Psychology Research and Practice Vol 3 Issue 1 2024
DOI: 10.37155/2972–3086–0301–4

Original Research Article

Open Access

Omniscient Pte.Ltd.

Retrospective Time Estimation: Comparing the Judgment of Resolving Time (JoRT) with Actual Resolving Time (ART) to Assess over- and Underconfidence

Lisa K. Son[1*], Seok-sung Hong[1], Jini Tae[2], Tae Hoon Kim[3], Yoonhyoung Lee[4]

[1]Department of Psychology, Barnard College
[2]Gwangju Institute of Science and Technology
[3]Kyungnam University
[4]Yeungnam University

***Corresponding to:** Prof. Lisa K. Son, Department of Psychology, Barnard College, Columbia University. Email: lson@barnard.edu.

Received: Jan 18, 2024. **Accepted:** May 10, 2024

How to cite: Lisa K. Son, Seok-sung Hong, Jini Tae, Tae Hoon Kim, Yoonhyoung Lee. Retrospective Time Estimation: Comparing the Judgment of Resolving Time (JoRT) with Actual Resolving Time (ART) to Assess over- and Underconfidence. *Psychology Research and Practice*, 2024; Vol 3(2024)
Doi: 10.37155/2972-3086-0301-4

Abstract: The commonly touted description of *hindsight bias*, where we believe that "we knew it all along," has us assume that after having learned something, we were, to some degree, a "natural." One's time estimation of a prior task, -- what we call the *Judgment of Resolving Time* (JoRT) --however, has not been tested. That is, do people "forget" all of the past time that they had invested into learning? Or, do they believe that they "knew it only somewhat faster" than the time it actually took to complete prior tasks? In the current study, we compared individual's JoRTs with time actually taken to resolve problems, and used the difference as a proxy for confidence. Specifically, we hypothesized that participants' JoRTs would be slightly shorter than the actual time it took to resolve problems, given the prevalence of the hindsight bias. Surprisingly, this overconfidence was not found. On the contrary, people's JoRTs, in both the United States (Experiment 1) and South Korea (Experiment 2), turned out to be *longer* than their actual resolving times, suggesting, we propose, a type of underconfidence. These results offer a potential new strategy for countering the bias -- retrospective time estimation -- while also providing a new tool in which to examine both over- and underconfidence.
Keywords: Hindsight bias; Time estimation, confidence; Overconfidence; Metacognitive judgments

1. Introduction

In the metacognitive field, confidence has been measured in a variety of ways. Researchers have examined different judgments that people make regarding their own knowledge, with the critical aim of discovering if and whether there is any systematic over- or underconfidence. For instance, the widely investigated *judgment of learning* (JOL) is one where individuals predict how well they will perform on a future test (Putnam et al., 2022). The accuracy of these judgments is critical because if one is overconfident, they are likely to cease study prematurely (Fiedler et al., 2019); if one is underconfident, they may overstudy, in vain (Karpicke et al., 2009). Data have shown that in many cases, such prospective judgments have shown that people tend to be overconfident, assessing their knowledge to be somewhat higher than what is found on a later test (Koriat et al., 1980). Similar findings have been shown for retrospective judgments. For example, one judgment that has been shown to be considerably stubborn is one where individuals display hindsight bias, which describes the act of overestimating the amount of knowledge that one had in the past, given updated current knowledge, also known as the "knew it all along effect" (Roese & Vohs, 2012). In the current study, we investigated what we believe to be related to the hindsight judgment, and what we call the *Judgment of Resolving Time* (JoRT), or one's time estimation, on average, as to how long it took for one to come to a resolution on a set of prior tasks. Given past evidence of overconfidence, we hypothesized that people would judge that the average time it took to complete prior tasks would be shorter than the actual average time it took to complete those tasks. In other words, people would mistakenly believe that they not merely "knew it *all along*" but, more accurately, "resolved it *some degree* faster." Such evidence would allow us to understand the extent to which people are overconfident, and provide further support for the notion that people believe that they are somewhat of "a natural" when it comes to the problems they face.

