LEARNING BY THINKING:
HOW REFLECTION CAN SPUR PROGRESS ALONG THE LEARNING CURVE

Giada Di Stefano
Bocconi University
Department of Management and Technology
Via Guglielmo Roentgen 1
20136 Milan, Italy
E-mail: giada.distefano@unibocconi.it

Francesca Gino
Harvard University
Negotiation, Organizations, and Markets Unit
Baker Library, Bloomberg Center 447
Boston, MA 02163, USA
E-mail: fgino@hbs.edu

Gary Pisano
Harvard University
Technology and Operations Management Unit
Morgan Hall 417
Boston, MA 02163, USA
Email: gpisano@hbs.edu

Bradley Staats
University of North Carolina
Kenan-Flagler Business School
CB #3490
Chapel Hill, NC 27599, USA
Email: bstaats@unc.edu

Acknowledgements: The authors are indebted to many for their important contributions during the eleven years of this paper's gestation. It has been a bumpy road, and we are grateful to Department Editor Greta Hsu and her team for seeing potential in the paper and pushing us to articulate it more clearly. A special thanks goes to the many who spent time providing comments throughout the process, especially Alfonso Gambardella, Andrea Coali, Chiara Bacilieri, Ella Miron-Spektor, Frederik Anseel, Hong Luo, Lamar Pierce, Maurizio Zollo, Michaéla Schippers, Saverio Favaron, Tomasz Obloj, and Riccarda Zezza. Thanks to participants in seminars at Bocconi University, Cornell University, Imperial College, INSEAD, London Business School, Warwick Business School, the Wharton School of Business, and WU Vienna, as well as attendees at a number of conferences over the years. Alfredo Desiderio, Chiara Taddei Santoni, Finia Jestaedt, and Sofia Santin, as well as Chiara Depalma and Saeid Kazemi, provided excellent research assistantship. The authors acknowledge financial support from Bocconi University, Harvard Business School, HEC Paris, and UNC Kenan-Flagler. This work would have not been possible without the substantial investment of time and attention of Devender Malhotra, Amit Rastogi, Rajesh Sehgal, Kartik Kamdar, Eben Samuel, and others at Wipro BPO.

Keywords: learning; reflection; knowledge articulation; knowledge codification; field experiment; laboratory experiment.

LEARNING BY THINKING:
HOW REFLECTION CAN SPUR PROGRESS ALONG THE LEARNING CURVE

ABSTRACT

It is common wisdom that practice makes perfect. And, in fact, we find evidence that when given a choice between practicing a task and reflecting on their previously accumulated practice, most people opt for the former. We argue in this paper that this preference is misinformed. Using evidence gathered in ten experimental studies (N = 4,340) conducted across different environments, geographies, and populations, we provide a rich understanding of the conditions under which the marginal benefit of reflecting on previously accumulated experience is superior to the marginal benefit of accumulating additional experience. We show that reflection has the potential to generate spillover effects to different but related tasks, and that reflection is mostly beneficial at the beginning of the learning curve, as long as one has accumulated a sufficient amount of experience on which to reflect. Interestingly, our study results also suggest that the way in which one engages in reflection may play a major role in its effectiveness as a learning tool. We test the robustness of the reflection effect to different tasks and its persistence over time in a series of additional studies.

INTRODUCTION

By three methods we may learn wisdom: First, by reflection, which is noblest; second, by imitation, which is easiest; and third, by experience, which is the bitterest.
Confucius

Imagine you have been given a math puzzle like the one in Figure 1 (initially developed by Mazar et al. 2008). You have 20 seconds to solve it by identifying the two numbers in the grid that sum to ten. After this initial practice run, you are given a choice of spending three minutes practicing more puzzles or three minutes reflecting on what you just did and how you could perform better in the future. Which option would you choose? We ran a pilot study to answer this question. A striking 82% of participants chose the first option (practice) over the second (reflection), presumably based on their intuition that gaining additional experience would lead to superior performance improvements as compared to reflection. What we argue in this paper is that this preference for experience over reflection is misinformed: Under some conditions, the marginal benefits of reflecting on previously accumulated experience outweigh those of accumulating additional experience. The question is, when?

- Insert Figure 1 about here -

The pilot experiment (reported more fully in Appendix 1) suggests individuals tend to think that the marginal returns of practice are superior to those of reflection.[1] This should not come as a surprise. The old

[1] Arguably, individuals could have engaged in cost/benefit considerations and chosen practice because of a higher perceived cost associated with reflection. For instance, one may find reflection more cognitively burdensome than practice, or one may simply prefer to keep doing the same activity instead of switching to a different one. Future work may be needed to unequivocally identify the drivers behind this choice, but the experimental design leads us to favor an interpretation based on considerations about expected

adage that "practice makes perfect" pervades a broad range of management theories and working methods, and the accumulation of experience has been extensively studied as a fundamental source of learning (e.g., Huckman and Pisano 2003, Argote and Miron-Spektor 2011, Clark et al. 2013, KC et al. 2013). The very idea of learning curves rests on the observation that experience fosters productivity and performance, both at the individual and organizational level (Argote and Epple 1990). But while experience is a fundamental source of learning, it is not the only one. Prior work suggests that reflecting on experience can also be a powerful tool for fostering progress along the learning curve (Seibert 1999, Ellis and Davidi 2005). The performance of individuals (Anseel et al. 2009, Ellis et al. 2014) and teams (Schippers et al. 2013, 2014) has been argued to improve when they are given the chance to engage in deliberate reflection on their experience. Evidently, experiential and reflective learning are deeply intertwined: Reflection can take place only if an individual has accumulated sufficient experience on which to reflect. That said, as soon as this condition is satisfied, experiential and reflective learning become mutually exclusive choices: Every additional minute spent reflecting is a minute of lost practice, and one more minute spent practicing means one less minute available for reflection. In light of this trade-off, we ask: What is the best way to allocate our learning time?

Approaching this question broadly, we argue that reflecting on accumulated experience generates higher performance outcomes than accumulating additional experience alone. We recognize that this main effect may be subject to a variety of moderating factors; we hence designed a series of experimental studies to provide evidence in support of a "reflection effect" and to explore its boundary conditions. We start with a field experiment (Study 1, N=101) carried out at Wipro BPO, an India-based global leader in the business-process outsourcing industry. We assigned Wipro agents to either a reflection or a practice condition during their on-the-job training and gathered evidence on their performance at the end of the training period. To overcome some inherent limitations in our design and provide a causal assessment of the effect of reflection on performance, we next designed an online experiment (Study 2, N=468) that replicated the pilot described at the start of this introduction, but this time with random assignment. We also included a control condition

returns. In fact, as explained in Appendix 1, participants were incentivized with a performance-based bonus, which plausibly led them to choose the strategy that could have helped them maximize their compensation.

aimed at excluding the possibility that our reflection effect was confounded with taking time off from the task at hand. To explore the boundary conditions of such a reflection effect, we ran a series of experiments through Prolific (Peer et al. 2017, Palan and Schitter 2018) after pre-registering with the Open Science Framework, where we shared materials, data, and statistical code.[2] In Study 3 (N=290), we removed the control condition and added another treatment for reflection, in which participants were asked to employ some of their reflection time to write down their reflections instead of just reflecting to themselves. In Study 4 (N=560), we used a more complex task to explore whether the benefits of reflection can spill over to related tasks. In Study 5 (N=1,203), we manipulated the amount of experience participants accumulated before engaging in the reflection effort. Table 1 provides an overview of the questions, motivations, findings, and conclusions of each study, including the pilot study reported at the beginning of this introduction and four additional studies discussed in the Appendix. Overall, the evidence in this paper is gathered from ten studies involving a total of 4,340 participants across different environments, geographies, and populations.

– Insert Table 1 about here –

Our findings support our conjecture that reflecting on previously accumulated experience generates higher performance outcomes compared to the accumulation of additional experience alone. We further show that reflection has the potential to generate spillover effects to related tasks, and that reflection is mostly beneficial at the beginning of the learning curve, as long as one has accumulated a sufficient amount of experience on which to reflect. Interestingly, our study results also suggest that the way in which one engages in reflection may play a major role in its effectiveness as a learning tool, with different results for participants who took time to simply think about their experience with a task and those who spent the same amount of time first thinking and then writing down their takeaways. We discuss the theoretical foundations of our work next.

THEORETICAL DEVELOPMENT

Much has been written about the role of experience in fueling both individual and organizational learning. People learn directly by acquiring experience and vicariously through its transfer from others (Bandura 1977,

[2] Please see the following link: https://osf.io/us92t/?view_only=4ef21e738e2741e6b599d24d534cc537.

Levitt and March 1988, Argote and Ingram 2000). Experience can accumulate from different tasks, from successes and failures, with more or less ambiguity, and at different points in time (Argote and Miron-Spektor, 2011). Experience can be a good or a bad teacher (March 2010), depending on whether we infer the "right" or "wrong" lessons from it (Levitt and March 1988). Independent of how we acquire it, the accumulation of experience tends to be associated with substantial performance improvements (Argote and Epple 1990). But as much as experience is necessary for learning, productivity, and performance, there are decreasing marginal returns to its accumulation. The question we ask in this paper is therefore simple: How can we maximize the learning benefit from our experience? Can we accelerate progress along the learning curve by substituting the accumulation of additional experience with a deliberate reflection exercise aimed at articulating and codifying the experience we have already accumulated?

We have some evidence that reflection can be a powerful learning tool. Building on prior work suggesting a role for reflection in generating performance improvements (Seibert 1999, Ellis and Davidi 2005), Anseel and colleagues (2009) designed two studies to test the benefits of receiving feedback on performance relative to reflecting on the feedback received. Their results suggest that feedback coupled with reflection generates higher performance improvements than feedback alone. Schippers and co-authors (2013) point out that team performance benefits from engaging in conscious reflection about how the team functions, especially when teams begin with relatively low levels of performance. Ellis et al. (2014) argue that in the case of successes, reflection can also be a prominent tool in learning from experience. Finally, Schippers and colleagues (2014) contend that deliberately discussing goals, processes, and outcomes in a team can reduce decision-making biases and errors, and hence potentially facilitate performance. A similar argument has been made at the organizational level, where prior work maintains that firms can improve performance by articulating and codifying their know-how (Zollo and Winter 2002). The concept has a direct application in the context of alliances and acquisitions, with Kale and Singh (2007) arguing that alliance managers can improve future alliance success if they spend time articulating actions and decisions from prior alliances.

Different mechanisms have been invoked to explain the effectiveness of reflection as a learning tool. The

phenomenon has neurobiological foundations: Research conducted with the use of fMRIs[3] (Nyberg et al. 2006, Olsson et al. 2008) shows that individuals can improve their performance on a task (i.e., motor ability) by simply imagining themselves executing it (i.e., mental training using motor imagery). One possible explanation for this finding is that internally focused thought (from daydreaming to effortful abstract thinking) activates a so-called default mode of neural processing that has been argued to be crucial in the development of cognitive abilities (Immordino-Yang et al. 2012). Beyond neurobiological explanations, prior literature has emphasized other advantages of reflection. For instance, Anseel and colleagues (2009) show that when individuals are given the chance to reflect on feedback about their prior performance, they experience higher self-efficacy, which may improve their subsequent performance. This idea resonates with a longstanding research tradition emphasizing that when individuals experience self-efficacy from an activity, they select more challenging tasks (Bandura and Schunk 1981), exert more effort (Schunk and Hanson 1985, Schunk et al. 1987), and have less adverse reactions when faced with difficulties (Bandura 1997), ultimately showing higher achievement (Multon et al. 1991). In a completely different context, Kale and Singh (2007) explain how reflection can facilitate ex-post sense-making, which may resolve causal ambiguity and facilitate a better understanding of cause-effect mechanisms. The idea behind after-action reviews (Ellis and Davidi 2005, Goh et al. 2013) is that providing feedback on both failures and successes enriches our mental models. Ellis and Davidi (2005, p. 859) provide a vivid depiction of this notion by describing our experiences as "databases from which [we] can make inferences to advance the generation of new propositions or draw evidence to confirm/refute old or new ones."

In short, prior work tells us that we need experience to learn, and that we can use reflection as a prominent tool to fuel progress along the learning curve. However, because time is a limited resource, when it comes to learning, we may face choices between accumulating more experience or reflecting on the experience previously accumulated—as in the choice between practicing with more math puzzles or reflecting on those you have solved so far. Or, borrowing from prior studies on learning curves (e.g., Edmondson et al.

[3] Functional magnetic resonance imaging, or functional MRI (fMRI), is used to detect physical changes in the brain resulting from increased neuronal activity.

2001, Huckman and Pisano 2003, KC et al. 2013), think about the case of a surgeon learning a less invasive surgical technique. For the surgeon to master the procedure as quickly as possible, would it be better for them to perform additional surgeries or to deliberately reflect on what they have learned and how they could do better? If we generalize from our pilot study, our hypothetical surgeon would choose the former, consistent with interview evidence collected by Edmondson et al. (2001) from 16 cardiac surgical teams attempting to learn a new minimally invasive heart procedure. Is such an approach the right strategy, however? The answer to this question, of course, depends on when the marginal returns of practice become inferior to the marginal returns of reflection—our next topic.

Comparing Reflective and Experiential Learning

In their influential piece on learning by doing *something else*, Schilling and colleagues (2003) show that learning rates are significantly greater when individuals work on related tasks than when they specialize in a single task or, alternatively, accumulate experience with different unrelated tasks. The researchers speculate this may be because of synergies between related learning efforts, which may trigger insights (Mayer 1996) and foster implicit learning (Wulf and Schmidt 1997). We build on a similar argument to suggest that reflecting on previously accumulated experience will yield higher returns compared to the accumulation of experience alone. Taking the time to articulate and codify previous experience, we argue, enables individuals to make the most of those experiences, as it allows them to unlock all the beneficial effects described by prior work: the activation of the default mode of neural processing, an increase in the perceived ability to complete the task at hand, and a richer mental model where relationships between cause and effect are clearer and surrounded by less uncertainty. These same benefits are simply not available to those who keep practicing. This is not to say, of course, that practice does not allow individuals to consciously abstract knowledge and create mental models; this is, indeed, the whole idea behind the insight and implicit learning mechanism argued for by Schilling et al. (2003). Our point is that *deliberately* engaging in reflection will yield superior returns. We hence hypothesize the following:

Hypothesis: Reflecting on previously accumulated experience generates higher performance outcomes than the accumulation of additional experience alone.

This hypothesis is clearly very broad in nature. Also, saying that combining experience and reflection yields higher performance benefits than accumulating experience alone makes intuitive sense. Still, we expect this main effect to be subject to a variety of moderating factors—conditions under which the overall benefit we believe is associated with combining experience and reflection may become more or less prominent.[4]

First, *how individuals reflect* may impact the effectiveness of reflection as a learning tool. In some prior work, reflection is defined as mere articulation: think, for instance, of the argument that the neurobiological foundations of reflection stem from internally focused thought, as per Immordino-Yang et al. (2012). In other studies, reflection entails both the articulation *and* codification of previously accumulated experience. An example is the theoretical model developed by Zollo and Winter (2002) emphasizing three learning mechanisms of experience accumulation, knowledge articulation, and knowledge codification. In simple terms, one first accumulates experience (i.e., do), then reflects by articulating previously accumulated experience (i.e., think), and finally can codify the results of such articulation efforts (i.e., write). Building on this three-pronged approach, we explore whether the addition of codification provides incremental value.

A second moderating factor to explore has to do with whether the marginal returns of reflection also benefit from the specialization versus related variation effect that Schilling and colleagues (2003) have documented for practice. According to the authors, learning by doing *something else* produces a superior boost in learning rates compared to learning by doing per se. Does the same apply to reflection? When we sit down and think about our experience with a task, are the benefits of such reflection exclusive to the original task, or can they spill over to different but related tasks?

Third, one may ask whether the marginal returns of reflection are superior to the marginal returns of practice at all *stages of the learning curve*. In other words, should we expect the marginal returns of reflection to change depending on how much experience one has accumulated? It could be argued that reflection is more beneficial than practice at the beginning of the learning curve, when an individual may need a boost in self-

[4] We are indebted to the editorial team for inspiring us to go beyond the main effect and explore potential moderating factors.

efficacy and there is room to resolve causal ambiguity. But it could also be argued that reflection will help more than practice when an individual is already familiar with a task and the marginal returns of experience are tapering off, allowing reflection to produce a more significant incremental increase in learning.

The above are the boundary conditions we will explore in this research. We recognize, of course, that there are many potential others. For instance, does the effect of reflection depend on the task at hand, on how accurately one understands it, or on how successful one has been at completing it? We leave these and other questions for future research to explore and focus our attention on (a) *how* individuals reflect—that is, whether reflection involves the mere articulation of accumulated experience or also a codification element of writing down what one is reflecting on; (b) *what* individuals reflect on—that is, whether the benefits of reflection are sticky or tend to spill over to related tasks; and (c) *when* individuals reflect—that is, whether the returns of reflection change depending on the amount of experience on which one can reflect. To provide evidence in support of our main hypothesis and explore these boundary conditions, we report the results of five experiments (N = 2,622), excluding the five additional experiments discussed in Appendices 1, 2, and 7, and featured in Table 1. The main features of these five focal experiments are summarized in Table 2, while Table 3 provides the exact wording of all our reflection manipulations across the five studies.

– Insert Table 2 and Table 3 about here –

STUDIES 1–2: IS THERE A REFLECTION EFFECT?

We have hypothesized that reflecting on previously accumulated experience will generate higher performance outcomes compared to the accumulation of additional experience alone. To provide evidence in support of this hypothesis, we designed two studies. Study 1 is a field experiment (N = 101) with agents undergoing on-the-job training at a large business-process outsourcing firm. In the study, we assigned participants to either a reflection or a practice condition, then observed their performance at the end of the training program and one month into their job. Study 2 is an online experiment (N = 468) that extends the pilot experiment described in the introduction. However, instead of giving participants the choice to practice or reflect, we randomly assigned them to either a reflection or a practice condition. We also added a control condition

aimed at excluding the possibility that our reflection effect was confounded with taking time off from the task at hand, since, according to Immordino-Yang et al. (2012), the same brain default system activated by reflection can be activated by rest.

Study 1: Is there evidence of the performance benefits of reflection vs. practice in the wild?

In Study 1, we explored the marginal benefits of reflection and practice in the field. Despite some limitations inherent to its design (on which we will elaborate), this experiment provides real-world evidence substantiating our conjecture about the relative performance returns of reflection and practice.

Design and procedure. Our field experiment was carried out at Wipro BPO, an India-based global leader in the business-process outsourcing industry. Wipro provides knowledge-based customer support and back-office services (e.g., data entry and data processing) for its global customer base. We conducted this study using one customer account, namely a Western technology company. The work for this account involved answering telephone calls from customers with technology-related support questions. The call center provided us with an excellent setting to study learning and performance outcomes at the individual level. Successful completion of the work requires technical knowledge on the part of Wipro employees. Questions can cover a wide range of topics; some can be answered easily, while others require a great deal of problem solving. To complete the work, Wipro not only recruits well-qualified agents (college graduates) but also trains them through a two-phase training program. As soon as an agent joins the firm, they start four weeks of "process training," a rather technical training fully delivered in the classroom. Next, agents transition to two weeks of "on-the-job training," a combination of classroom training and answering actual calls under the supervision of a trainer. After these six weeks of full-time training, agents transition into customer service. Our field study involved agents who completed their on-the-job training between June and August 2013.

