lofty aims, straightforward samples, and the implications of nonexpert performance, J. Forensic Ident. 69 (3) (2019) 281–298.
[46] A.J. Koertner, H.J. Swofford, Comparison of latent print proficiency tests with latent prints obtained in routine casework using automated and objective quality metrics, J. Forensic Ident. 68 (3) (2018) 379–388.