2. The Lay Person's Question

Practice makes perfect. We all know this phrase – essentially, it means that the amount of time we take on a task will be correlated with how well we perform. While there are other variables that will influence accuracy – the depth of processing (Craik & Simon, 1980), the spacing or scheduling of study (Kornell, 2009), and the amount of active versus passive study (Gureckis & Markant, 2012), to name a few – in general, the amount of time we spend on a task is important (Tullis & Benjamin, 2011). But is one's *assessment* of the time that is spent on a task important as well? This retrospective time estimation judgment, while assumed, has not directly been tested in the metacognitive realm. We believe that one's judgment of how long it took to complete a task may even be more crucial than the actual time it takes, because our judgment, like all other metacognitive judgments, is likely to influence how we approach a subsequent task (Metcalfe & Finn, 2008). Consider the following anecdotal examples.

First imagine a task that takes a relatively long time, say, learning French as a second language. It is quite different from English, and it takes you a good couple of months to learn. But through determination and hard work, you reach a level of proficiency that allows you to converse with native French speakers. Then, at some later time, you decide you want to learn German as a third language. As you begin to study – perhaps after a few weeks – you find yourself very frustrated and judge that you will "never be able to learn it." Soon after, you are ready to quit. You think to yourself, "French was so easy, but I'm not getting the German. I must not be a language person after all." Could this frustration, and eventual giving up, be due to an inaccurate judgment of how long it took to learn a new language previously? More specifically, could it be that your judgment of your past learning of French indicates that you "knew it faster," and, in turn, that you should now, with a completely new language, "know it faster?"

Here is another example, one that would require a much shorter time than in the example above. Imagine buying a "Where's Waldo" book for your child. Upon receiving the book, your child immediately begins to search for Waldo on page one, and after 15 minutes, succeeds in finding him. When your child's sibling enters the room, child says to the sibling, "Hey, try finding Waldo." The sibling begins to search, and after about 10 minutes, the child grows impatient, unaware of the time it had taken them to find Waldo in hindsight, and cries, "Still searching? But it's so easy!"

Like before, if a child judges that they had found Waldo faster (say, in 10 minutes) than their actual search time (15 minutes), this might affect not only their own frustration, but the reactions from other naïve learners as well. It's not difficult to imagine the sibling being embarrassed, wondering if they are too slow, or, sadly, giving up.

Finally, consider an example that has been tested where researchers can measure hindsight bias that occurs for tasks requiring very short periods of time. One procedure, related to the Waldo example above, has been replicated in the laboratory and tests the notion of what has been called *visual hindsight bias*. Harley et al. (2004) presented degraded celebrity faces to participants and asked them to identify each face. They continued to show gradually clearer pictures of each celebrity and recorded the moment at which the participant was able to correctly identify the face. Afterwards, participants were asked to identify the level of blur at which they had been first able to identify the celebrity. Their results showed that, with familiar celebrity faces, visual hindsight bias was significant. That is, they thought that they had identified the face much earlier than they had actually identified the face. This and other laboratory examples of visual hindsight bias (Bernstein et al., 2004) suggest that once we have "seen" someone or something, we may be overconfident in that we feel that the road to "seeing" that someone or something was shorter than it had actually been. In other words, just as we had "known it all along," we seem to also think that we "saw it all along."

As in all of the above examples, we imagine that people's JoRTs, as compared to the actual times needed to resolve a prior task, will support the notion of prevalent overconfidence, while also supplementing much of the literature on hindsight bias. In the current research, we consider that the harm of such a bias – as measured by mis-judged time estimation – can include the fact that when people are not accurate at remembering past time, the value of time may decrease, just when we need to depend on (the idea of) them for subsequent tasks. Thus, in the current research, we asked the core question that must be asked: Can the difference between actual resolving time (ART) on past tasks and JoRTs be one of the reasons for why overconfidence, and, consequently, a mistargeted time allocation judgment on a subsequent task, might occur?