Agents joined in six batches of 9 to 28 people, with each batch assigned to a different trainer.[5] We assigned each batch to one of two conditions: (1) reflection (treatment group) or (2) practice (control group).[6]

[5] Trainer 1: 10 agents; Trainer 2: 22 agents; Trainer 3: 24 agents; Trainer 4: 9 agents; Trainer 5: 10 agents; Trainer 6: 28 agents.
[6] We collected additional data on a third condition, in which we asked participants to spend the last 15 minutes of their training time (i) articulating and codifying the experience accumulated in the past (10 minutes) and (ii) sharing this knowledge with another

The first three batches (joining in June) were assigned to practice, while the last three (two joining in July and one in August) were assigned to reflection. The choice of administering our treatment at the batch level was motivated by our desire to decrease the risk of contamination. Our discussions with Wipro management made it evident that if we were to assign different treatments within batches, this would have likely induced agents to talk about the treatment and trainers to potentially interact with agents differently because of the treatment. We thus administered the same treatment to all agents assigned to the same trainer, even if this raised the issue of unobserved trainer heterogeneity, which we will return to later.

The final sample included 103 agents; two (from Trainer #6's group) left the program, leaving us with 101 participants, 56 in the treatment and 45 in the control group. Participants within each condition represented a similar profile of agents. They all went through the same process training and had identical on-the-job training experiences, except for our intervention, which we administered to the treatment group only. Starting from the sixth day of the two weeks of on-the-job training, agents assigned to the treatment group were given a paper journal and asked to spend the last 15 minutes of their training time articulating and codifying the experience they accumulated during the day. This was active training time substituted with reflection, prompted by instructions provided by their Wipro trainer (see Table 3). We kept the text of the manipulation broad enough to allow participants to reflect at the end of a full day handling customer calls, which arguably requires a broad set of skills. Agents assigned to the control group instead spent the last 15 minutes of their training time on their normal training activities—that is, they accumulated additional experience with the tasks associated with their job. The only difference between conditions was whether the last 15 minutes were spent reflecting (reflection condition, i.e., our treatment group) or doing the regular tasks associated with training (practice condition, i.e., our control group).

Measures. Our primary independent variable of interest was *reflection*, as manipulated in the experimental condition described above. Our dependent variable was participants' performance, which we measured in two ways. First, we looked at the score participants received in the test they took at the end of their training (*final*

participant (5 minutes). We included this condition to examine whether sharing would further increase the benefits of reflection. The results show that this is not the case. For ease of discussion, we do not report this additional analysis here, but in Appendix 3.

assessment). This is a test administered by Wipro to assess the extent to which trainees have learned the main lessons taught during the training. Scores, ranging from 0 to 100, were provided to us directly by the company. Second, we looked at customer satisfaction data, as rated by randomly sampled customers after each agent transitioned into their customer service responsibilities. In particular, we looked at a common measure in call centers, "top box," which is the percentage of randomly sampled callers who indicate their satisfaction in the highest category. *Customer satisfaction* captures this measure in the first month of customer-service work. Unfortunately, we could collect this measure for only 57 of the 101 participants.

In our analyses, we also included a series of controls for age (years), gender (male = 1, female = 0), and previous work experience (months), both at the level of the participant and at the level of the trainer to whom participants were assigned (see Appendix 3 for additional details). Table 4 reports mean comparisons across participants allocated to the treatment group versus the control group. Participants in the treatment group did not significantly differ from those in the control group in terms of individual characteristics. However, they differed when it came to the trainers to whom they were allocated: Trainers for the treatment group were older but had less work experience at Wipro compared to those for the control group. As for gender, five out of six trainers were male, with the only female trainer being allocated to one of the three batches in the treatment group. Effect sizes of such differences are not negligible, particularly for experience, thus suggesting a significant and sizable difference across trainers. This difference, coupled with the fact that our assignment was at the batch (i.e., trainer) level, suggests caution in interpreting the results of our analyses; we cannot completely rule out that the effect we estimate for reflection could be confounded with a trainer effect (see analyses reported and discussed below and in Appendix 3).

– Insert Table 4 about here –

Results. The t-tests reported at the bottom of Table 4 show that participants in the treatment group performed better both in the assessment test at the end of the training program ($M_{reflection}$ = 71.536, SD = 9.785; $M_{practice}$ = 54.422, SD = 20.715; $p < 0.001$) and in the customer satisfaction surveys completed for the first month after the training was over ($M_{reflection}$ = 0.912, SD = 0.236; $M_{practice}$ = 0.765, SD = 0.285; $p = 0.045$). The size of the effects is remarkable, with +16.934 over 100 points in the assessment test, and +0.147 over a

score of 1 in the customer satisfaction survey. To further explore the previously discussed differences across trainers, Figure 2 provides a graphical representation of the performance of participants by condition and by trainer using box plots and kernel density plots. We observed that among participants in the control group, the nine participants with Trainer #1 reached the highest performance, with five members of this batch being the only ones achieving a score equal to or greater than 85/100 on the final assessment test. These differences disappear in the long run, with participants in this batch being outperformed by two of the batches of the treatment group, including the batch assigned to Trainer #4, which did not perform particularly well in the final assessment test. Given these differences across trainers, we ran an ordinary least squares (OLS) regression with robust standard errors and controls at the trainer level, which we report in Table 5 and further discuss in Appendix 3. Participants in the reflection condition scored an average of 13.976 points more on their final assessment test than did participants in the practice condition (Model 1: $p < 0.001$, CI: 7.960, 19.992), which corresponds to a 21.9% increase with respect to the average score for the entire sample (63.911). However, we do not observe any significant difference one month after the training was completed (Model 2), though we suggest caution in drawing conclusions from this result, given the small N.

– Insert Figure 2 and Table 5 about here –

Taken together, the results from Study 1 seem to be in line with our conjecture that reflecting on previously accumulated experience will generate higher performance outcomes compared to the accumulation of additional experience alone. When completing the assessment test at the end of their training program, agents assigned to the reflection condition substantially outperformed their counterparts in the practice condition. However, the limitations inherent to our empirical design suggest caution in interpreting these results as a proper test of our hypothesis. We hence characterize them as suggestive of a reflection effect and move next to a laboratory experiment in which the treatment was administered at the individual level, thus allowing us to causally assess the effect of reflection on performance.

Study 2: Are the performance benefits of reflection vs. practice causal?

We designed a second study to gather causal evidence of the performance effects of reflection over practice, and to exclude the possibility that our finding could be influenced by the fact that reflection allowed participants to take time off from the task at hand. To these ends, we randomized participants into three conditions: reflection, practice, and control. This study also allowed us to test the robustness of our results with a broader subject pool (adults online versus participants in the Wipro training program) and a different incentive system (flat pay plus performance-based bonus versus no incentive, as per Study 1).

Design and procedure. We recruited 468 adults online to complete a brain teaser under time pressure in exchange for $1 and the potential to earn an additional bonus based on performance. Participants were recruited through Amazon Mechanical Turk (see Buhrmester et al. 2011 for a description), where they were provided with a brief description of the study and informed that 10% of them would receive a bonus of $1 based on their performance. The brain teaser consisted of a series of sum-to-ten games similar to the one described in the introduction. We gave participants 20 seconds per puzzle and let them first complete a practice round of five different puzzles. After completing each puzzle, participants were told if their answers were correct. Participants were then randomly assigned to one of three conditions: reflection, practice, or control. Compared to the manipulation used for Study 1 (Table 3), we manipulated reflection in a way that translates the broad concept of "lessons learned" to a very specific task. The goal was to make participants generate examples of presumed behavior based on past experience (Anseel et al. 2009) to deepen their information processing (Zollo and Winter 2002). We manipulated practice by asking participants to practice on some additional puzzles for three minutes. Finally, participants in the control group were instructed to watch an unrelated video (about cooking) that lasted about three minutes and were told they would be asked questions about it afterward. After three minutes, participants completed two other rounds of five puzzles each. We concluded by asking a few demographic questions.

Results. We conducted an ANOVA using participants' average performance on the brainteaser in the last two rounds as the dependent variable and condition as the independent variable, while controlling for performance on the first round (before our manipulation occurred). We found evidence of an effect, F (2, 464) = 10.46, $p < .001$, $\eta^2_p = .04$. Participants correctly solved more puzzles in the reflection condition

($M_{reflection}$ = 4.79, SD = 1.82) than in the practice condition ($M_{practice}$ = 4.07, SD = 2.29; p = .006) and the control condition ($M_{control}$ = 3.81, SD = 2.45; p < .001). Performance did not differ for participants in the practice versus control conditions.[7] As one might expect, performance in the first (practice) round predicted performance in the last two, F (2, 464) = 182, p < .001, η^2_p = .28. We interpret these results as providing causal support for our conjecture that reflecting on previously accumulated experience generates higher performance outcomes compared to the accumulation of additional experience alone.

STUDIES 3–5: UNPACKING THE EFFECT OF REFLECTION

We designed three experiments to explore three potential moderators of the reflection effect. Specifically, we attempted to understand whether the effect of reflection might depend on (a) *how* individuals reflect—that is, whether reflection involves the mere articulation of accumulated experience or also a codification element of writing down what one is reflecting on (see Study 3); (b) *what* individuals reflect on—that is, whether the benefits of reflection are sticky or tend to spill over to related tasks (see Study 4); and (c) *when* individuals reflect—that is, whether the returns to reflection change across stages of the learning curve (see Study 5). We next discuss the results of each study separately and then provide a general discussion integrating all findings.

Study 3: Do the performance benefits of reflection vs. practice depend on *how* individuals reflect?

Our two previous studies left participants free to manage the time they had for reflection without examining the process or outcomes of their reflection efforts. Study 3 tackled this point by comparing participants who took time to simply think about their experience with a task (a reflection condition we will refer to as articulation) with participants who spent the same amount of time first thinking and then writing down their takeaways (a reflection condition we will refer to as codification). The goal was to investigate whether reflection can be made more effective by acting on *how* individuals reflect—for instance, by requiring them

[7] The absence of a significant difference in performance across participants in the practice and control conditions raises a concern that participants in the practice condition did not use the time they had available to practice but rather wasted it doing nothing. We tried to mitigate this potential concern by introducing the performance-based bonus, which should have motivated participants to engage with the treatment to improve their performance, and hence the chance to double their compensation. In Additional Studies #3 and #4 (reported in Appendix 7), where we used a similar incentive scheme (flat fee + performance-based bonus for 10% of participants), results from a manipulation check show that participants accurately reported spending their time practicing.

not only to articulate their experience with a task but also to develop guidelines to better execute it. In other words, we aimed at understanding whether adding codification to articulation provided any incremental value.

Design and procedure. We recruited a total of 290 adults to complete a brain teaser under time pressure in exchange for £2.50 and the potential to earn an additional £1.00 bonus based on performance.[8] Participants were recruited through Prolific, which provided them with a brief description of the study and explained that 10% of them would receive the extra bonus based on their performance. To be eligible, participants had to be located in an English-speaking country and have a good track record on the platform.[9] We used the same math puzzles employed for Study 2 (sum-to-ten games), with the same basic structure across three rounds: five math puzzles (practice round), five math puzzles (round 1), and five math puzzles (round 2). We decided to contrast practice with two reflection conditions.[10] For the first reflection condition (articulation), participants were asked to think about what was required of them and reflect on their performance. This articulation effort has been argued to deepen information processing (Zollo and Winter 2002) by making participants generate examples of presumed behavior based on past experience (Anseel et al. 2009). Compared to the manipulation used for Study 2, where we asked about strategies (see Table 3), here we asked participants to reflect on what was required of them and on their performance. This is arguably a more basic formulation than the one used in the previous study, but it helped us better distinguish articulation from codification by focusing on the idea of articulating past behavior. For the second reflection condition, we stepped up the learning effort from a simple sharing of experiences to a codification of learning (Zollo and Winter 2002). In this case, we asked participants to first engage in articulation and then to codify this knowledge by developing guidelines for the execution of the task. By doing so, we encouraged them to discover linkages between actions and associated performance outcomes (Zollo and Winter 2002). Finally, participants in the practice condition were given three minutes to practice the task they had just completed.

[8] We recruited 304 participants and rejected those who failed the four attention checks we had inserted, a total of 14 participants.
[9] We selected participants who (a) had taken part in at least ten studies, (b) had an approval rate of 100%, (c) held at least a high school diploma, and (d) could participate using a laptop or PC.
[10] We collected data on an additional manipulation to start exploring the existence of spillover effects for the benefits of reflection—a conjecture we further explored in Study 4. Participants in this condition received the same 3 x 4 grids, but in the very last round they were asked to find the smallest and largest numbers among the twelve displayed on the grid–a task participants found much easier and for which reflection did not help. For ease of discussion, we do not report this additional analysis here but in Appendix 4.

Measures. We manipulated *articulation*, *codification*, and *practice* as described above, and identified each treatment with a dummy variable equal to 1 in the case of "high" and 0 otherwise. Our dependent variable, *final performance,* was a count variable ranging between 0 and 10 corresponding to the number of math puzzles solved correctly in the last two rounds. We controlled for performance in the practice round (*initial performance*). We gathered data about age, gender, education, employment, and socioeconomic status. Giving credit to the exceptional time during which the study was conducted (i.e., during the Covid-19 pandemic), we also measured neuroticism, a personality trait that has been argued to play a crucial role in the extent to which an individual's behavior is affected by a pandemic (Taylor 2019). Results from a series of t-tests show that all characteristics for which we controlled were evenly distributed among participants.[11]

Results. To compare the performance of participants across the different treatment groups, we first looked at model-free evidence. Figure 3 provides a graphical representation of the performance of participants by condition using box plots and kernel density plots. We contextually ran a series of t-tests to better understand the differences between means. We observed that participants in the articulation condition performed better than those in both the practice condition ($M_{articulation}$ = 6.402, SD = 2.115; $M_{practice}$ = 5.830, SD = 2.265; p = 0.036) and the codification condition ($M_{articulation}$ = 6.402, SD = 2.115; $M_{codification}$ = 5.747, SD = 2.443; p = 0.023). The graphs show the upward shift in performance, with the bulk of the distribution displaying higher minimum values and smaller variance in performance outcomes. Interestingly, not only were participants in the codification condition at a disadvantage compared to those in the articulation condition but they also did not seem to experience any substantial advantage compared to those who practiced ($M_{codification}$ = 5.747, SD = 2.443; $M_{practice}$ = 5.830, SD = 2.265; p = 0.596). To investigate the performance effects of our two treatments more formally, we ran a series of regression analyses, which we discuss in Appendix 4. Results across models show that *articulation* has a positive effect on *performance* but that participants assigned to the *codification* condition only experienced an increase in performance when they had an accurate understanding of the task at hand. This connects to our original conjecture that the marginal

[11] We detected two significant but small differences in the codification group, where participants were slightly older (Cohen's d = 0.21) and had a slightly different representation of nationalities (Cohen's d = 0.18).

benefits of reflection may differ at different stages of the learning curve—a conjecture we explore in Study 5.

– Insert Figure 3 about here –

Study 4: Do the performance benefits of reflection vs. practice depend on *what* individuals reflect on?

We designed a replication of Study 3 to explore whether the benefits of reflection are sticky or can spill over to related tasks. To this end, we devised a more complex experimental task that allowed us to explore some variations of it. Specifically, instead of the math puzzles used in the previous two studies, which test simple addition skills, we employed Raven matrices, which are used to assess abstract reasoning more broadly (Raven 1989, 2000). Given that organizations use them widely to screen applicants, a performance uplift on the matrices could potentially have real-life implications for job performance. Importantly for our purposes, there are different types of Raven matrices, whose solutions are based on the application of different strategies. We leveraged this feature to design a study in which, after going through a practice round and the administration of our treatment, participants were asked to solve the same task or to move to a different but related one.[12]

Design and procedure. We recruited 560 adults to complete a brain teaser under time pressure in exchange for £2.63 and the potential to earn an additional £1.00 bonus based on performance.[13] Participants were recruited through Prolific, which provided them with a brief description of the study and explained that 10% of them would receive the extra bonus based on performance in the study. To be eligible, participants had to be located in an English-speaking country, have a good track record on the platform,[14] and not have participated in any study we had previously run. To keep the overall duration of the study similar, and since Raven matrices take longer to solve (we gave participants 45 seconds per matrix with a mandatory five-second break in between), we adapted the structure of Study 3: We used a first round of five matrices to measure initial performance and a second round of five matrices to observe the effect of our treatments.

---

[12] While running Study 4, we collected additional data to start exploring potential mechanisms behind the effect of reflection. The analysis is exploratory in nature since we measured, rather than manipulated, potential mediating variables. This is standard procedure in the field (Spencer et al. 2005), but it has been consistently shown to bias estimates through measurement error and omitted variable bias (Pierce and Snyder 2020). As a result, we present the analyses of mediation as preliminary, descriptive evidence of the role of potential mechanisms that have been put forward by prior work. We report the analyses in Appendix 5.

[13] We recruited 601 participants and rejected 41 participants who failed the four attention checks we inserted.

[14] We selected participants who (a) had taken part in at least ten studies, (b) had an approval rate of 100%, (c) held at least a high school diploma, and (d) could participate using a laptop or PC.

We used a 3 (condition: articulation vs. codification vs. practice) x 2 (task: same vs. related) between-subject factorial design. The first treatment contrasted practice with articulation and codification using the exact same manipulations adopted in Study 3. To observe whether the beneficial effects of reflection spill over to related tasks, we introduced a second treatment that we administered by acting on the matrices participants were asked to solve: All participants attempted to solve the same five matrices for the practice round. These consisted of Raven matrices of the "distribution of three values" type (Carpenter et al. 1990): The nine figures that appear in the grid represent all the possible combinations of two features that take three different values. Matrix (a) in Figure 4 is characterized by the distribution of three types of background (white, diagonal, black) for both the vertical and horizontal shapes. After receiving the treatment, participants in the same-task condition received five other matrices of the same category, while participants in the related-task condition received five matrices of the "figure addition and subtraction" type (Carpenter et al. 1990): The nine figures that appear in the grid are built by adding or subtracting features from one another. Matrix (b) in Figure 4 is an example of figure addition: The bottom cell of each column is formed by adding the images in the two cells above. Matrix (c) in Figure 4 is an example of figure subtraction: The bottom cell of each column is derived by subtracting from the top cell of the column the image in the cell below. To make the two sets of matrices as comparable as possible, we relied on the probability of solving each of them as measured by Wytek et al. (1984) and Arthur and Day (1994). Given the impossibility of achieving the exact same probability for the two groups, we preferred to keep the related matrices slightly easier: The probability of solving the matrices correctly was 76% for the same task and 82% for the related task.

- Insert Figure 4 about here -

***Measures*.** We manipulated condition and task as described above, identifying each treatment with a dummy variable equal to 1 in the case of "high" and 0 otherwise. Our dependent variable, *final performance,* was a count variable ranging between 0 and 5 corresponding to the number of Raven matrices that were solved correctly in the last round. We controlled for performance in the practice round (*initial performance*). We

gathered the same demographic data as in Study 3, including our pandemic control. Results from a series of t-tests show that all characteristics for which we controlled were evenly distributed among participants.[15]

Results. To compare the performance of participants across the different treatment groups, we first looked at model-free evidence. Figure 5 provides a graphical representation of the performance of participants by condition using box plots and kernel density plots. We contextually ran a series of t-tests to better make sense of the differences between means. Overall, we observed that participants in the practice condition were outperformed by participants in both the articulation ($M_{articulation}$ = 3.650, SD = 1.167; $M_{practice}$ = 3.354, SD = 1.428; p = 0.014) and the codification ($M_{codification}$ = 3.578, SD = 1.227; $M_{practice}$ = 3.354, SD = 1.428; p = 0.052) conditions. When we distinguished between participants based on the type of task (same vs. related), we observed, not surprisingly, that participants overall performed better when they were asked to solve the same type of matrices in the second round ($M_{sametask}$ = 3.864, SD = 1.064; $M_{relatedtask}$ = 3.172, SD = 1.397; $p < 0.001$), independent on whether participants were asked to practice, articulate, or codify. When participants were asked to solve the related type of matrices, however, those assigned to the reflection conditions did better than those assigned to practice, independent on whether they engaged in articulation ($M_{articulation}$ = 3.400, SD = 0.126; $M_{practice}$ = 2.802, SD = 0.163; p = 0.002) or codification ($M_{codification}$ = 3.315, SD = 0.142; $M_{practice}$ = 2.802, SD = 0.163; p = 0.009).

To investigate the performance effects of our two treatments more formally, we next ran regression analyses, the results of which we report in Table 6. More specifically, we ran an OLS regression with robust standard errors where we examined differences in *final performance* while controlling for *initial performance*. We started by estimating the effects of *articulation* and *codification* vis-à-vis *practice*, as well as the effect of *same* vis-à-vis *related task* (Model 1). We next explored the interaction between the type of task assigned and *articulation* as well as *codification* (Model 2). To visualize differences across tasks more intuitively, we also ran a split sample analysis where we separately looked at the effects of *articulation* and *codification* vis-à-vis *practice* for participants who completed the same vis-à-vis *related task* (Model 3 and Model 4, respectively). Results from our

[15] We detected a significant but small difference related to lower neuroticism for participants assigned to practice (Cohen's d = 0.17). We also detected two smaller differences in the same task group, where participants who identified as women were slightly more represented (Cohen's d = 0.14) and participants came from a slightly less advantaged socioeconomic background (Cohen's d = 0.11).

regression analyses show that, compared to *practice*, *articulation* had a positive effect on *final performance* (Model 1: ß = 0.268, SE = 0.118, p = 0.024), and *codification* marginally did as well (Model 1: ß = 0.230, SE = 0.122, p = 0.059). Not surprisingly, participants performed considerably worse when asked to solve a different type of matrix in the second round (Model 1: ß = -0.724, SE = 0.096, p < 0.001). However, in line with the model-free evidence, we observe this disadvantage to attenuate for participants who were assigned to articulation (Model 2: ß = 0.498, SE = 0.052, p = 0.026) and to do so marginally for those assigned to codification (Model 2: ß = 0.534, SE = 0.238, p = 0.081).

- Insert Figure 5 and Table 6 about here -

Study 5: Do the performance benefits of reflection vs. practice depend on *when* individuals reflect?

We designed a replication of Study 4 to explore whether the marginal returns of reflection change depending on the amount of experience on which one can reflect. To this end, we employed the same experimental task but manipulated the amount of experience we let participants accumulate before engaging in the reflection effort. To avoid any confusion between experience accumulated with the task and our practice condition, we will refer to the former as the level of *familiarity* of our participants with the task at hand.

Design and procedure. We recruited a total of 1,203 adults to complete a brain teaser under time pressure in exchange for £3.00 and the potential to earn an additional £1.20 bonus based on performance. Participants were recruited through Prolific, which provided them with a brief description of the study and explained that 10% of them would receive the extra bonus based on their performance. To be eligible, participants had to be located in an English-speaking country, have a good track record on the platform,[16] and not have participated in any study we had previously run. To allow some participants to accumulate more experience with the task, we adapted the structure of Study 4 and designed a first round of up to eight matrices, followed by a second round of four matrices to observe the effect of our treatments. Our experimental design was a 4 (familiarity: two, four, six, or eight matrices) x 3 (condition: articulation vs. codification vs. practice) between-subject

[16] We selected participants who (a) had an approval rate of at least 90%, (b) held at least a high school diploma, and (c) could participate using a laptop or PC.

factorial design. To observe the effect of familiarity, we treated the number of matrices participants were asked to solve in the practice round. In particular, depending on the experimental cell to which they were assigned, participants were asked to solve two, four, six, or eight matrices, with 45 seconds to solve each matrix and a mandatory five-second break in between. To compensate for the higher fatigue associated with solving more matrices, participants assigned to the first three familiarity levels were given, respectively, six, four, or two filler tasks to solve before they got started on the matrices. Filler tasks were simple puzzles that required participants to find the piece that fit with the graphical pattern, as per Figure 6. We selected this task because it looked similar enough to the experimental task but required a different type of reasoning (matching of a graphical patterns vs. understanding of a logical sequence).

- Insert Figure 6 about here -

After completing six filler tasks and two matrices (familiarity level 1), four filler tasks and four matrices (familiarity level 2), two filler tasks and six matrices (familiarity level 3), or zero filler tasks and eight matrices (familiarity level 4), we introduced our main treatment, which contrasted practice with articulation and codification using the manipulations adopted in Studies 3 and 4, with slightly simplified wording (see Table 3). More substantially, for the codification treatment, we removed the separation between articulation and codification: Instead of giving participants one minute to articulate and two minutes to codify, we instructed them to articulate and codify but left them free to administer their three minutes according to their own preferences so that they could reflect more organically, potentially going back and forth between articulation and codification. After receiving this treatment, participants were asked to solve a second round of four matrices, which we used to compare performance across the different experimental cells. A crucial choice we had to make was related to the selection of matrices our participants would practice on as well as the ones they would be asked to solve in the final round. As in Study 4, we selected matrices based on the probability of solving each of them (Arthur and Day, 1994). Matrices were ordered from the easiest to the most difficult, with participants facing slightly more difficult matrices in the four familiarity groups: The probability of correctly solving the matrices was, respectively, 93.32% for group 1, 92.45% for group 2, 91.99% for group 3,

and 91.09% for group 4. The difficulty of matrices increased substantially in the second round, where participants were asked to solve matrices that had a probability of being correctly solved equal to 80.69%.

Measures. We manipulated articulation, codification, and practice as described above, and identified each treatment with a dummy variable equal to 1 in the case of "high" and 0 otherwise. As for familiarity, we created a variable ranging from 1 to 4, depending on the number of matrices (two, four, six, or eight) participants solved before we administered our main treatment. Our dependent variable, *final performance,* was a count variable ranging between 0 and 4 corresponding to the number of Raven matrices solved correctly in the last round. We controlled for performance in the practice round, in this case averaged for the number of matrices participants had to solve depending on the familiarity treatment (*initial performance*). We gathered the same demographic data as in Studies 3 and 4, including our pandemic control. Results from a series of t-tests show that all characteristics for which we controlled were evenly distributed among participants.[17]

To further explore if how participants reflected impacted performance, we codified the reflection logs of participants assigned to the codification condition. To this end, we first went through the reflection logs and inductively developed meaningful coding categories. We then hired two independent coders to codify two broad classes of information. First, they measured the *length* of the reflection log expressed in number of words (count). Reflection logs were 29 words long on average, with a minimum word count of three (i.e., one participant simply wrote: "Observation and deduction") and a maximum of 157. The following is an example of a reflection log of remarkable length:

> *First, I took a global look at the pattern, then separated my vision to focus on one column at a time. I noticed how the first column on the left progressed from top to bottom, adding a new element to the image in each row. Elements were added symmetrically, so I applied the same logic to complete the missing piece in the third column, maintaining elements of the same pattern. To be effective, you have to observe the data provided, form a hypothesis, and check that hypothesis against the available data. Then, using the answers offered, see if any of those answers might fit the hypothesis correctly. If so, then I used process of elimination to choose the answer that would correctly fit my hypothesis/observation. Elements are usually added or taken away from the images.*

Length is certainly suggestive of a more informative reflection log. However, we came across many examples of shorter but still very informative reflection logs. For instance, one participant simply wrote: "*I look for movements, repetitions, and orders. I also try to eliminate some choices.*" This observation pushed us to also

[17] We detected a significant but small difference related to higher education for participants assigned to practice (Cohen's d = 0.04).

codify the content of the reflection logs (see Appendix 6) in terms of whether they mentioned the *task* at hand (dummy), the *strategy* used to solve it (dummy), and the participant's *performance* on the task (dummy). To code the content of the reflection logs, our coders examined the first few logs independently, discussed cases of disagreement, and then independently coded all 403 reflection logs, one for each of the participants assigned to the *codification* condition. The inter-rater agreement was substantial, with a percent agreement of at least 83% with an almost perfect (80–100%) benchmark interval.

Results. To compare the performance of participants across the different treatment groups, we first looked at model-free evidence. Figure 7 provides a graphical representation of the performance of participants by condition (articulation vs. codification vs. practice) across different familiarity levels (i.e., whether participants practiced on two, four, six, or eight matrices during the practice round). We plotted the average second-round scores for the 12 experimental groups and observed that the benefits of reflection versus practice change over time: Participants in the practice condition got a jump start on those in the reflection conditions when the treatments were administered at the very beginning of the learning curve (familiarity level 1). However, when reflection happened after participants had gained some experience with the task (familiarity level 2), those who reflected (particularly through codification) benefited the most. Differences across conditions seem to become immaterial for higher levels of familiarity (familiarity levels 3 and 4). To investigate these broad patterns more formally, we next ran regression analyses. We investigated differences in *final performance* while controlling for *initial performance* by running OLS regressions with robust standard errors, the results of which are shown in Table 7. We started by estimating the effects of *articulation* and *codification* vis-à-vis *practice*, as well as the effect of *familiarity* (Model 1). We next included the interaction terms of *familiarity* and *articulation* as well as the interaction terms of *familiarity* and *codification* (Model 2). We further unpacked such interactions to observe differences across all the different levels of *familiarity* (Model 3). Finally, we replicated the same analysis using the time taken to solve the matrices in the final round as our dependent variable (Model 4). The results show that, not surprisingly, *familiarity* had a positive effect on performance (Model 1: ß = 0.173, SE = 0.029, $p < 0.001$), with no obvious interaction with *articulation* (Model 2: ß = 0.062, SE = 0.072, p = 0.386) or *codification* (Model 2: ß = 0.028, SE = 0.074, p = 0.700). Our results

further suggest that participants who had the least experience with the task (familiarity level 1) benefited more from *practice* compared to *articulation* (Model 3: ß = -0.360, SE = 0.161, p = 0.025) or *codification*, even if marginally (Model 3: ß = -0.290, SE = 0.171, p = 0.090) in the latter case. On the other hand, participants who reflected after they gained a more substantial level of experience with the task at hand (familiarity level 2) experienced a significant performance boost. This is particularly true for those who engaged in *codification* (Model 3: ß = 0.577, SE = 0.236, p = 0.015) and marginally so for those who engaged in *articulation* (Model 3: ß = 0.444, SE = 0.240, p = 0.065). As the level of previously accumulated experience further increased (familiarity levels 3 and 4), we found no difference across conditions, with *familiarity* having the most impact on performance for participants who had gained the most experience with the task (familiarity level 4; Model 3: ß = 0.388, SE = 0.170, p = 0.023). Looking at the time participants took to solve the second-round matrices, we observe that when participants were least familiar with the task (familiarity level 1), those who reflected were substantially slower than those who practiced, independent of whether they reflected through *articulation* (Model 4: ß = 10.076, SE = 3.885, p = 0.010) or *codification* (Model 4: ß = 12.109, SE = 4.727, p = 0.011). Indeed, participants in the reflection conditions took between 10 and 12 seconds more to complete a round than those who practiced, with the latter group taking an average 115 seconds. The time disadvantage disappears when the level of familiarity increases.

- Insert Figure 7 and Table 7 about here -

Next, we focused on participants assigned to the codification condition only and examined associations between characteristics of their reflection logs and their performance. In Table 8, we present descriptive evidence of the average length of reflection logs as well as the extent to which participants focused their logs on a discussion of the task at hand, the strategy used to solve it, and their performance with it. We find it interesting that the figures are quite similar across familiarity levels, thus suggesting no specific patterns in the relationship between how much experience one has gained with the task and what one tends to focus on when reflecting. To formally explore associations with performance, Table 9 shows what happened when we ran an OLS regression with robust standard errors looking at the effect of *length* of the reflection logs, as well as their content in terms of *task*, *strategy*, and *performance* for participants assigned to the codification condition

across familiarity levels and for each familiarity level. Results from the analysis show that participants who wrote longer reflection logs (ß = 0.007, SE = 0.002, p < 0.001) and explicitly elaborated a strategy for solving the task at hand (ß = 0.414, SE = 0.198, p = 0.037) also performed better. The effect of *length* is mostly driven by participants who had the least experience with the task (familiarity level 1: ß = 0.014, SE = 0.004, p < 0.001), while the effect of *strategy* mostly results from the experience of participants who reflected after they had gained a more substantial level of experience with the task at hand (familiarity level 2: ß = 0.756, SE = 0.344, p = 0.030). As the level of previously accumulated experience increased further (familiarity levels 3 and 4), we found no evidence of any association between how participants reflected and how they performed.

- Insert Table 8 and Table 9 about here -

GENERAL DISCUSSION: WHAT DO OUR STUDIES TELL US ABOUT REFLECTION?

We opened the paper with a broad question: If you had to choose between accumulating additional experience with a task and reflecting on your previously accumulated with the task, what would be the best way to allocate your learning time? We broadly hypothesized that reflecting on accumulated experience would generate higher performance outcomes than accumulating additional experience alone. Yet we also recognized that this main effect might be subject to a variety of moderating factors. Thus, we conducted a series of experimental studies aimed at providing evidence in support of a "reflection effect" and exploring its boundary conditions.

We started with a field experiment (Study 1, N=101) and an online experiment (Study 2, N=468) aimed at showing the existence of a reflection effect descriptively in the wild and causally in the lab. Results from both studies provided support for our conjecture: When allocated to the reflection condition, Wipro trainees and mTurk workers who were asked to spend time reflecting on the experience they had accumulated so far outperformed counterparts who focused their efforts on accumulating more practice. Two additional studies reported in Appendix 2 successfully tested the robustness of the reflection effect to different types of tasks and incentive schemes. The pilot study reported in Appendix 1, moreover, suggests that the superior performance benefits associated with reflection are not intuitive, as an overwhelming majority of participants

chose practice over reflection when given the chance to make their own selection.

Once we established a reflection effect, we started exploring potential boundary conditions. In Study 3 (N=290), we investigated the performance returns associated with different ways in which individuals can reflect. To this end, we kept the same experimental task (math puzzles) but compared participants who practiced with those who took time to simply think about their experience with a task (the "*articulation*" condition) and those who spent the same amount of time first thinking and then writing down their takeaways (the "*codification*" condition). Results from our analyses suggest that articulation generates superior performance benefits compared to practice as well as codification. Interestingly, codification did not appear to bring any benefit compared to practice or mere articulation.

In Study 4 (N=560), we examined whether benefits to reflection are sticky—that is, whether they are helpful only when dealing with the exact same task or can spill over to related tasks. To this end, we kept the distinction between articulation and codification, but changed the experimental task to a more complex one (Raven matrices) to introduce a condition in which, after practice or reflection, participants faced a variation on the main task (that is, a type of Raven matrix that required a different strategy to be solved). Results from our analyses suggest that participants performed better overall when asked to solve the same type of matrices in the second round, independent of whether they were asked to practice, articulate, or codify. When participants were asked to solve the related type of matrices, however, those assigned to practice were substantially outperformed by those assigned to articulation, and marginally outperformed by those assigned to codification. Note that in this case, different from Study 3, codification appeared to bring some benefits compared to practice.

In Study 5 (N=1,203), we explored how the marginal returns of reflection change depending on the amount of experience one has accumulated before engaging in the reflection effort. To this end, we kept the more complex task (Raven matrices) as well as the distinction between articulation and codification, but for the codification treatment, instead of giving participants one minute to articulate and two minutes to codify, we instructed them to articulate and codify but left them free to administer their three minutes according to their own preferences. Our results show that participants who had little experience upon which to reflect

benefited the most from additional practice and that practice and reflection were perfect substitutes for expert participants. Reflection seemed to make the most significant difference when participants were at the beginning of their learning curve but had accumulated enough experience on which to reflect. In these cases, we observed that codification, and to some extent articulation, provided a significant boost in performance without coming at the expense of speed. Note that in this case, different from Study 3 and Study 4, codification brought more benefits than practice as well as mere articulation.

Many explanations may drive the partially inconsistent results related to whether it is more beneficial to engage in articulation or codification when reflecting, including the fact that we changed the task from Study 3 (i.e., math puzzles) to Studies 4–5 (i.e., Raven matrices) and slightly modified the experimental manipulation for codification from Studies 3–4 (i.e., one minute to articulate and two minutes to codify) to Study 5 (i.e., three minutes to articulate and codify). With respect to the experimental task, our data suggest that participants found Raven matrices more difficult than math puzzles: The results from Studies 3 and 4 (same task condition only) show that the performance increase between the round of practice and the last round equaled +33% for math puzzles (-11% of time per puzzle) compared to +5% for Raven matrices (-1% of time per matrix). In light of this observation, we interpret our experimental results as suggesting that articulation alone is sufficient when dealing with a relatively simpler task (as in Study 3),[18] but that engaging in a more systematic codification effort may be beneficial when facing a more complex task (as in Studies 4 and 5). Results from two additional studies, reported in Appendix 7, seem to support this conjecture, showing that the benefits of codification emerged only when participants felt challenged by the task with which they were engaged.

If codification may be beneficial only under some conditions, taking time away from articulation to engage in codification may not always be worth it. In Study 5, by removing the mandatory break between the time participants were articulating their thoughts and the time they were writing them down, we allowed codification to happen in a more organic way and not necessarily at the expense of articulation, with participants potentially

[18] The results from Appendix 4 seem to suggest that articulation may not be beneficial in those cases in which, instead of being challenged to make progress along the learning curve, one is asked to step back and engage with an easier task.

going back and forth between articulating and codifying. The fact that we observed a stronger codification effect with this manipulation seems to support our conjecture that the way in which participants engage in codification may play a major role in the effectiveness of codification as a reflection approach. Results from descriptive analyses seem to suggest that participants who were just learning about a task benefited from elaborating about it in more detail. On the other hand, trying to explicitly elaborate a strategy for solving the task at hand was particularly helpful for participants who were reflecting after having gained a more substantial level of experience.

CONCLUSIONS

This paper aims to identify the best way to allocate one's time to promote learning. The results of a number of studies carried out across different environments, geographies, and populations helped us identify some of the conditions under which reflecting on accumulated experience generates higher performance outcomes compared to the accumulation of additional experience alone.