Time estimations and time estimation discrepancies (between one's JoRTs and one's ARTs) have not been considered as a means of supplementing the literature on hindsight bias. However, this very prevalent bias has been described in a too-general manner, where data have not been able to say explicitly how severe the bias tends to be. In the above-mentioned studies on visual hindsight bias, there seems to be evidence that participants did, indeed, acknowledge that there was some positive amount of time spent on identifying the celebrity faces (Bernstein et al.; Harley et al., 2004). That is, participants did not select the very first, extremely degraded, photo as the point at which they had identified the celebrity. Rather, they knew that some time had passed before successful identification. And yet, the hindsight bias continues to be described as a bias in which one "saw it *all along*" (Schill et al., 2023, italics added), "expected it *all along*" (Greene et al., 2023, italics added), or "knew it *all along*" (Hom Jr, 2023) (italics added). While our primary purpose is to understand how the JoRTs may be used to measure accuracy in retrospective time estimations of completing a task, we also believe that this judgment may be a novel way in which to qualify the notion of "all along." That is, we would be able to see *how much more* quickly, than actual, one tends to estimate past time on task. Thus, in the following section, we provide a visual framework describing how JoRTs fit into a model of hindsight bias.

3. Using JoRTs as a Supplement to test for Hindsight Bias

Hindsight bias is a robust phenomenon that occurs when outcome knowledge interferes with the ability to accurately recall judgments made in a previous, naïve state (Bernstein, 2021; Chen et al., 2021; Greene et al., 2023; McDermott et al., 2020; Welsh, 2020; Zimdahl & Undorf, 2021). Found in a variety of situations, it refers to the act of overestimating knowledge one had in the past, given current knowledge. In 2003, Hoffrage and Pohl describes the bias as "a projection of new knowledge into the past accompanied by a denial that the outcome information has influenced judgment." In short, the bias supports the notion that learners had never really learned something, but, rather, had "known it all along" (Musch & Wagner, 2007). Fischoff originally defined the term as "creeping determinism,"

describing that as more information is accumulated as one advances in experience and time, one gets closer to getting the final, or fuller, story, a story made up of logical causal links that, in the end, seem "predetermined" or "inevitable" (Fischhoff, 1975). According to this definition, there might not even be a "past" to which to travel back. In other words, as Fischoff writes, "upon receipt of outcome knowledge, judges immediately assimilate it with what they already know about the event in question" (p. 310). Given these perspectives of hindsight bias, one might assume that all of one's past time that had been spent learning had been utterly forgotten. In the realm of problem solving, it could mean that the actual time it had taken to solve a problem, whether that was 5 minutes or 5 days, could be evaluated as having been zero time. Similarly, for a search task, it would mean that resolving a visual problem could be evaluated as being quite a bit shorter than it had actually been, after the fact.