Our findings have theoretical, empirical, and practical implications. From a theoretical standpoint, we contribute to the literature on the role of experience as the main driver behind learning curves (e.g., Huckman and Pisano 2003, Argote and Miron-Spektor 2011, Clark et al. 2013, KC et al. 2013) by illuminating some of the conditions under which the marginal returns of practice can be inferior to the marginal returns of reflection. Also, for scholarship suggesting that engaging in reflection after action is an important source of learning (e.g., Zollo and Winter 2002, Ellis and Davidi 2005, Anseel et al. 2009, Schippers et al. 2013), we shed light on some of the conditions under which reflective learning can generate more significant performance improvements. From an empirical standpoint, we combine the fidelity of lab experiments and the real-world representativeness of field studies. In this respect, our paper provides an example of how scholars can strike a balance between internal and external validity by leveraging the complementary strengths of different methods (Di Stefano and Gutierrez 2019) and overcoming some of their intrinsic limitations (Henrich et al. 2010). Finally, our results have practical implications. Organizations that rely on reflective learning tools, such as after-action reviews and post-mortems (Gino and Pisano 2011, Catmull 2014), are the

exception more than the norm. Results from our field study suggest that taking time away from training to reallocate it to reflective learning efforts improved individual performance. Companies may encourage the use of tools such as learning journals to support reflective learning in training and regular operations (Amabile and Kramer 2011a, Ciampa 2017). While our personal experience suggests that not everyone will take these exercises seriously, our findings reveal that they should. With training estimated to be a $165 billion industry in the United States and a $360 billion industry worldwide,[19] we highlight that it may be possible for people to train and learn smarter, not harder. We encourage additional work to clarify how reflective learning can be incorporated more broadly into both training and regular operations.

Our results are subject to several limitations, and the routes to overcome them suggest interesting new avenues for future research. First, despite our best efforts at exploring the boundary conditions of the reflection effect, our findings are still contextual, and further work is needed to generalize them more broadly. Think, for instance, about the differential benefits of articulation and codification highlighted across Studies 3–5. Future work could better unpack the sensitivity of our results to different tasks, manipulations, and experimental protocols. Second, our research focused on individual learning, with participants being removed from social interactions. We believe it would be fascinating to explore how social interaction may aid or detract from the performance outcomes of reflective learning and how the effects we document in this paper may interact with group dynamics, for instance, in the case of formal after-action reviews (Ellis and Davidi 2005, Goh et al. 2013). If we move away from the individual level, we also see value in engaging with the question of how individual learning by thinking affects organizational learning and, potentially, organizational performance. In a diverse array of contexts, competitive advantage critically depends on the skills of individual contributors (Chan et al. 2014; Laureiro-Martinez et al. 2015). We encourage future work to explore this link in more detail. Future research could also extend the study of reflective learning to other outcome variables. For instance, the notion that reflection favors learning may inform research on employee motivation and the role of small wins as one of its drivers (Amabile and Kramer 2011b).

[19] Estimates for 2020. See: https://trainingindustry.com/wiki/outsourcing/size-of-training-industry/

Other limitations of our study are more closely related to the specific deficiencies of our experiments. Study 1 would benefit from replication in a context where experimental assignment can be done at the individual level (as opposed to the trainer level, as it was in our case). Study 4 could be replicated with a protocol where potential mechanisms behind the reflection effect would be manipulated instead of measured (which is why we present our mediation results as descriptive evidence in Appendix 5). The pilot study we presented in the introduction (see also Appendix 1) was not conceived to provide any causal evidence, but we could easily imagine a replication disentangling the effect of choice from the effect of reflection. Similarly, one could explore the extent to which the choice participants made was driven by considerations about expected performance benefits or the perceived costs associated with reflection or even just switching activity. Finally, some of the studies would benefit from being replicated in a physical laboratory with a protocol aimed at ensuring high levels of compliance (see discussion in Appendix 7).

Despite all these limitations, we believe there is value in starting the conversation, and we are hopeful this paper will pave the way for more research on the role of reflection as a mechanism for learning. In our constant battle against the clock, taking time to step back from our daily activities and engage in a deliberate reflection effort may sound like a luxurious pursuit.[20] Reflection does entail high opportunity cost of one's time, yet we argue that taking time to articulate and codify previously accumulated experience is no idle pursuit: It can powerfully enhance the learning process, and do so more than the accumulation of additional experience alone.

[20] Data show that between 1973 and 2000, "the average American worker added an additional 199 hours to his or her annual schedule—or nearly five additional weeks of work per year (assuming a 40-hour workweek)" (Schor 2003, p.7). Meanwhile, between 1969 and 2000, "the overall index of labor productivity per hour increased about 80 percent, from 65.5 to 116.6" (Schor 2003, p.10). As a result, productivity and efficiency have become significant concerns in modern Western societies, with time being "the ultimate scarcity" (e.g., Gross, 1987)—a valuable resource to guard and protect (Gleick 2000; Zauberman and Lynch 2005).

REFERENCES

Amabile T, Kramer S. 2011a. Four reasons to keep a work diary. *Harvard Bus. Rev.* See: https://hbr.org/2011/04/four-reasons-to-keep-a-work-di.

Amabile T, Kramer S. 2011b. *The progress principle: Using small wins to ignite joy, engagement, and creativity at work.* Cambridge, MA: Harvard Business Review Press.

Anseel F, Lievens F, Schollaert E. 2009. Reflection as a strategy for enhancing the effect of feedback on task performance. *Organ. Behav. Human Decision Processes*, 110: 23–35.

Argote L, Epple D. 1990. Learning curves in manufacturing. *Sci.,* 247(4945): 920–924.

Argote L, Ingram P. 2000. Knowledge transfer: A basis for competitive advantage in firms. *Organ. Behav. Human Decision Processes*, 82(1): 150–169.

Argote L, Miron-Spektor E. 2011. Organizational learning: From experience to knowledge. *Organ. Sci.*, 22(5): 1123–37.

Arthur W, Day DV. 1994. Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational Psych. Measurement*, 54(2): 394–403.

Bandura A. 1977. Self-efficacy: Toward a unifying theory of behavior change. *Psych. Rev.*, 84: 191–215.

Bandura A. 1997. *Self-efficacy: The exercise of control.* New York, NY: Freeman.

Bandura A, Schunk DH. 1981. Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *J. Personality and Social Psych.*, 41: 586–598.

Buhrmester M, Kwang T, Gosling SD. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives Psych. Sci.*, 6: 3–5.

Carpenter PA, Just MA, Shell P. 1990. What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psych. Rev.*, 97(3): 404–431.

Cassiman B, Veugelers R. 2006. In search of complementarity in innovation strategy: Internal R&D and external knowledge acquisition. *Management Sci.,* 52(1): 68–82.

Catmull E. 2014. *Creativity, Inc.* New York, NY: Random House.

Chan TY, Jia L., Pierce L. 2014. Learning from peers: Knowledge transfer and sales force productivity growth. *Marketing Sci.*, 33(4): 463–484.

Ciampa D. 2017. The more senior your job title, the more you need to keep a journal. *Harvard Bus. Rev.* See: https://hbr.org/2017/07/the-more-senior-your-job-title-the-more-you-need-to-keep-a-journal

Clark J, Huckman RS, Staats BR. 2013. Customer specificity and learning: Individual and organizational effects in outsourced radiological services. *Organ. Sci.*, 24(5): 1539–1557.

Dewey J. 1933. *How We Think.* Boston, MA: D. C. Heath and Co.

Di Stefano G, Gutierrez C. 2019. Under a magnifying glass: On the use of experiments in strategy research. *Strategic Organ.*, 17(4): 497–507.

Edmondson A, Bohmer R, Pisano G. 2001. Disrupted routines: Team learning and new technology implementation in hospitals. *Admin. Sci. Quart.*, 46(4): 685–716.

Ellis S, Davidi I. 2005. After-event reviews: Drawing lessons from successful and failed experience. *J. Appl. Psych.* 90(5): 857–871.

Ellis S, Carette B, Anseel F, Lievens F. 2014. Systematic reflection: Implications for learning from failures and successes. *Current Directions in Psych. Sci.*, 23: 67–72.

Gino F, Pisano GP. 2011. Why leaders don't learn from success. *Harvard Bus. Rev.* See: https://hbr.org/2011/04/why-leaders-dont-learn-from-success.

Gleick J. 2000. Faster: The acceleration of just about everything. Vintage Press, New York, NY.

Goh KT, Goodman PS, Weingart LR. 2013. Team innovation processes: An examination of activity cycles in creative project teams. *Small Group Res.*, 44: 159–194.

Gross BL. 1987. Time scarcity: Interdisciplinary perspective and implications for consumer behavior. *Res. Cons. Beh.*, 2.

Henrich J, Heine SJ, Norenzayan A. 2010. The weirdest people in the world? *Behav. Brain Sci.*, 33(2–3): 61–83.

Huckman R, Pisano GP. 2003. The firm specificity of individual performance: Evidence from cardiac surgery. *Management Sci.*, 52(2): 473–488.

Immordino-Yang MH, Christodoulou JA, Singh V. 2012. Rest is not idleness: Implications of the brain's default mode for human development and education. *Perspectives on Psych. Sci.*, 7(4): 352–365.

Kale P, Singh H. 2007. Building firm capabilities through learning: The role of the alliance learning process in alliance capability and firm-level alliance success. *Strategic Management J.*, 28: 981–1000.
KC D, Staats BR, Gino F. 2013. Learning from my successes and others' failure: Evidence from minimally invasive cardiac surgery. *Management Sci.*, 59(11): 2435–2449.
Laureiro-Martinez D, Brusoni S, Canessa N, Zollo M. 2015. Understanding the exploration-exploitation dilemma: An fMRI study of attention control and decision-making performance. *Strategic Management J.*, 36: 319–338.
Levitt B, March J. 1988. Organizational learning. *Ann. Rev. Soc.,* 14: 319–340.
March JG. 2010. *The Ambiguities of Experience.* Cornell University Press, Ithaca, NY.
Mayer RE. 1996. The search for insight: Grappling with Gestalt psychology's unanswered questions. R. J. Sternberg, J. E. Davidson, eds. *The Nature of Insight*. MIT Press, Cambridge, MA.
Mazar N, Amir O, Ariely D. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *J. Marketing Res.*, 45: 633–644.
Multon KD, Brown SD, Lent RW. 1991. Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *J. Counseling Psych.*, 18: 30–38.
Nyberg L, Eriksson J, Larsson A, Marklund P. 2006. Learning by doing versus learning by thinking: an fMRI study of motor and mental training. *Neuropsych.*, 44: 711–717.
Olsson CJ, Jonsson B, Nyberg L. 2008. Learning by doing and learning by thinking: An fMRI study of combining motor and mental training. *Frontiers Hum. Neurosci.*, 2: 5.
Palan S, Schitter C. 2018. Prolific.ac—A subject pool for online experiments. *J. Behav. Exp. Finance* 17: 22–27.
Peer E, Brandimarte L, Samat S, Acquisti A. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *J. Exp. Soc. Psychol.* 70: 153–163.
Pierce L, Snyder JA. 2020. Historical origins of firm ownership structure: The persistent effects of the African slave trade. *Acad. Management J.,* 63(6): 1687–1713.
Raven J. 1989. The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *J. Educational Measurement*, 21(1): 1–16.
Raven J. 2000. Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psych.*, 41, 1–48.
Schilling M, Vidal P, Ployhart RE, Marangoni A. 2003. Learning by doing something else: Variation, relatedness, and the learning curve. *Management Sci.*, 49:1, 39–56.
Schippers MC, Homan AC, van Knippenberg D. 2013. To reflect or not to reflect: Prior team performance as a boundary condition of the effects of reflexivity on learning and final team performance. *J. Organ. Behav.*: 6–23.
Schippers MC, Edmondson AE, West MA. 2014. Team reflexivity as an antidote to information processing failures. *Small Group Res.*, 5:731–769.
Schor J. 2003. The (even more) overworked American. In J. de Graaf (ed.), *Take back your time: Fighting overwork and time poverty in America*. Berrett-Koehler Publishers, San Francisco, CA; 6-11.
Schunk DH, Hanson AR. 1985. Peer models: Influence on children's self-efficacy and achievement behaviors. *J. Educational Psych.*, 77: 313–322.
Schunk DH, Hanson AR, Cox PD. 1987. Peer model attributes and children's achievement behaviors. *J. Educational Psych.*, 79: 54–61.
Seibert KW. 1999. Reflection-in-action: Tools for cultivating on-the-job learning conditions. *Organ. Dynamics*, 27: 54–65.
Spencer SJ, Zhanna MP, Fong GT. 2005.Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *J Pers. Soc. Psychol.* 89(6): 845–851.
Taylor R. 2019. *The Psychology of Pandemics* (Cambridge Scholars Publishing, Cambridge, UK).
Wulf G, Schmidt RA. 1997. Variability of practice and implicit motor learning. *J. Experiment. Psych.: Learn. Memory, and Cognition* 23: 987–1006.
Wytek R, Opgenoorth E, Presslich O. 1984. Development of a new shortened version of Raven's Matrices Test for application rough assessment of present intellectual capacity within psychopathological investigation. *Psychopathology*, 17(2): 49–58.
Zauberman G, Lynch JG. 2005. Resource slack and discounting of future time versus money. *J. Experimental Psych.: General*, 134(1): 23–37.
Zollo M, Winter SG. 2002. Deliberate learning and the evolution of dynamic capabilities. *Organ. Sci.*, 13(3): 339–351.

TABLE 1.
Overview of Empirical Evidence

	Question	Motivation	Findings	Conclusion
		Motivating our study		
Pilot study (N=256, Amazon Mechanical Turk) – *See Appendix 1*	Are the performance benefits of reflection vs. practice obvious?	Examining people's belief regarding the benefits of reflection vs. practice.	82% of our participants chose practice over reflection.	We interpret these results as suggesting that the existence of a reflection effect is not obvious and that it is most people's intuition that gaining additional experience leads to superior performance improvements compared to reflecting on previously accumulated experience.
		Providing support to our hypothesis		
Study 1 (N=101, Wipro BPO)	Is there any evidence of the performance benefits of reflection vs. practice in the wild?	Providing preliminary field evidence in support of the existence of a reflection effect.	Agents who were given time to reflect on their experience outperformed those who were given the same amount of time to practice in a final assessment test administered by the firm.	We interpret these results as suggestive of a "reflection effect." However, limitations inherent to our empirical design (treatment administered at the trainer level) suggest caution in interpreting these results as a proper test of our hypothesis.
Study 2 (N=468, Amazon Mechanical Turk)	Are the performance benefits of reflection vs. practice causal?	Providing causal evidence of the marginal benefits associated with reflection vs. practice vs. control (watch cooking video).	Participants assigned to the reflection condition improved their performance significantly more than those who spent the same amount of time either practicing or taking a break from the task.	We interpret these results as providing causal support in favor of our hypothesis: Reflecting on previously accumulated experience generates higher performance outcomes compared to the accumulation of additional experience alone.
Additional Study #1 (N=109, US university) – *See Appendix 2*	Are the performance benefits of reflection task specific?	We test the effects of reflection vs. control using an in-person, involved task (karaoke singing in front of an experimenter).	Participants assigned to the reflection condition improved their singing accuracy compared to those who spent the same amount of time taking a break from the task.	Results are consistent with Study 2.
Additional Study #2 (N=453, Amazon Mechanical Turk) – *See Appendix 2*	Are the performance benefits of reflection task specific?	We test the effects of reflection vs. practice vs. control using a non-mathematical task (tumor cell count).	Participants assigned to the reflection condition improved their counting accuracy compared to those who spent the same amount of time either practicing or taking a break from the task.	Results are consistent with Study 2.

		Exploring boundary conditions to the reflection effect		
Study 3 (N=290, Prolific)	Do the performance benefits of reflection vs. practice depend on how individuals reflect?	Investigating whether one can make reflection more effective by acting on how individuals reflect. We separate reflection into articulation of accumulated experience and codification of guidelines to better execute the task at hand.	Participants in the articulation condition performed better than those in both the practice and the codification conditions. Participants in the codification condition did not seem to experience any substantial advantage compared to those in the practice condition.	We interpret these results as suggestive of the fact that, when one has limited time to allocate to reflection, it is more effective to focus one's reflection efforts on the articulation of accumulated experience than to try and abstract more general guidelines on how to better execute the task at hand.
Study 4 (N=560, Prolific)	Do the performance benefits of reflection vs. practice depend on what individuals reflect on?	Investigating whether the benefits of reflection can spill over to related tasks. We look at the effect of articulation and codification, as per Study 3, on the same task vs. a related task.	Participants performed better when they were asked to solve the same type of matrices in the second round, independent of whether they were asked to practice or reflect. When participants were asked to solve the related type of matrices, however, participants assigned to reflection were put at a significant advantage compared to participants assigned to practice.	We interpret these results as suggestive of the fact that reflection has the potential to generate spillover effects to related tasks.
Study 5 (N=1,203, Prolific)	Do the performance benefits of reflection vs. practice depend on when individuals reflect?	Investigating whether the benefits of reflection change depending on the amount of experience on which one can reflect. We look at the effect of articulation and codification and manipulate the amount of experience we let participants accumulate before engaging in practice vs. reflection.	Participants in the practice condition outperformed those who engaged in reflection after having accumulated little experience. The opposite held true if participants reflected after they had gained a more substantial level of experience. We remarked no significant differences once participants had accumulated a substantial amount of experience.	We interpret these results as suggestive of the fact that reflection is mostly beneficial at the beginning of the learning curve, as long as one has accumulated a sufficient amount of experience on which to reflect.
Additional Study #3 (N=599, Prolific) and Additional Study #4 (N=301, Prolific) – *See Appendix 7*	Do the performance benefits of reflection vs. practice change over time?	Investigating whether the benefits of reflection change as one engages in reflection repeatedly. We look at the effect of articulation and codification, and have participants engage in three subsequent rounds of reflection. We ran two versions of the study with the difficulty of the task increasing across rounds (Additional Study #3) or only at the end (Additional Study #4).	Participants in the codification condition outperformed those in the practice condition. The moment at which this happened, however, changed across the two studies. The benefits associated with reflection materialized earlier in Additional Study #3, taking more time to materialize for Additional Study #4.	We interpret these results as suggestive of the fact that one cannot extract identical benefits from reflection (or practice) over time when engaging repeatedly with the same task. Interestingly, they seem to suggest the existence of a correlation between the ability to benefit from reflection and the extent to which individuals are challenged by the task at hand.

Note: The table provides an overview of the empirical evidence we gather in support of our findings, including four additional studies featured in Appendices 1, 2, and 7.