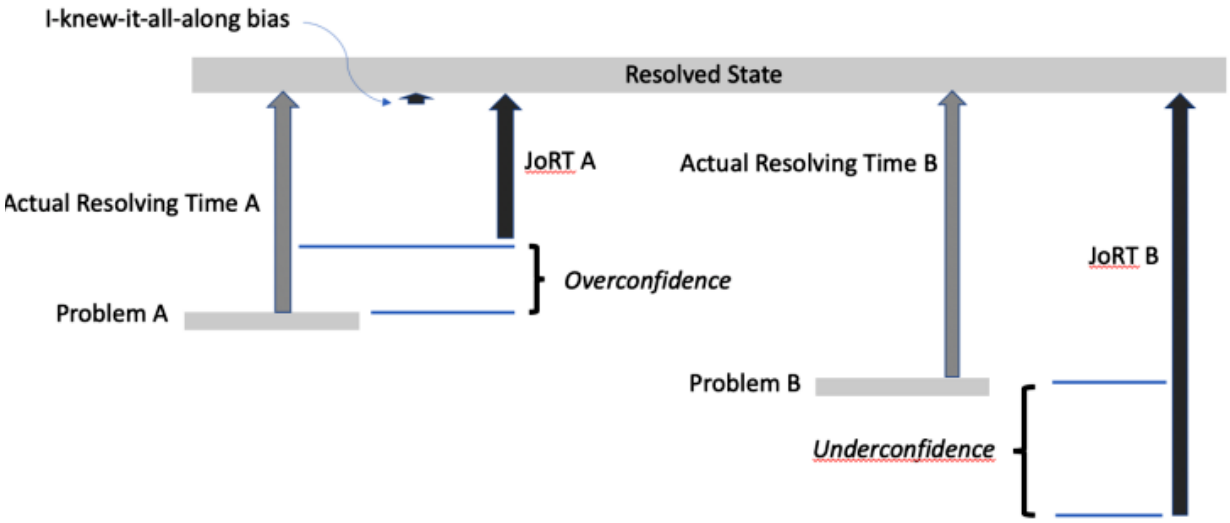

Figure 1. Caption: A Model comparing Judgments of Resolving Time (JoRTs) to Actual Resolving Time (ARTs) for two hypothetical problems, A and B. As can be seen, the ART for problem B was longer than the ART for problem A. The length of the gray arrows to the resolved state represents time taken. The JoRTs are represented by two additional (darker) arrows, and indicate the amount of time one could have, hypothetically, judged their resolving time to be. In these scenarios, the JoRT for problem A was *shorter* than ART for A; the JoRT for problem B was *longer* than the ART for B. These discrepancies can be used as a proxy for over- and underconfidence, respectively (Had the ART and JoRT been equal, this would be an indication of accurate metacognition). This framework provides a specific mechanism for the hindsight bias, or the *knew-it-all-along* bias. Here, the notion of "knowing it all along" is represented by the very short arrow, a hypothetical scenario where one would neglect all of one's prior resolving time – the most extreme example of overconfidence, and the current un-quantified definition of hindsight bias.

In **Figure 1**, we draw a hypothetical model of the actual time it had taken to resolve two problems, A and B. As can be seen, taking the simplest perspective possible, problem B took longer to resolve than it took to resolve problem A – the actual path from problem B to the resolved state is longer than that of problem A. After having knowledge of the resolution, the literature on the hindsight bias might predict that when someone has the illusion that they "knew it all along," then the time paths to the solution would be neglected. Thus, the hindsight bias might be indicated by a judgment where one believed that they had always been in a resolved state – represented in the figure by the short arrow with no time path. However, other than the assumptions stemming from connotations of "knowing it all along," there is no particular reason that the hindsight bias path need be "zero" in judged time. It could simply be a relatively shorter amount of time (than actual resolving time). To illustrate this, we include in the model the JoRT, which is shown as a judgment of time that can be either shorter or longer than (or equal to) the actual time it took to resolve a problem. In the figure,

the judgment of how long it took to resolve problem A was shorter than the actual amount of time it took to resolve problem A. On the flipside, the JoRT for problem B is shown to be longer than the actual time it took to resolve problem B. The novel proposal we put forth here is that the direction of the discrepancies can be used as a logical assessment for over- and underconfidence, respectively.

By using this type of retrospective judgment – the JoRT – we are able to target a particular range of time, and, therefore, a level of over- or underconfidence people might be exhibiting when it comes to the hindsight bias. More generally, we feel this may be a good measure of confidence in that it is a way to require people to travel back in time to the past, and see themselves, or the reality of their learning, more clearly. Indeed, recent work on hindsight bias has shown that narcissists show a larger bias than non-narcissists in that narcissists refuse to re-assess their past (Howes et al., 2020) -- suggesting that such a paradigm obliging individuals to re-assess their past might potentially be a way in which to mitigate the bias. Despite any possible mitigation, given the patterns found in the literature, we hypothesized that it would be more likely that we would find relatively shorter JoRTs, supporting the general overconfidence literature as well as the stubbornness of the hindsight bias. Overconfidence would also be in line with the phrase that has now become synonymous with the bias – knowing it all along – while at the same time providing additional data to allow us to understand the degree to which the bias occurs. With our hypothesis, we would also be adding to the metacognitive theories suggesting that people often discount, or *want to* discount, time, or sometimes, effort, that has been exerted in the past. We discuss the complexity of discriminating between time and effort below.

4. Estimations of time and effort

While the current study focuses on (retrospective) *time* estimation, one might look at "effort" as a way of understanding the current investigation between actual time spent and judged time spent on a task. In other words, the amount of effort exerted may not be "judged" to be equally valuable by the learner. Literature has already pointed to the reality that effort benefits learning. However, effort, or one's judgment of effort, doesn't necessarily translate into "feel good learning" for the individual. On the contrary, in many cases, learners are prone to believe that exertion of effort can mean a *lack of* learning (Hong et al., 1999). In one study, participants studied cue-target pairs. Then, in a second learning session, they either re-read the pairs again, or were tested on the pairs. While the latter required more effort, people believed that the former would lead to better performance on a later test. In the end, this disconnect between effort and learning was confirmed: People had performed worse after re-reading (Kornell & Son, 2009). The primary difference between re-reading and self-testing is one that seems to affect the judgment of effort, and perhaps time. Given that people mistakenly believe that more effort indicates a lack of learning, they could then easily believe that once learning had occurred, there probably wasn't that much effort involved in the first place. This could be one of the reasons for the hindsight bias, where people who have now gained knowledge correlate that knowledge with very little effort, or none at all (Agarwal et al., 2008) - consequently a feeling of "I knew it all along."