TABLE 2.
Overview of Design Features

	Environment	Participants	Conditions	Task	Process	Incentive	Performance
Study 1	Field study, Wipro BPO (India)	101 agents joining one customer account	1. Reflection 2. Practice	Training program	1. 5 days of regular training 2. 10 days of training with treatment	None	1. Score on Final Assessment test, 0-100 2. Score on Customer Satisfaction measure "top box," 0-1
Study 2	Amazon Mechanical Turk	468 adults	1. Reflection 2. Practice 3. Control	Math puzzles	1. 5 puzzles 2. Treatment 3. 5 puzzles 4. 5 puzzles	Flat pay of $1 + performance-based bonus of $1 for 10% of participants	Number of correctly solved puzzles after treatment, 0-10
Study 3	Prolific	290 adults	1. Articulation 2. Codification 3. Practice	Math puzzles	5. 5 puzzles 6. Treatment 7. 5 puzzles 8. 5 puzzles	Flat pay of £2.50 (20 minutes at £7.50/hour) + performance-based bonus of £1 for 10% of participants	Number of correctly solved puzzles after treatment, 0-10
Study 4	Prolific	560 adults	1. Articulation 2. Codification 3. Practice	Raven matrices	1. 5 matrices 2. Treatment 3. 5 matrices (either same or related type)	Flat pay of £2.63 (21 minutes at £7.50/hour) + performance-based bonus of £1 for 10% of participants	Number of correctly solved matrices after treatment, 0-5
Study 5	Prolific	1,203 adults	1. Articulation 2. Codification 3. Practice	Raven matrices	1. 2 to 8 matrices (four levels of experience) 2. Treatment 3. 4 matrices	Flat pay of £2.55 (17 minutes at £9.00/hour) + performance-based bonus of £1.15 for 10% of participants	Number of correctly solved matrices after treatment, 0-4

Note: The table provides an overview of the main design features of the five experimental studies featured in the paper.

TABLE 3.
Overview of Reflection Manipulations

Study	Manipulation
Study 1	Please take the next 15 minutes to **reflect on the training day** you just completed. **Please write about the main lessons you learned** as you were completing your training. Please reflect on and write about at least two key lessons. Please be as specific as possible.
Study 2	Please take the next few minutes to **reflect on the task** you just completed. **Please write about what strategies, if any, you used** as you were working on the task. **Also, please write about what you think one can do to be effective** in solving the math problems included in this task. Please be as specific as possible. You will have 3 minutes to engage in this reflection.
Study 3/4	Articulation: Please take the next few minutes to **reflect on the task** you just completed. **Please reflect about what you think was required of you** in solving the puzzles(/matrices) **and your performance** in this round. Please be as specific as possible. You will have 3 minutes to engage in this reflection. Articulation + Codification: Please take the next few minutes to reflect on the task you just completed. Please reflect about what you think was required of you in solving the puzzles (/matrices) and your performance in this round. You will have 1 minute to engage in this reflection. Once the minute is over, you will be asked to **write down what you think one can do to be effective** in solving the puzzles (/matrices) included in the task. Please be as specific as possible. You will have 2 minutes to write down your thoughts.
Study 5	Articulation: Please take the next few minutes to **reflect on the task** you just completed. **Please think about what strategies, if any, you used** as you were working on the task. Please be as specific as possible. You will have 3 minutes to engage in this reflection. Articulation + Codification: Please take the next few minutes to reflect on the task you just completed. **Please think about what strategies, if any, you used** as you were working on the task. **Also, please write about what you think one can do to be effective** in solving the matrices included in this task. Please be as specific as possible. You will have 3 minutes to engage in this reflection.

Note: The table provides a summary of the different reflection manipulations used across studies. We have highlighted parts of the manipulations that may be interesting to contrast across studies. We comment on such differences in the text when introducing each manipulation.

TABLE 4.
Univariate Tests across Conditions, Study 1

	Treatment: Reflection (N=56)		Control: Practice (N=45)		T-test (Two-tailed)		Cohen's d
	Mean	S.D.	Mean	S.D.	T	P-Value	D
Control Variables							
Participant: Age	24.768	3.618	25.787	3.839	-1.385	0.169	0.273
Participant: Gender	0.732	0.447	0.872	0.337	-1.769	0.080	0.354
Participant: Prior experience	31.245	33.549	26.930	27.786	0.702	0.484	0.140
Trainer: Age	33.679	6.764	30.936	2.839	2.593	0.011	0.529
Trainer: Gender	0.821	0.386	1.000	0.000	-3.165	0.002	0.659
Trainer: Prior experience	63.286	1.546	94.936	27.335	-86.562	0.000	1.635
Dependent Variables							
Final Assessment	71.536	9.785	54.422	20.715	5.474	0.000	1.056
Customer Satisfaction	0.912	0.236	0.765	0.285	2.051	0.045	0.562

Note: In the case of Customer Satisfaction, N = 23 for treatment and N = 36 for control.

TABLE 5.
The Effect of Reflection on Final Assessment and Customer Satisfaction, Study 1

	Final Assessment			**Customer Satisfaction**		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Reflection	13.976	3.030	0.000	0.286	0.251	0.259
Participant: Age	-0.855	0.507	0.095	-0.006	0.010	0.546
Participant: Male	-2.328	2.061	0.262	0.083	0.109	0.451
Participant: Experience	0.164	0.061	0.009	0.002	0.001	0.140
Trainer: Age	-0.260	0.203	0.203	0.007	0.014	0.625
Trainer: Male	21.770	4.464	0.000	-0.078	0.189	0.683
Trainer: Experience	-0.193	0.082	0.021	0.005	0.004	0.195
Cons	78.399	14.257	0.000	0.175	0.444	0.695
N		101			59	
F	11.170		0.000	2.180		0.052
Adjusted R2		0.431			0.217	

Note: We investigated differences in final performance by running an OLS regression with robust standard errors and controls at the trainer level. Final Assessment (a score ranging from 0 to 100) is our dependent variable for Model 1, while Customer Satisfaction (percentage of top satisfied customers) is our dependent variable for Model 2. We provide a thorough discussion of our choice not to cluster the standard errors in Appendix 3.

TABLE 6.

The Effects of Articulation and Codification on Final Performance for Same and Related Tasks, Study 4

	Model 1			Model 2			Model 3			Model 4		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	0.268	0.118	0.024	0.011	0.138	0.939	0.011	0.138	0.937	0.541	0.194	0.006
Codification	0.230	0.122	0.059	0.022	0.142	0.879	0.019	0.142	0.895	0.448	0.200	0.026
Related task	-0.724	0.096	0.000	-1.044	0.183	0.000						
…X Articulation				0.498	0.052	0.026						
…X Codification				0.534	0.238	0.081						
Initial Performance	0.502	0.052	0.000	0.430	0.246	0.000	0.471	0.074	0.000	0.527	0.071	0.000
_cons	1.845	0.221	0.000	2.013	0.233	0.000	2.114	0.317	0.000	0.864	0.293	0.003
N	560			560			287			273		
F	39.766		0.000	27.726		0.000	13.885		0.000	23.399		0.000
Adjusted R^2	0.224			0.229			0.179			0.163		

	Model 1			Model 2			Model 3			Model 4		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	0.268	0.118	0.024	0.011	0.138	0.939	0.011	0.138	0.937	0.541	0.194	0.006
Codification	0.230	0.122	0.059	0.022	0.142	0.879	0.019	0.142	0.895	0.448	0.200	0.026
Related task	-0.724	0.096	0.000	-1.044	0.183	0.000						
…X Articulation				0.498	0.052	0.026						
…X Codification				0.534	0.238	0.081						
Initial Performance	0.502	0.052	0.000	0.430	0.246	0.000	0.471	0.074	0.000	0.527	0.071	0.000
_cons	1.845	0.221	0.000	2.013	0.233	0.000	2.114	0.317	0.000	0.864	0.293	0.003
N	560			560			287			273		
F	39.766		0.000	27.726		0.000	13.885		0.000	23.399		0.000
Adjusted R^2	0.224			0.229			0.179			0.163		

Note: We investigated differences in *final performance* while controlling for *initial performance* by running an OLS regression with robust standard errors. We started by estimating the effects of *articulation* and *codification* vis-à-vis *practice*, as well as the effect of *same* vis-à-vis *related task* (Model 1). We next explored the interaction between the type of task assigned with *articulation* and *codification* (Model 2). To visualize differences across tasks more intuitively, we then ran a split sample analysis where we separately looked at the effects of *articulation* and *codification* vis-à-vis *practice* for participants who completed the *same* vis-à-vis *related task* (Model 3 and Model 4, respectively).

TABLE 7.

The Effects of Articulation and Codification on Final Performance for Different Levels of Familiarity, Study 5

	Model 1			Model 2			Model 3			Model 4		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	-0.125	0.082	0.126	-0.282	0.201	0.161	-0.360	0.161	0.025	10.076	3.885	0.010
… X Familiarity				0.063	0.072	0.386						
… X Familiarity group 2							0.444	0.240	0.065	-9.576	6.284	0.128
… X Familiarity group 3							0.235	0.222	0.290	-9.168	5.936	0.123
… X Familiarity group 4							0.269	0.227	0.235	-5.860	5.989	0.328
Codification	-0.036	0.082	0.662	-0.107	0.206	0.604	-0.290	0.171	0.090	12.109	4.727	0.011
… X Familiarity				0.028	0.074	0.700						
… X Familiarity group 2							0.577	0.236	0.015	-6.194	5.369	0.249
… X Familiarity group 3							0.265	0.234	0.257	-5.463	5.514	0.322
… X Familiarity group 4							0.190	0.233	0.415	-7.440	5.510	0.177
Familiarity	0.173	0.029	0.000	0.143	0.054	0.008						
Familiarity group 2							-0.251	0.177	0.158	6.281	3.943	0.111
Familiarity group 3							0.041	0.164	0.804	-1.053	3.874	0.786
Familiarity group 4							0.388	0.170	0.023	-8.701	3.912	0.026
Initial Performance	2.055	0.127	0.000	2.056	0.127	0.000	2.069	0.127	0.000	2.278	0.124	0.000
_cons	0.180	0.147	0.220	0.255	0.188	0.175	0.554	0.163	0.001	58.091	3.959	0.000
N		1,203			1,203			1,203			1,203	
F		73.981	0.000		50.119	0.000		25.897	0.000		33.307	0.000
Adjusted R^2		0.182			0.181			0.184			0.269	

Note: We investigated differences in *final performance* while controlling for *initial performance* by running an OLS regression with robust standard errors. We started by estimating the effects of *articulation* and *codification* vis-à-vis *practice*, as well as the effect of *familiarity* (Model 1). We next included the interaction terms of *familiarity* and *articulation* as well as the interaction terms of *familiarity* and *codification* (Model 2). We further unpacked such interactions to observe differences across all the different levels of *familiarity* (Model 3). Finally, we replicated the same analysis using the time taken to solve the matrices in the final round as dependent variable (Model 4).

TABLE 8.

Characteristics of Reflection Logs across and within Different Levels of Familiarity, Study 5

	Overall	Familiarity Level 1	Familiarity Level 2	Familiarity Level 3	Familiarity Level 4
Length (words)	55	52	60	53	57
Content: Task	5.33%	3.88%	5.50%	4.46%	7.58%
Content: Strategy	84.86%	83.50%	86.00%	85.15%	84.85%
Content: Performance	10.79%	9.22%	12.50%	10.89%	10.61%

Note: We examined the average length of reflection logs, as well as the extent to which participants focused their logs on a discussion of the task at hand, the strategy used to solve it, and their performance with it. We first report these figures for all participants assigned to the codification condition across familiarity level, and then for the sub-samples of participants assigned to each of the four familiarity levels.

TABLE 9.

Characteristics of Reflection Logs and Final Performance for Different Levels of Familiarity, Study 5

	Overall			Familiarity Level 1			Familiarity Level 2			Familiarity Level 3			Familiarity Level 4		
	Coef	SE	p-value	Coef	SE	p-value	Coef	SE	p-value	Coef	SE	p-value	Coef	SE	p-value
Length (words)	0.007	0.002	0.000	0.014	0.004	0.000	0.002	0.003	0.478	0.008	0.005	0.084	0.005	0.005	0.304
Content: Task	0.002	0.349	0.995	-0.912	0.625	0.148	0.577	0.446	0.199	-0.906	0.655	0.170	0.621	0.556	0.267
Content: Strategy	0.414	0.198	0.037	0.086	0.338	0.799	0.756	0.344	0.030	0.240	0.525	0.650	0.337	0.378	0.375
Content: Performance	0.069	0.212	0.744	0.250	0.451	0.580	-0.098	0.344	0.777	-0.267	0.439	0.546	0.540	0.357	0.133
Initial Performance	1.821	0.217	0.000	1.410	0.455	0.003	1.862	0.351	0.000	2.106	0.443	0.000	2.130	0.486	0.000
_cons	-0.009	0.203	0.963	-0.011	0.404	0.979	-0.046	0.363	0.900	-0.037	0.516	0.943	0.113	0.424	0.790
N	403			103			100			101			99		
F	29.515		0.000	9.635		0.000	14.294		0.000	10.002		0.000	7.583		0.000
Adjusted R^2	0.216			0.237			0.270			0.214			0.178		

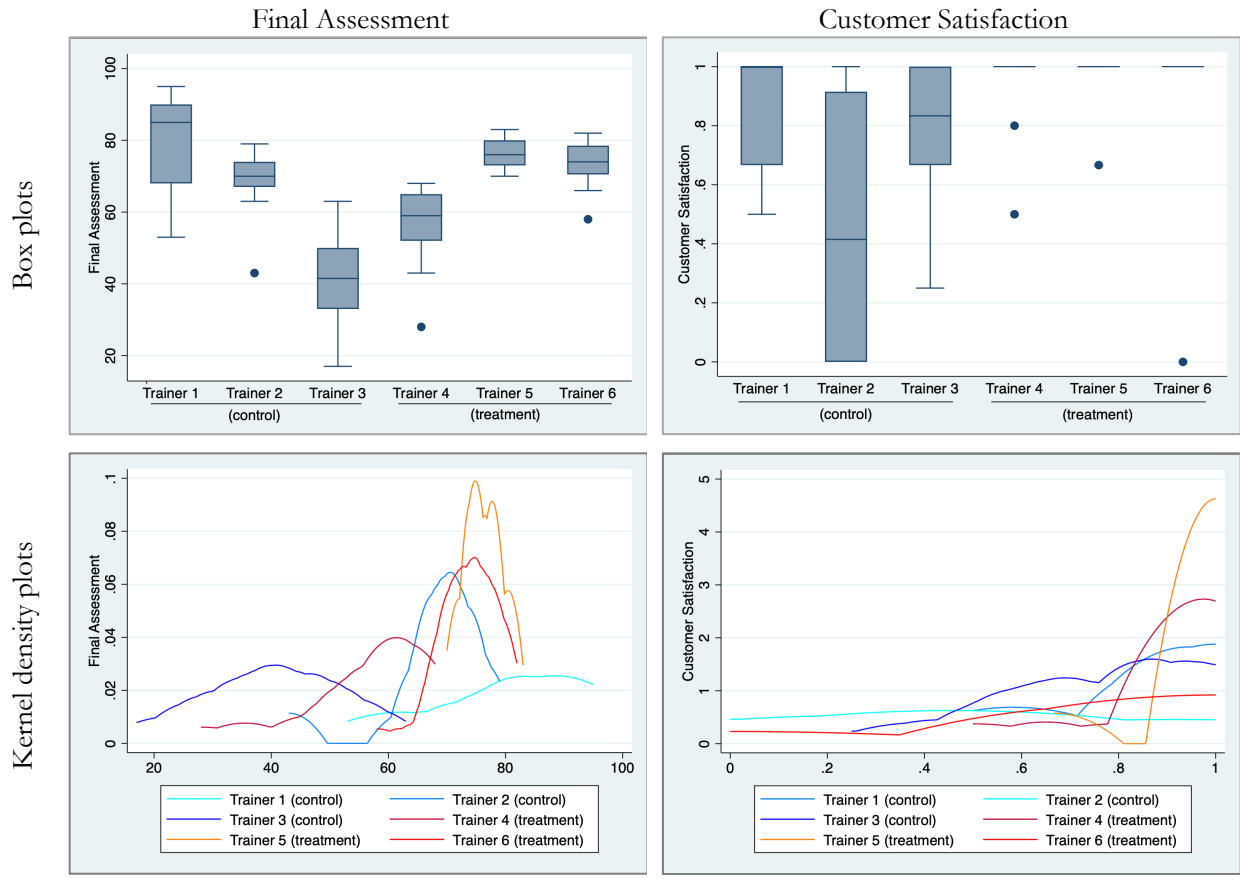
Note: We investigated differences in *final performance* while controlling for *initial performance* by running an OLS regression with robust standard errors. We first looked at all participants assigned to the codification condition across familiarity level, and then re-ran the analysis on the sub-samples of participants assigned to each of the four familiarity levels.

FIGURE 1.
Example of Math Puzzle Participants Were Asked to Solve

8.18	9.01	3.97
5.2	4.56	9.12
0.28	2.92	6.59
1.12	6.93	9.72

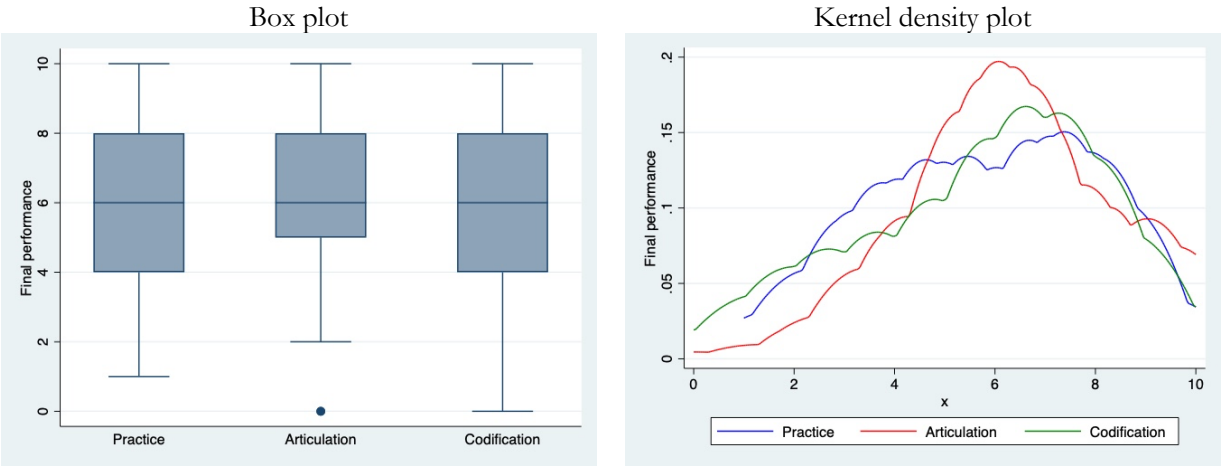
Note: The experimental task consisted of a series of sum-to-ten games, a math puzzle in which two out of twelve numbers sum to ten (see Mazar et al. 2008).

FIGURE 2.
Performance by Condition and Trainer, Study 1



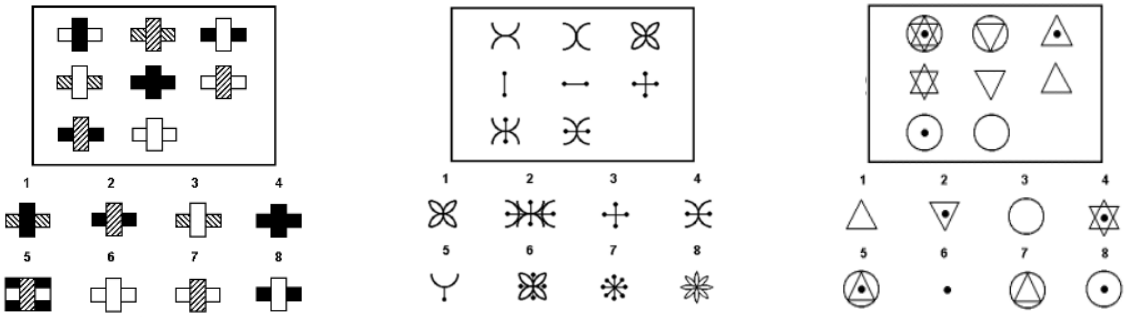
Note: Panels in the figure show box plots (top) or kernel density plots (bottom) for performance measured as Final Assessment (left) or Customer Satisfaction (right). Both types of graph help visualize the distribution of performance outcomes for participants by condition and trainer. The box plots do so by displaying the interquartile range (represented by the box itself, the median value marked with a line) as well as the minimum and maximum values (shown by the whiskers). Dots represent outliers. The Kernel density plots on the other hand provide a visualization that is more similar to a histogram, only with kernel smoothing to plot the values.

FIGURE 3.
Performance by Condition, Study 3



Note: Panels in the figure show box plots (left) and kernel density plots (right) for performance by condition (practice, articulation, and codification). Both types of graph help visualize the distribution of performance outcomes for participants by condition. The box plots do so by displaying the interquartile range (represented by the box itself, the median value marked with a line) as well as the minimum and maximum values (shown by the whiskers). Dots represent outliers. The Kernel density plots on the other hand provide a visualization that is more similar a histogram, only with kernel smoothing to plot the values.

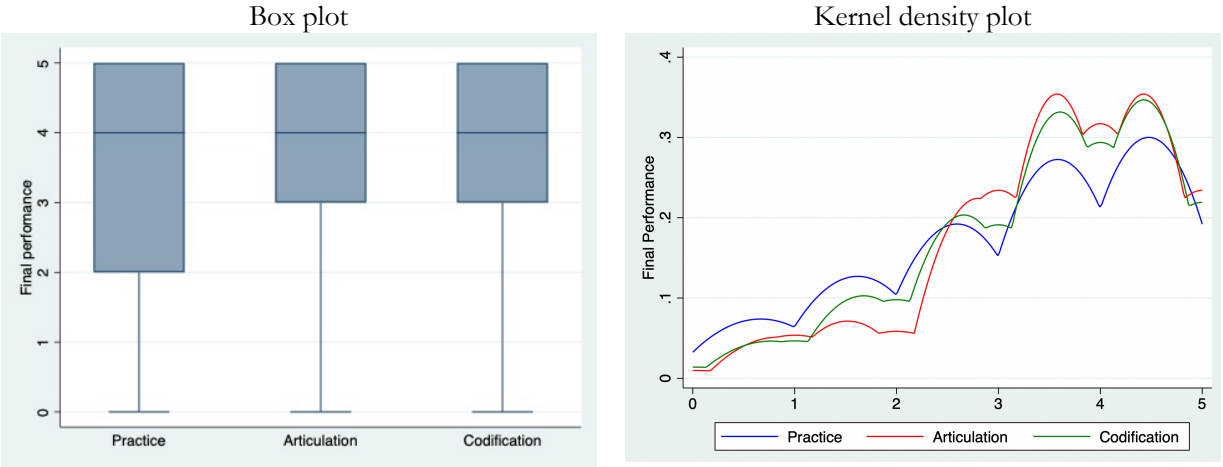
FIGURE 4.
Example of Raven Matrices Participants Were Asked to Solve, Study 4



(a) (b) (c)

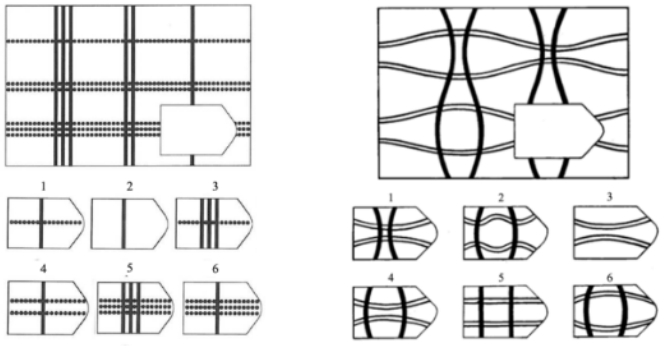
Note: Matrix (a) is an example of Raven matrix of the "distribution of three values" type (Carpenter et al. 1990) received by all participants in the practice round and by participants in the same-task condition in the second round. During the second round, participants in the related-task condition received five matrices of the "figure addition and subtraction" type (Carpenter et al. 1990). Matrix (b) is an example of figure addition, while Matrix (c) is an example of figure subtraction.

FIGURE 5.
Performance by Condition, Study 4



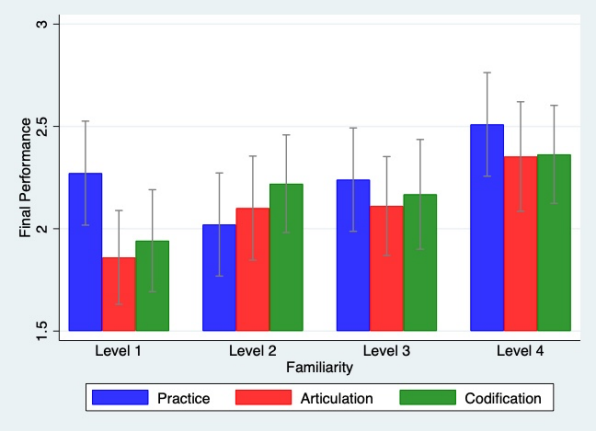
Note: Panels in the figure show box plots (left) and kernel density plots (right) for performance by condition (practice, articulation, and codification). Both types of graph help visualize the distribution of performance outcomes for participants by condition. The box plots do so by displaying the interquartile range (represented by the box itself, the median value marked with a line) as well as the minimum and maximum values (shown by the whiskers). Dots represent outliers. The Kernel density plots, on the other hand, provide a visualization that is more similar to a histogram, only with kernel smoothing to plot the values.

FIGURE 6.
Example of Filler Tasks Participants were Asked to Solve, Study 5



Note: To compensate for the higher fatigue associated with solving more matrices, participants assigned to the first three groups (i.e., participants who had to complete two, four, or six matrices) were assigned, respectively, six, four, or two filler tasks to solve before they got started on the matrices. Filler tasks were simple puzzles, where participants were required to find the piece that fitted with the graphical pattern. We selected this task so that it looked similar enough to the main task of the study but required a different type of reasoning (matching of a graphical patterns vs. understanding of a logical sequence).

FIGURE 7.
Performance by Familiarity and Condition, Study 5



Note: The panel in the figure shows the mean and confidence intervals for the second-round scores of the 12 experimental groups, thus providing a graphical representation of the performance of participants by condition (articulation vs. codification vs. practice) across different familiarity levels (i.e., whether participants practiced on two, four, six, or eight matrices during the practice round).

Online Appendix for
"Learning by thinking: How reflection can spur progress along the learning curve"
Giada Di Stefano, Francesca Gino, Gary Pisano, Bradley Staats

List of contents

Appendix 1
Are the benefits of reflection obvious?

In this pilot study, we examined people's beliefs regarding the benefits of reflection versus practice. We gave participants a choice regarding how to allocate their time after gaining some experience on a task and before working on it further. Participants could choose between (1) gaining additional experience on the task in question and (2) articulating and codifying the experience they had accumulated so far. The purpose of this study was not to test for any causal link between reflection and performance. Rather, we were interested in understanding whether the effects of reflection are obvious to the majority of individuals.

Design and procedure. We recruited 256 adults (56.3% male, M_{age} = 31.67, SD = 8.46) on Amazon Mechanical Turk to participate in an online study in exchange for $1 and the potential to earn an additional bonus based on performance ($1 to 10% of participants). After receiving welcoming instructions, participants were asked two questions used as attention checks. Participants who failed one of the attention checks were redirected to a page telling them they could not participate in the study. Participants who correctly answered both attention checks moved on to a screen with instructions. The experimental task consisted of a series of sum-to-ten games, a math puzzle in which two out of 12 numbers sum to ten, as per Figure 1 in the paper (see Mazar et al. 2008).

We gave participants 20 seconds per puzzle and let them first complete a practice round to gain familiarity with the task. Participants were then asked to complete a first round of five different puzzles. After completing each puzzle, they were told whether the answers they selected were correct or not. After this first round, participants received the following instructions:

> *We'll soon be asking you to engage in a second round of the math brain teaser (i.e., five more math puzzles). Before you start round 2, you can choose how to spend the next 3 minutes. You have two choices:*
> *1) You can spend 3 minutes thinking and writing about strategies you used in the first round.*
> *2) You can spend 3 minutes practicing on another set of math puzzles (the same type of math puzzles as the ones you solved in the first round).*
> *Please choose how you want to spend the next 3 minutes to best prepare for round 2 of the math brain teaser.*

Depending on the choice they then made, participants were redirected to one of two different screens. Participants who had chosen to reflect received the following instructions:

> *Please take the next few minutes to reflect on the task you just completed. Please write about what strategies, if any, you used as you were working on the task. Also, please write about what you think one can do to be effective in solving the math problems included in this task. Please be as specific as possible. You will have 3 minutes to engage in this reflection. The study will advance to the next stage once the 3 minutes are over.*

Participants who had chosen to practice received the following instructions:

> *Please take the next few minutes to practice some more on the task you just completed. Below you'll see a few puzzles that you can try to solve. (You can keep track of your performance on a piece of paper if you'd like.) You will have 3 minutes to practice on the puzzles. The study will advance to the next stage once the 3 minutes are over.*

After three minutes spent on either of the two conditions (reflection vs. practice), participants completed two additional rounds of five puzzles each and then answered a few demographic questions.

Results. Eighty-two percent of participants (210 out of 256) chose practice, while the remaining 18 percent (46 out of 256) chose to reflect on the experience they accumulated in the first round, $X^2(1) = 105.6$, $p < .001$. Despite the overwhelming preference for practice, reflection resulted in higher levels of performance over rounds 2 and 3 of the brain teaser. To show this, we ran an ANOVA using participants' performance on the second and third rounds of the brain teaser as the dependent variable and choice (i.e., whether people decided to engage in reflection or practice) as the independent variable, controlling for performance on the first round (before the choice occurred). We found a significant effect of our manipulation on performance in the second and third rounds, $F(1, 253) = 5.24$, $p = .023$, $\eta^2_p = .02$. Participants correctly solved more puzzles in the second and third rounds when they decided to reflect on the experience they accumulated in the first round (M = 5.33,

SD = 2.31) rather than gain additional experience (*M* = 4.37, *SD* = 2.39). As one might expect, performance in round 1 predicted performance over rounds 2 and 3, $F(1, 253) = 56.39$, $p < .001$, $\eta^2_p = .18$. We also conducted independent sample t-tests and found no significant performance difference in round 1 between those who later chose reflection versus practice ($p = .37$).

Discussion. We ran this study to determine whether the "superiority" of reflective learning is intuitive. Our results show that the overwhelming majority of participants decided to gain additional experience rather than take time to reflect on what they had learned from prior experience. Their preference for "experiential learning" over "reflective learning" was presumably based on the premise that gaining additional experience would lead to a superior performance improvement compared to engaging in a deliberate articulation and codification effort.[21] In other words, individuals chose to practice because they expected it would enable them to perform better in subsequent rounds. However, results from the other studies we carried out for this paper (where participants were randomly assigned to our experimental conditions) show just the opposite: Participants who were asked to reflect on their past experience consistently outperformed those who opted for additional practice.

[21] As noted in the manuscript, we base our interpretation on the fact that participants were incentivized with a performance-based bonus, which plausibly led them to choose the strategy that could have helped them maximize their compensation. We invite future work to unequivocally disentangle the extent to which the choice participants made was driven by considerations about expected performance benefits or the perceived costs associated with reflection or even just switching activity.

Appendix 2.
Are the benefits of reflection task-dependent?

We report results from two additional studies exploring the robustness of the reflection effect uncovered with Study 2 to different types of tasks. In Additional Study #1, we used an in-person, involved task (Karaoke singing) and compared the effect of reflection versus control. In Additional Study #2, we used a non-mathematical task (tumor cell count) and compared the effect of reflection versus practice versus control. Compared to Study 2, both studies were incentivized with a flat fee only.

Additional Study #1: Reflection vs. control in karaoke (N = 109). We recruited 109 students (48% male, M_{age} = 20.72, SD = 2.55) from local universities in a city in the Southeastern United States. Participants were asked to perform karaoke songs in exchange for a participation fee of $20. More specifically, we asked them to sing karaoke on a Nintendo Wii video game console in front of an experimenter using the "Karaoke Revolution: Glee" program. Participants completed the study one at a time for a total of 30 to 45 minutes. At the beginning of the study, the experimenter told participants they would be singing songs on a karaoke program. We selected the following three songs: (1) "Haven't Met You Yet," by Michael Bublé; (2) "Don't You Forget about Me," by Simple Minds; and (3) "I Will Survive," by Gloria Gaynor. Next, a second experimenter accompanied the participant into a different room where a Nintendo Wii was set up with a microphone and a television screen, as per Figure A1.

Figure A1. Experimental Setting

To eliminate potential demand effects, both experimenters were blind to the experimental condition and hypotheses. The experimenter handed the microphone to the participant and said: "You will sing into this microphone. The lyrics will appear across the bottom of the screen. Make sure that you keep the microphone close to your mouth so that the software can record your performance." The participant sang the first song while the experimenter sat in front of them, watching. Upon completion of this task, participants were administered our treatment and then asked to sing the other two songs. We concluded by asking a few demographic questions.

Participants in the control group were simply instructed to wait for five minutes for the next round to start. Participants in the reflection group were given the following instructions:

> *Please take the next five minutes to reflect on the task you just completed. Please write about what strategies, if any, you used as you were working on the task. Also, please write about what you think one can do to be effective in completing this task. Please be as specific as possible. When done, click on "Next."*

At the end of each song, the karaoke program's voice recognition software provided an objective performance score (0–100) measuring singing accuracy. The score was computed as an equally weighted average

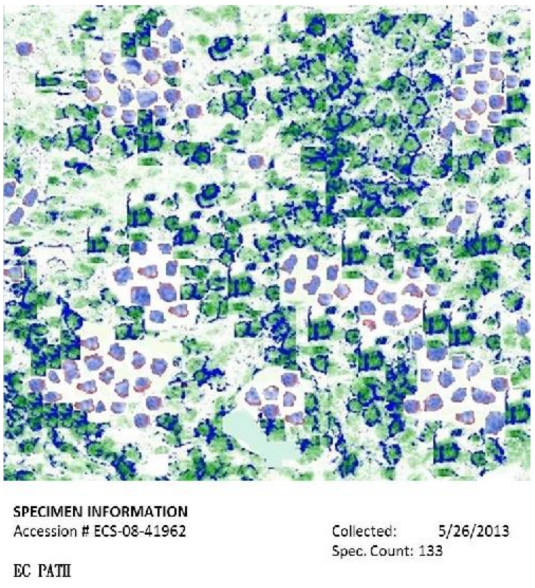
of the software's measurement of volume (quiet-loud), pitch (distance from true pitch), and note duration (accuracy of breaks between notes).

To estimate our results, we conducted an ANOVA using participants' performance on the second and third rounds of the karaoke as the dependent measure and condition as the independent variable. Since we found evidence of learning occurring across rounds, $F(1, 106) = 173.07$, $p < .001$, $\eta 2p = .62$, we also added performance on the first round (before our manipulation occurred) as a control. Consistent with our expectation, we found that participants in the treatment group sang more accurately in the rounds following our manipulation compared to participants in the control group ($M_{reflection} = 41.85\%$, $SD = 22.03\%$ vs. $M_{control} = 30.00\%$, $SD = 26.41\%$), $F(1, 106) = 7.48$, $p = .007$, $\eta 2p = .07$.

Additional Study #2: Reflection vs. practice vs. control in tumor cell count (N = 453). We recruited 453 adults on Amazon Mechanical Turk (49.4% male, $M_{age} = 32.52$, $SD = 7.94$) to participate in an online study in exchange for $2. To be eligible for the study, participants were required to be located in the United States and to pass a colorblindness test (necessary to effectively complete the study's task). Participants were randomly assigned to one of three between-subjects conditions: reflection, practice, or control. We told participants that they were participating in a task called "Tumor cell count task." In the task, they would be counting the number of tumor cells that appeared within an image. The instructions informed participants that "Although we already have initial counts from one source, we need humans to verify the count of tumor cells within each image. Given the nature of the task, you may not participate if you are colorblind." This task was adapted from the paradigm used by Chandler and Kapelner (2013). We used this task to provide a context that would be challenging and unfamiliar to participants, thus making it more likely that they would learn across rounds.

After answering demographic questions, including the colorblindness test mentioned earlier, we included two attention filters. Participants who failed one or both attention filters were automatically brought to a screen telling them that, given their answers, they could not participate in the study. Their data was not recorded. Next, participants received the instructions for the tumor cell count task and were given the opportunity to see an example of the task they would be completing. When they felt sufficiently prepared, participants could advance to the first part of the task, in which they saw six images, each representing a blood smear that contained several tumor cells (see Figure A2 for an example). For each image they saw on the screen, participants had to indicate the number of tumor cells present in the blood smear.

Figure A2. Example of Blood Smear Participants Were Asked to Analyze



After analyzing six blood smears in the first round, participants were randomly assigned to one of three conditions: reflection, practice, or control. In the *reflection* condition, participants read the following:

Please take the next few minutes to reflect on the task you just completed. Please write down your reflections and be as specific as possible. You will have THREE minutes to engage in this reflection. The study will advance to the next stage once the THREE minutes are over.

In the *practice* condition, participants were provided the following instructions:

Please take the next few minutes to practice some more on the task you just completed. Below you'll see a few images. For each, you can count the number of tumor cells that are present. You will have THREE minutes to practice on the images. The study will advance to the next stage once the THREE minutes are over.

In the *control* condition, we instructed the participants as follows:

Please take the next few minutes to read the short story below. We'll ask you a few questions about it after you're done reading it. You will have THREE minutes to read the story. The study will advance to the next stage once the THREE minutes are over.

After completing the task used as manipulation, participants advanced to a second round of the tumor-cell count task with a series of six different blood smears. As our main dependent measure, we computed participants' performance on the second round of the task in terms of their accuracy (scoring their tumor cell count against the correct count) for each of the six blood smears analyzed in the round. Note that lower values indicate greater accuracy. Additionally, we captured their first-round score (which we used as a control in our analyses). We also computed the difference in these two accuracy scores to identify the (likely) improvement in accuracy observed across rounds. Table A1 below reports descriptive statistics by condition.

Table A1. Means (and Standard Deviations) by Condition

Condition	Accuracy score round 1	Accuracy score round 2	Accuracy improvement
Control	127.37	98.14	29.22
	(126.16)	(106.37)	(86.12)
Practice	117.21	92.46	24.76
	(116.97)	(115.33)	(68.31)
Reflection	135.75	87.16	48.58
	(135.51)	(114.78)	(74.07)

Results. We conducted an ANOVA using participants' accuracy on the second set of the tumor-cell count task (after our manipulation took place) as the dependent measure and condition as the independent variable, controlling for accuracy on the first round (before our manipulation occurred). We found a significant effect of condition on the accuracy score in the rounds after the manipulation occurred, F (2, 449) = 3.52, p = .031, η^2_p = .015. As one may expect, the accuracy score in round 1 predicted the accuracy score in round 2, F (2, 449) = 802, p < .001, η^2_p = .64. Similarly, improvement in accuracy across rounds (comparing the rounds before and after the manipulation occurred) varied by condition, F (2, 450) = 4.15, p = .016, η^2_p = .018. When examining differences in improved accuracy across conditions, we found that participants in the reflection condition showed greater improvement in their accuracy of counted tumor cells (M = 48.58, SD = 74.07) as compared to those in the practice (M = 24.76, SD = 68.31; p = .007) and control (M = 29.22, SD = 86.12; p = .028) conditions. The improvement in accuracy did not differ for participants in the practice condition compared to those in the control condition (p = .610).

References

Chandler D, Kapelner A. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *J. Econ. Behav. Organ.*, 90: 123–133.

Appendix 3.
Additional evidence from Study 1

In Study 1, we explored the effect of reflection in the context of an actual organization by means of a field experiment with employees at a large business-process outsourcing firm. The design and procedure are detailed in the manuscript, along with some general results. Here we provide additional analyses and details on an additional condition on which we gathered data.

Additional analyses. In Table 5 in the paper, we show the results of an OLS regression in which we estimated the performance effect of treatment (reflection) versus control (practice) while controlling for trainer-level characteristics. We included robust standard errors and considered the possibility of clustering the standard errors at the trainer level. However, an evident complication with our estimation of the standard errors was that we only had six clusters (the six trainers). The issue of reliability of clustered standard errors with a limited number of clusters has been long debated (see MacKinnon and Webb 2017 for a recent overview). Since our number of clusters was below 10, we could employ wild cluster bootstrapping with the additional adjustment proposed by Webb (2014). When we performed the boottest with 10,000 replications and the Webb weightype (Roodman et al 2019), the bootstrapped p-value for the t statistics from the Wald test was unfortunately above the 5% threshold. We read more about this, and we were particularly interested to learn that Canay et al. (2021) discourage the use of weighting schemes other than Rademacher. We hence looked at recent papers with similar set-ups, such as the forthcoming article by Ghani and Reed (2021), who study the ice industry in Sierra Leone, West Africa. Their empirics are based on 18 months of panel data involving five retailers interacting with 153 fishermen across three wharves. The authors cluster their standard errors at the level of the fisherman (and delivery date), then add fixed effects for retailers in a subsequent model. We hence decided to avoid the clustering altogether and simply control for trainer characteristics instead.

Looking at the results of the regression analyses, it may be worth reminding the reader that participants in the reflection condition were trained by older trainers with fewer months of work experience at Wipro, as per Table 4 in the paper. In Table 5 in the paper, we observe that trainers' age has a negative effect on performance, thus raising no particular concern. However, the sign of coefficient on trainers' work experience suggests that participants assigned to trainers with fewer months of work experience at Wipro (as was the case for trainers assigned to the treatment group) performed better. Given the size of coefficients, it would take a trainer with 72 additional months of experience (13.976/0.193) to completely wipe out the positive effect of reflection on performance. Considering that there is a 36-month difference in experience of trainers across experimental conditions ($M_{reflection}$ = 63.286, SD = 1.546; $M_{practice}$ = 94.936, SD = 27.335), one could conjecture that even if the treatment had been administered by the more experienced trainers, the effect of work experience (+36 months) would not have been high enough to compensate for the increase in the final assessment performance (equivalent to +72 months).

Additional condition. While running Study 1, we collected additional data on a third condition, in which we asked participants to spend the last 15 minutes of their training (1) articulating and codifying the experience accumulated in the past (10 minutes) and (2) sharing this knowledge with another participant (five minutes). We included this condition to examine whether sharing would further increase the benefits of articulation and codification. Forty-four agents were assigned to this condition and instructed as follows:

> *Please take the next 10 minutes to reflect on the training day you just completed. Please write about the main key lessons you learned as you were completing your training. Please reflect on and write about at least two key lessons. Please be as specific as possible. When done, you will be given another 5 minutes to explain these to another participant who is completing the training process with you.*

Table A2 reports mean comparisons across participants allocated to the reflection versus sharing groups. As shown in the table, and similar to what we observed when comparing the reflection and practice groups, participants in the sharing group did not significantly differ from those in the reflection group, except for gender (more male trainees in the sharing group), gender of the trainer (only male trainers in the sharing group), and experience of the trainer (more experienced trainers in the sharing group). In light of these results, in our regression analyses we controlled for trainer characteristics, as we did when comparing reflection to practice.

The t-tests reported at the bottom of the table provide us with a first taste of the differences in performance across sharing and reflection. We observe lower performance for participants in the sharing group both in the assessment test at the end of the training program and in the customer satisfaction surveys completed for the first month after the training was over.

Table A2. Univariate Tests by Condition, Study 1: Reflection vs. Sharing

	Reflection (N=56)		Sharing (N=44)		T-test (Two-tailed)		Cohen's d
	Mean	S.D.	Mean	S.D.	T	P-Value	D
Control Variables							
Participant: Age	24.768	3.618	25.341	3.396	-0.808	0.421	0.163
Participant: Gender	0.732	0.447	0.932	0.255	-2.643	0.009	0.549
Participant: Prior experience	31.245	33.549	29.118	27.780	0.347	0.729	0.069
Trainer: Age	33.679	6.764	33.636	2.252	0.039	0.968	0.008
Trainer: Gender	0.821	0.386	1.000	0.000	-3.061	0.003	0.659
Trainer: Prior experience	63.286	1.546	105.454	4.055	-71.567	0.000	13.742
Dependent Variables							
Final Assessment	71.536	9.785	71.232	10.263	0.149	0.881	0.030
Customer Satisfaction	0.912	0.236	0.600	0.377	3.393	0.001	0.992

Note: In the case of Customer Satisfaction, N = 23 for reflection and N = 27 for sharing.

Next, we replicated the OLS regression we ran before, but this time we included the sharing condition, for a total of 144 observations. Table A3, Model 1 and Model 2 mirror Models 1 and 2 in Table 5 in the paper, with the difference that this time we included both manipulations (*reflection* and *sharing*) and assessed their effects in contrast to practice. Results show that reflection and sharing both improved performance on the final test. As for the effects on customer satisfaction, it is interesting to note that while reflection had no significant effect, sharing seems to have had somewhat of a negative effect—a puzzling finding that we intend to explore in future studies. One important point to clarify would be, for instance, the effectiveness of our manipulation, which could not be assessed in the context of the field study.

Table A3. Results from OLS Regressions, Study 1: Reflection vs. Sharing vs. Practice

	Final Assessment			**Customer Satisfaction**		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Reflection	14.328	2.872	0.000	0.242	0.223	0.281
Sharing	19.487	3.637	0.000	-0.209	0.092	0.026
Participant: Age	-0.769	0.369	0.039	0.006	0.011	0.578
Participant: Male	-2.728	1.780	0.128	0.009	0.103	0.930
Participant: Experience	0.146	0.048	0.003	0.001	0.001	0.389
Trainer: Age	-0.282	0.183	0.126	0.009	0.012	0.455
Trainer: Male	21.767	4.230	0.000	-0.069	0.167	0.682
Trainer: Experience	-0.187	0.078	0.018	0.004	0.004	0.254
Cons	77.122	11.408	0.000	-0.055	0.463	0.907
N		144			84	
F	10.110		0.000	2.590		0.015
R-squared		0.408			0.212	

References

Canay IA, Santos A, Shaikh AM. 2021. The wild bootstrap with a "small" number of "large" clusters. *Rev. Econ. Statistics* 103(2): 346–363.

Ghani T, Reed T. 2021. Relationships on the rocks: Contract evolution in a market for ice. *Amer. Econ. J: Microecon.* Forthcoming. See: https://www.aeaweb.org/articles?id=10.1257/mic.20190166&&from=f

MacKinnon JG, Webb MD. 2017. Wild bootstrap inference for wildly different cluster sizes. *J. Appl. Econ*. 32: 233–254.

Roodman D, Nielsen MØ, MacKinnon JG, Webb MD. 2019. Fast and wild: Bootstrap inference in Stata using boottest. *Stata J*. 19(1): 4–60.

Webb MD. 2014. Reworking wild bootstrap based inference for clustered errors. Queen's Economics Department Working Paper No. 1315.

Appendix 4.
Additional evidence from Study 3

In Study 3, we explored the effect that different ways of reflecting have on performance. To this end, we moved from our generic manipulation, asking participants to elaborate on the strategies they used, to a more specific manipulation, asking them to reflect on their performance and what they thought was required of them to solve the task. The design and procedure are detailed in the manuscript, together with some general results. Here we provide additional analyses and details on an additional condition on which we gathered data.

Analyses. In Table A4, we examine the performance effects of our two reflection treatments by running an OLS regression with robust standard errors. We first estimated the effects of the two treatments (*articulation* and *codification*) vis-à-vis control (*practice*) on *final performance* while controlling for *initial performance* (Model 1). Leveraging the insight that experience can be a good or a bad teacher (March 2010), depending on whether we infer the "right" or "wrong" lessons from it (Levitt and March 1988), we next explored whether the effect of reflection is contingent on participants having an accurate understanding of what they are learning. To this end, we measured the extent to which participants *actually* understood the task relative to *perceiving* that they understood the task. To capture actual task understanding, after the last math puzzle from the last round was solved, we asked participants to answer four questions: an open text question that asked them to write down what they understood about the task, a multiple-choice question based on a verbal description of the task, and two multiple choice questions based on a visual description of the task (since the task was completed starting from a visual stimulus, i.e., the 3 x 4 grid). Two independent coders went through the open text, coded it as correct or not, and then discussed the cases of disagreement until they reached consensus. We aggregated the four measures to generate the variable *actual task understanding*. To capture perceived task understanding, right after the manipulation occurred, we asked participants to assess the following three items (1 = strongly disagree, 7 = strongly agree): (1) "I now understand how to perform this task better," (2) "It is now clearer to me how this task works," and (3) "I now know how to correctly identify the two numbers accurately in this task." We averaged the items into one measure of *perceived task understanding* (α = .84). To better compare our objectively measured *actual task understanding* (min: 0; max: 4) to *perceived task understanding* (min: 1; max: 7), we transformed the variables into dummies taking the value of 1 if actual/perceived task understanding was above the average (M_{actual} = 3.57, SD = 0.64; N = 185; $M_{perceived}$ = 5.16, SD = 1.41; N = 155) and 0 otherwise (N_{actual} = 105; $N_{perceived}$ = 135). The dummy *accurate* identifies participants who scored 1 on both measures for a total of 107 accurate and 183 inaccurate learners in our sample. Out of the 183 inaccurate learners, 48 understood less than they thought (high perceived understanding, low actual understanding), 78 understood more than they thought (low perceived understanding, high actual understanding), and 57 did not understand the task either way (low perceived understanding, low actual understanding). We next re-ran our regression analysis by including our measures for *actual* and *perceived task understanding* (Model 2) as well as the *accurate* dummy (Model 3), and then interaction terms between *accurate* and *articulation* as well as *codification*, first individually (Model 4 and Model 5, respectively) and then jointly (Model 6).

Table A4. Results from OLS Regressions, Study 3: Articulation vs. Codification vs. Practice

	Model 1			Model 2			Model 3		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	0.455	0.262	0.083	0.624	0.258	0.016	0.636	0.265	0.017
Codification	0.069	0.273	0.800	0.246	0.259	0.342	0.257	0.259	0.322
Initial Performance	0.864	0.076	0.000	0.719	0.083	0.000	0.718	0.084	0.000
Actual Task Understanding				0.488	0.166	0.004	0.465	0.196	0.018
Perceived Task Understanding				0.226	0.087	0.010	0.214	0.101	0.034
Accurate							0.079	0.324	0.808
_cons	3.871	0.248	0.000	1.174	0.686	0.088	1.286	0.860	0.136
N	290			290			290		
F	49.204		0.000	34.530		0.000	29.040		0.000
Adjusted R2	0.300			0.333			0.330		

	Model 4			Model 5			Model 6		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	0.904	0.314	0.004	0.558	0.263	0.035	0.656	0.321	0.042
Codification	0.319	0.255	0.213	-0.144	0.323	0.656	-0.083	0.330	0.801
Initial Performance	0.718	0.083	0.000	0.701	0.083	0.000	0.703	0.082	0.000
Actual Task Understanding	0.452	0.194	0.021	0.469	0.193	0.016	0.464	0.193	0.017
Perceived Task Understanding	0.213	0.101	0.036	0.231	0.102	0.025	0.229	0.102	0.026
Accurate	0.316	0.353	0.372	-0.288	0.358	0.422	-0.171	0.415	0.681
…X Articulation	-0.708	0.494	0.153				-0.240	0.535	0.654
…X Codification				1.134	0.472	0.017	1.021	0.511	0.046
_cons	1.207	0.855	0.159	1.412	0.853	0.099	1.373	0.852	0.108
N		290			290			290	
F	26.469		0.000	25.280		0.000	22.941		0.000
Adjusted R^2		0.333			0.340			0.338	

Results across all models show that *articulation* had a positive effect on *performance*, while the same did not hold true for *codification*. Once we included the interaction with the dummy *accurate*, we found no effect in the case of *articulation* (Model 4: ß = -0.708, SE = 0.494, p = 0.153) but a positive interaction effect with *codification* (Model 5: ß = 1.134, SE = 0.472, p = 0.017). In other words, participants assigned to the codification condition experienced an increase in performance only when they had an above-average *perceived task understanding* matched with an above-average *actual task understanding*.

Additional condition. While running Study 3, we collected data on an additional manipulation to start exploring whether there are spillover benefits for the reflection effect, a conjecture we further explored in Study 4. To this end, we collected data from 286 participants (48% male, average age group = 30–34) who went through the exact same experimental protocol, except for in the very last round of math puzzles (round 2). Participants in this condition received the same 3 x 4 grids as other participants, but in the very last round, instead of the usual sum-to-ten game, they were asked to find the smallest and largest numbers among the 12 displayed on the grid. To this end, they were told:

> *The final round is about to start. It consists of 5 puzzles to be solved. You will be solving one puzzle at a time, and for each puzzle you will have 30 seconds to solve it and a 5 seconds break.*
> *IMPORTANT: For this round, you will have to correctly select the two cells corresponding to the biggest and the smallest number within the table. This is different from what you have been doing so far.*

Similar to Study 3, we ran a series of t-tests on the demographic data we gathered at the end of the study. Results show that all the characteristics we controlled for were evenly distributed among participants across different cells.[22] Importantly, participants who were assigned to this condition did not differ significantly from those who were assigned to the main study described in the paper.[23]

In Table A5, we show what happened when we replicated Model 1 and Model 2 from Table A4 on these additional data. Contrary to what we observed in the main study, where participants benefited from being assigned to the articulation condition, participants in this additional study were somewhat disadvantaged when asked to reflect instead of practicing: The effect of *articulation* compared to *practice* was negative and marginally significant (Model 1: ß = -0.448, SE = 0.244, p = 0.068), but we found no evidence of any effect when we controlled for the extent to which participants understood the task at hand (Model 2: ß = -0.243, SE = 0.257, p = 0.346). To make better sense of this finding, we looked at data on average performance and completion time across the two different tasks. This comparison revealed that participants found this alternative task much easier: They took an average of 13.47 seconds and correctly solved 4.28 puzzles, relative to 20.47 seconds and 3.29 puzzles for participants who kept solving sum-to-ten games. This seems to suggest that the marginal benefits of reflection may become inferior to those associated with practice when engaging with a rather trivial

[22] We detected one significant but small difference in the practice group, where participants came from a slightly less advantaged socioeconomic background (Cohen's d = 0.19).

[23] We detected two significant but small differences: Participants in the related condition were more likely to identify as women (Cohen's d = 0.18) and had a slightly higher education level (Cohen's d = 0.13).

task that does not require effort to progress along the learning curve. This is an interesting conjecture that may be worth exploring in future studies.

Table A5. Results from OLS Regressions, Study 3: Alternative Task

	Model 1			Model 2		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	-0.448	0.244	0.068	-0.243	0.257	0.346
Codification	-0.221	0.248	0.375	-0.072	0.239	0.763
Initial Performance	0.702	0.082	0.000	0.559	0.089	0.000
Actual Task Understanding				0.583	0.156	0.000
Perceived Task Understanding				0.207	0.090	0.023
_cons	5.702	0.300	0.000	2.809	0.754	0.000
N		286			286	
F	27.063		0.000	21.941		0.000
Adjusted R^2		0.242			0.290	

References

Levitt B, March J. 1988. Organizational learning. *Ann. Rev. Soc.,* 14: 319–340.
March JG. 2010. *The Ambiguities of Experience*. Cornell University Press, Ithaca, NY.

Appendix 5.
Additional evidence from Study 4

In Study 4, we explored whether the benefits of reflection are sticky or can spill over to related tasks. To this end, we modified the experimental task, devising a more complex task and a study in which, after going through a practice round and the administration of our treatment, participants were asked to solve the exact same task or to move to a related, but different, one. The design and procedure are described in the manuscript, along with detailed results. Here we provide preliminary results on potential mechanisms behind the effect of reflection.

Potential mechanisms. While running Study 4, we collected additional data to start exploring potential mechanisms behind the effect of reflection. The analysis is exploratory in nature since we did not manipulate, but rather measured, potential mediators. This is standard procedure in the field—to the point that the measurement of mediation has been defined as the gold standard of mediation studies (Spencer et al. 2005)—but it has been consistently shown to bias estimates through measurement error and omitted variable bias (Pierce and Snyder 2020). As a result, we invite the reader to consider the results of our mediation analyses as preliminary descriptive evidence of the role of potential mechanisms put forward by prior work.

The first mechanism we looked at was perceived self-efficacy. As discussed in the paper, prior work by Anseel and colleagues (2009) has argued that when individuals are given feedback on their prior performance, they experience higher self-efficacy and perform better in the future. We measured self-efficacy as perceived by participants after they were administered our manipulation. In particular, we assessed *perceived self-efficacy* using a four-item measure adapted from Bandura (1990), asking participants to rate the extent to which they agreed with four statements (from 1 = Strongly disagree, to 7 = Strongly agree), namely: (1) "Right now, I feel capable"; (2) "Right now, I feel competent"; (3) "Right now, I feel able to make good judgments"; and (4) Right now, I think I can manage to solve difficult problems if I try hard enough." We averaged the items into one measure ($\alpha = .95$).

The second mechanism we looked at was the reduction of causal ambiguity. We have seen how reflection increases the ability to understand the causal relationship between actions and outcomes (Zollo and Winter 2002, Kale and Singh 2007) and generate richer mental models (Ellis and Davidi 2005, Goh et al. 2013). We speculate that reflection will strengthen one's understanding of the task by reducing a person's experience of causal ambiguity, or the perceived uncertainty about the causal relationships that link actions and associated outcomes. As a result of this decreased uncertainty, individuals will perform better. To capture the improved mental models of our participants, we used our measure of *actual task understanding*, which we previously described as being assessed through four questions we asked participants to answer after they finished solving the last matrix.

To start exploring the mediating role of *perceived self-efficacy* and *actual task understanding*, we constructed bias-corrected confidence intervals based on 10,000 random samples with replacement from the full sample (Preacher and Hayes 2004). We ran the bootstrap procedure twice, once for articulation and once for codification. Given our empirical design, we tested for the indirect effect of perceived self-efficacy and actual task understanding on the interaction between articulation (or codification) and same task. Results should hence be read as estimating the indirect effect in the case of same task compared to related one. We report the 95% bias-corrected confidence intervals in Table A6.

In the case of articulation, the results suggest the presence of a mediation effect of actual task understanding [0.113, 0.364] and an overall mediation effect when one includes both perceived self-efficacy and actual task understanding [0.103, 0.376]. In other words, participants who were asked to reflect about what was required to solve the task at hand ended up understanding the task better.

In the case of codification, we observed a mediation pattern both for perceived self-efficacy [-0.129, -0.008] and actual task understanding [0.077, 0.334]. The total indirect effect was also significant [0.008, 0.293]. We find it interesting that the indirect effect of self-efficacy was negative in the case of codification, thus suggesting that participants who were asked to write down what one can do to effectively solve the task at hand

experienced a decrease in self-efficacy while, at the same time, understanding the task better. The positive sign of the total indirect effect suggests a prevalence of the effect of actual task understanding.

Table A6. An Exploration of Mediation Effects

	Indirect effect of:		Total Indirect Effect
	Perceived Self-Efficacy	Actual Task Understanding	
Articulation X Same task	[-0.043, 0.046]	[0.113, 0.364]	[0.103, 0.376]
Codification X Same task	[-0.129, -0.008]	[0.077, 0.334]	[0.008, 0.293]

Overall, these results suggest there may be merit in further exploring these two mechanisms as drivers for the beneficial effects of articulation and codification on performance. We hesitate to make any causal claim, given our approach to assessing mediation, but see potential in advancing this line of enquiry.

References

Anseel F, Lievens F, Schollaert E. 2009. Reflection as a strategy for enhancing the effect of feedback on task performance. *Organ. Behav. Human Decision Processes*, 110: 23–35.

Bandura A. 1990. *Multidimensional scales of perceived academic efficacy*. Stanford, CA: Stanford University Press.

Ellis S, Davidi I. 2005. After-event reviews: Drawing lessons from successful and failed experience. *J. Appl. Psych.* 90(5): 857–871.

Goh KT, Goodman PS, Weingart LR. 2013. Team innovation processes: An examination of activity cycles in creative project teams. *Small Group Res.*, 44: 159–194.

Kale P, Singh H. 2007. Building firm capabilities through learning: The role of the alliance learning process in alliance capability and firm-level alliance success. *Strategic Management J.*, 28: 981–1000.

Pierce L, Snyder JA. 2020. Historical origins of firm ownership structure: The persistent effects of the African slave trade. *Acad. Management J.,* 63(6): 1687–1713.

Preacher KJ, Hayes AF. 2004. SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behav. Res. Methods Instruments Computers*, 36: 717–731.

Spencer SJ, Zhanna MP, Fong GT. 2005.Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *J Pers. Soc. Psychol.* 89(6): 845–851.

Zollo M, Winter SG. 2002. Deliberate learning and the evolution of dynamic capabilities. *Organ. Sci.*, 13(3): 339–351.

Appendix 6.
Additional evidence from Study 5

In Study 5, we explored whether the marginal returns of reflection change depending on the amount of experience on which one can reflect. To this end, we also codified the reflection logs of participants assigned to the codification condition. We describe the coding procedure and categories in the manuscript, along with detailed results. Here we provide some additional details on the coding categories related to the content of the reflection logs. As a reminder, we decided to codify the content of the reflection logs in terms of whether it mentioned the task at hand (dummy), the strategy used to solve it (dummy), and the participant's performance on the task (dummy). Table A7 includes a collection of exemplary reflection logs for each of the three categories of content.

Table A7. Exemplary Reflection Logs Focused on Different Categories of Content, Study 5

Content	Example
Task	I did not use any strategy. This really confused me, but it was intriguing. Just study the pattern and how it affects the others. I attempted to solve the puzzles by thinking about the sequence of the patterns involved and the way in which they were presented. Some are more difficult, whilst others seem to jump out at you. It was very easy to assess the various rows and columns to see which pattern logically followed in the sequence. None of the puzzles were challenging though I am well aware that this is only the very first section.
Strategy	I look for movements, repetitions, and orders. i also try to eliminate some choices. The most effective way to complete the puzzles is to split the overall image into sections rather than having to deal with it as a whole. I split up the images into their components. Then look for patterns of repetition, add, deletion, rotation. Then look at the other elements and apply the same to them. Gradually I build up the missing shape.
Performance	I just tried to use the other figures in the matrix to match up the last one. I am not going great at it though. I hope the way i am trying will work next round. The first one i looked at the additional piece attached, the second one i wasn't too sure i got right, that was trickier for me. it was between 2 for me and i chose the wrong one. I was looking for the pattern to follow, which is sometimes easier than at other times! My brain works well with letters and words, so shapes are a bit problematic, so I need to focus more and be more logical. the first three tasks made sense, but I just couldn't see the correct answer for the fourth problem, so I just guessed.

Appendix 7.
Do the benefits of reflection change over time?

We report results from two additional studies exploring whether the performance benefits of reflection vs. practice change over time as one repeatedly engages with reflection vs. practice. To this end, we designed a replication of Study 5 in which we required participants to engage with the same experimental task in multiple rounds, each followed by the same experimental treatment of reflection or practice.

Design and procedure. We recruited 900 adults to complete a brain teaser under time pressure in exchange for £3.45 and the potential to earn an additional £1.40 bonus based on performance. Participants were recruited through Prolific, which provided them with a brief description of the study and explained that 10% of them would receive the extra bonus based on their performance. As for Study 5, to be eligible, participants had to be located in an English-speaking country and have a good track record on the platform. We adapted the structure of Study 5 and designed the study around four rounds of matrices, with our experimental manipulation (articulation vs. codification vs. practice) following each of the first three rounds. This means that participants alternated between practicing for about two minutes and engaging in either reflection (articulation vs. codification) or additional practice (five matrices) for three minutes three consecutive times. After these three rounds, all participants solved a final set of matrices.

We manipulated our three experimental conditions (articulation vs. codification vs. practice) by using the same manipulations adopted in Study 5. As before, we selected matrices based on the probability of solving each of them (Arthur and Day, 1994). We ran two separate studies that were identical in all aspects except for the matrices participants were asked to solve. In Additional Study #3 (N = 599), we created four rounds of two matrices each with substantial increases in difficulty between rounds: The probability of correctly solving the matrices was, respectively, 90.84% for round 1, 88.37% for round 2, 83.91% for round 3, and 72.53% for round 4. We then ran Additional Study #4 (N = 301), a replication based on easier matrices so that the increments in difficulty between rounds became more substantial only at the end, with respectively 94.31% for round 1, 90.84% for round 2, 93.07% for round 3, and 88.12% for round 4 (for which we used four instead of two matrices). We ran this replication because when we examined the results of Additional Study #3, we noticed that all participants, independent of the experimental condition, experienced a sharp decline in performance in the last two rounds. The comments left by participants at the end of the experiment made it clear that they had experienced stress and frustration by the end of the study due to the increasing difficulty. For instance, one participant wrote: "*I found the last test very difficult, partly because I became more aware of the time pressure, and I felt very stressed. However, while my brain had difficulty focusing because of stress, I also think the last round of tests were more difficult to solve - at least for me!*" Another remarked: "*Wow!! That was stressful! I expected more from myself, though*." A third one noted: "*Found the task quite difficult and stressful, made me feel tense as time counted down*." In light of these remarks, we immediately launched a follow-up study that was identical in all aspects but the increase in the level of difficulty of matrices participants had to deal with across rounds. We discuss results from both below.

Measures. We manipulated condition as described above, identifying each treatment with a dummy variable equal to 1 in the case of "high" and 0 otherwise. Our dependent variable, *final performance,* was a count variable ranging between 0 and 2 (Additional Study #3) or 4 (Additional Study #4) and corresponding to the number of Raven matrices that were solved correctly in the last round. We controlled for performance in the practice round (*initial performance*). We gathered the same demographic data as in Studies 3–5, including our pandemic control. Results from a series of t-tests show that all characteristics were evenly distributed among participants.[24]

A problematic aspect of this study was that, by design, participants had to go through three rounds of reflection or practice for a total of nine minutes spent on the manipulations. This made us particularly concerned about compliance, as it seemed likely that some participants could have become distracted and taken the manipulation less seriously over time. To control for this aspect, we introduced a question after the experimental task was completed and before we collected demographic information, asking participants to describe how they spent the three minutes in between rounds. We asked two independent coders to go through

[24] We detected a significant but small difference related to higher neuroticism for participants to Additional Study #3 who were assigned to the articulation condition (Cohen's d = 0.18).

these texts and code whether they accurately described the manipulation to which participants had been assigned, as well as whether they provided any hints that participants did not fully comply. The two coders went through the texts independently and then discussed the very few cases of disagreement until they reached consensus. This exercise delivered two interesting results. First, it showed that almost all participants (Additional Study #3: 96%; Additional Study #4: 98%) reported doing, at least for some time, what they had been asked to do. Interestingly, it also showed that a non-trivial percentage of participants (Additional Study #3: 19%; Additional Study #4: 23%) explicitly reported using part of the time allocated to the experimental manipulation to do something else. Admittedly, we were surprised by the candor with which participants reported being involved in activities ranging from resting ("*I sat and thought for a minute about my strategy and wrote it down. usually this only took 1 or 2 minutes so i just sat and had a rest for the remaining time*") to doing other tasks at home ("*Practiced the matrices till I had completed them all, then did a bit of tidying while I waited for the 3 minutes to end*") or getting distracted by their phone or computer ("*I went through all of the matrices (which usually took about a minute and a half) and then I looked at my phone during the remaining time*"). In the light of these observations, we decided to run our analyses both including and excluding the sub-sample of participants for whom we had evidence of non-compliance. We report results of both analyses next.

Results. We performed regression analyses where we investigated differences in *final performance* while controlling for *initial performance* by running OLS regressions with robust standard errors. We report the results for Additional Study #3 in Table A8 (excluding non-compliers) and Table A9 (full sample), while those for Additional Study #4 are reported in Table A10 (excluding non-compliers) and Table A11 (full sample). In all tables, we started by estimating the effects *articulation* and *codification* vis-à-vis *practice* on *final performance* in round 2, round 3, and round 4. We then replicated the same analysis using the time taken to solve the matrices in rounds 2, 3, and 4, respectively, as our dependent variable.

Starting with Additional Study #3, Table A8, our results show that participants in the *codification* condition outperformed those in the *practice* condition in round 2 (ß = 0.213, SE = 0.072, p = 0.003). The effect disappears over the following rounds, where codification comes at the cost of more time spent solving the matrices, marginally in round 3 (ß = 4.406, SE = 2.344, p = 0.061) and significantly in round 4 (ß = 5.494, SE = 1.711, p = 0.001). As for *articulation*, we found no evidence of a significant performance increase compared to participants who engaged in practice, but our results suggest that by the time they reached the last round, these participants were taking more time to solve the matrices (ß = 3.637, SE = 1.702, p = 0.033).

Table A8. Results from OLS Regressions, Additional Study #3: Excluding non-compliers

	Score								
	Round 2			**Round 3**			**Round 4**		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	0.114	0.076	0.134	-0.022	0.086	0.801	0.008	0.079	0.917
Codification	0.213	0.072	0.003	0.071	0.084	0.399	0.030	0.079	0.710
Initial Performance	0.338	0.044	0.000	0.376	0.049	0.000	0.306	0.040	0.000
_cons	0.868	0.086	0.000	0.718	0.093	0.000	0.334	0.072	0.000
N	485			485			485		
F	23.897		0.000	19.810		0.000	19.765		0.000
Adjusted R^2	0.134			0.110			0.084		
	Time								
Articulation	-0.152	1.687	0.928	2.796	2.219	0.208	3.637	1.702	0.033
Codification	-1.710	1.712	0.318	4.406	2.344	0.061	5.494	1.711	0.001
Initial Performance	0.495	0.040	0.000	0.567	0.054	0.000	0.380	0.040	0.000
_cons	33.237	2.750	0.000	20.728	3.717	0.000	44.582	2.818	0.000
F	52.817		0.000	39.342		0.000	34.497		0.000
Adjusted R^2	0.236			0.190			0.173		

As shown in Table A9, the results did not change when we replicated the analyses on the full sample—that

is, including participants who explicitly admitted non-compliance. Participants in the *codification* condition outperformed those in the *practice* condition in round 2 (ß = 0.186, SE = 0.064, p = 0.004). The effect disappeared over the following rounds, where codification came at the cost of more time spent solving the matrices in round 3 (ß = 4.426, SE = 2.066, p = 0.033) and significantly in round 4 (ß = 5.507, SE = 1.504, p < 0.001). This was similar for *articulation*, for which we found that by the time participants reached the last round, they were taking more time to solve the matrices (ß = 3.756, SE = 1.518, p = 0.014).

Table A9. Results from OLS Regressions, Additional Study #3: Full sample

	Score								
	Round 2			**Round 3**			**Round 4**		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	0.076	0.066	0.252	-0.008	0.075	0.913	-0.023	0.070	0.740
Codification	0.186	0.064	0.004	0.065	0.074	0.382	-0.003	0.071	0.962
Initial Performance	0.365	0.041	0.000	0.400	0.045	0.000	0.301	0.038	0.000
_cons	0.863	0.081	0.000	0.725	0.087	0.000	0.366	0.069	0.000
N		599			599			599	
F	30.261		0.000	26.723		0.000	21.597		0.000
Adjusted R^2		0.144			0.120			0.078	
	Time								
Articulation	-0.389	1.512	0.797	2.116	1.958	0.280	3.756	1.518	0.014
Codification	-2.316	1.516	0.127	4.426	2.066	0.033	5.507	1.504	0.000
Initial Performance	0.494	0.035	0.000	0.600	0.047	0.000	0.391	0.035	0.000
_cons	33.280	2.423	0.000	18.227	3.154	0.000	43.923	2.415	0.000
N		599			599			599	
F	66.833		0.000	57.624		0.000	47.228		0.000
Adjusted R^2		0.246			0.219			0.185	

We next look at the results from Additional Study #4, a replication based on easier matrices so that the increments in difficulty between rounds became more substantial only at the end.

Table A10. Results from OLS Regressions, Additional Study #4: Excluding non-compliers

	Score								
	Round 2			**Round 3**			**Round 4**		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	-0.033	0.095	0.727	0.076	0.084	0.366	0.199	0.175	0.256
Codification	0.075	0.086	0.383	0.112	0.083	0.179	0.444	0.170	0.010
Initial Performance	0.410	0.070	0.000	0.314	0.069	0.000	0.548	0.105	0.000
_cons	0.962	0.136	0.000	1.142	0.132	0.000	1.677	0.221	0.000
N		231			231			231	
F	12.232		0.000	7.630		0.001	11.391		0.000
Adjusted R^2		0.183			0.136			0.121	
	Time								
Articulation	-0.046	2.561	0.986	2.753	2.570	0.285	9.240	4.885	0.060
Codification	1.938	2.556	0.449	-0.845	2.474	0.733	2.835	4.775	0.553
Initial Performance	0.482	0.056	0.000	0.527	0.065	0.000	0.790	0.105	0.000
_cons	21.173	3.286	0.000	11.173	3.649	0.002	63.951	7.066	0.000
N		231			231			231	
F	25.350		0.000	23.171		0.000	20.172		0.000
Adjusted R^2		0.196			0.246			0.180	

Starting with Table A10, our result show that participants in the *codification* condition only outperformed

those in the *practice* condition in round 4 (ß = 0.444, SE = 0.170, p = 0.010), with no apparent cost in terms of additional time spent on the matrices. As for *articulation*, we found no evidence of a significant performance increase and only a marginal increase in the time taken to solve the matrices in the very last round (ß = 9.240, SE = 4.885, p = 0.060).

Table A11 replicates the same analyses using the full sample of participants. We note one important difference from the results excluding non-compliers: We no longer found a significant effect of *codification* in round 4—not surprising, given the problem with compliance and the smaller N of this study (301 vs 599 participants). The results do not change in the case of *articulation*, for which we only observed a marginal increase in the time taken to solve the matrices in the last round (ß = 7.841, SE = 4.165, p = 0.061).

Table A11. Results from OLS Regressions, Additional Study #4: Full sample

	Score								
	Round 2			**Round 3**			**Round 4**		
	Coef	**SE**	**p-value**	**Coef**	**SE**	**p-value**	**Coef**	**SE**	**p-value**
Articulation	-0.002	0.077	0.980	0.032	0.067	0.627	0.134	0.148	0.365
Codification	0.022	0.076	0.772	0.038	0.071	0.590	0.189	0.149	0.208
Initial Performance	0.402	0.065	0.000	0.324	0.064	0.000	0.590	0.098	0.000
_cons	0.997	0.128	0.000	1.187	0.124	0.000	1.805	0.206	0.000
N	301			301			301		
F	12.925		0.000	8.609		0.000	12.259		0.000
Adjusted R^2	0.175			0.146			0.115		
	Time								
Articulation	-0.175	2.217	0.937	3.493	2.187	0.111	7.841	4.165	0.061
Codification	3.162	2.329	0.176	1.742	2.215	0.432	5.253	4.176	0.209
Initial Performance	0.477	0.053	0.000	0.504	0.058	0.000	0.743	0.092	0.000
_cons	20.928	3.074	0.000	11.245	3.204	0.001	64.339	5.878	0.000
N	301			301			301		
F	28.146		0.000	26.271		0.000	22.897		0.000
Adjusted R^2	0.195			0.230			0.164		

Overall, our results show that the benefits of reflection tend to taper off over time and that the time at which the benefits of reflection kicked in changed depending on how challenged participants were by the task at hand. In Additional Study #3, where the level of difficulty of the matrices grew constantly and sharply, participants were able to grasp the benefits of reflection right at the beginning of their learning curve, with reflection providing no advantage over practice afterwards. In Additional Study #4, where the level of difficulty grew minimally across rounds, the benefits of reflection became substantial only later, when the task at hand started to become more challenging.

We also want to note the absence of significant performance benefits associated with articulation—a finding that should be taken with caution, given the compliance issues we experienced with the study. Participants were asked to spend a total of nine minutes engaging with our treatment. But, while participants in the practice condition were given new matrices to solve and those in the codification condition had to write down the results of their reflection efforts, participants in the articulation condition were simply asked to reflect to themselves—admittedly, an activity from which it was easier to be distracted. Despite our attempts to account for compliance (and lack thereof), we recommend replicating this study in a laboratory setting, where it would be more difficult for participants to get distracted during the administration of the treatment.