The same idea can be found in other studies related to "desirable difficulties" (Bjork & Bjork, 2020). People systematically provide higher judgments for learning conditions that were relatively less effortful than for conditions that were more effortful. In a massed-versus-spaced paradigm, acquisition was found to be slower for the spaced items than for the massed items (although at final test, the spacing effect still transpired – (Bahrick et al., 1993). Thus, people seem to have an illusion that because their current learning is slower for spaced than for massed items, the difference will hold for future memory performance. Researchers have shown evidence of this illusion of increased confidence during massed study (Zechmeister & Shaughnessy, 1980), and have also suggested that when items are massed, people might believe that encoding occurred on the initial presentation and further effort on the item is not needed (Jacoby, 1978). In short, people may exhibit an overconfidence effect on massed items based on the *judgment* of investment of effort or time.

Might time act in the same way that effort does? Maybe or maybe not. A long period of time may be allocated with very little effort, or attention, leading to no learning at all, while a very short period of

high effort may lead to strong performance later on. On the other hand, the definitions of "time" and "effort" have meant the same thing -- "duration" -- for various researchers and fields. In fact, Halkjelsvik and Jørgensen (2012) , with the aim of avoiding the issue, decided to use the term "performance time" for their meta-review on the estimates people make for various subsequent tasks, from seconds long to hours long. While they were looking at predictions of time duration, their analysis pointed to the idea that people are sometimes optimistic and, at other times, realistic, about how long a future task would take. While the mechanisms of when they were overconfident may be related to a number of variables, one finding seemed consistent: Researchers found what is known as the *decomposition effect* (Connolly & Dean, 1997) or the *segmentation effect* (Forsyth & Burt, 2008), where people seem to be more accurate (i.e. more realistic) about time estimation as long as they are able to "decompose" the whole project down to smaller units or sub-tasks. In other words, the shorter the task would take in general, the more likely they were to target (rather than underestimate, which might have been typically expected) the amount of time it would take to complete. The same shift has been found with tasks that are relatively simple, as compared to complex (Thomas et al., 2003).

Whether the above findings would hold for prior time estimates for already completed tasks is an open question. In some sense, while they may seem related, the time estimations can be qualitatively different. Predictive estimates are not based on direct information of the actual task; retrospective judgments have the potential to be based wholly in the actual event's time, and, thus, more accurate. Logically, then, the JoRTs that we record here, we hoped, by default, would be relatively accurate. In addition, any deviations from accuracy could be used as a measure of over- and under-confidence in how long it had taken to resolve a task. Thus, overall, our aim was to use a simple task (i.e. "already largely decomposed" task) where it would be easy for participants to make realistic estimations of time. At the same time, with a simple task, our hypothesis slightly wavered. At the outset we hypothesized that people's JoRTs would be shorter than the actual times it took to complete prior tasks -- a display of overconfidence. However, given that in the studies discussed above people made more realistic estimations of time (not as optimistic), we also were not to be surprised if the JoRTs did not exhibit as much overconfidence as first had been expected.

5. The Stroop Task

The longer-term goal of this study was to understand the consequences of having the illusion of believing in faster learning than what was real. We believe that these illusions are likely to occur for learning that could take some time, such as in the examples we mention above – say, language learning. When we think about the situations in which people are at risk of "giving up," complicated learning tasks are brought to mind – math problems, reading problems, etc. But to begin the examination of an individual's JoRTs, we here focused on a short, controlled task – the Stroop task (Stroop, 1935) -- similar to the visual search procedures used to test the "saw it all along" effect. The Stroop task is one that is generally consistent across adult individuals, and, given the very short time it takes to resolve each problem, we expected that people would have more of a chance to reach high levels of accuracy in their resolving time judgment.

Thus, in the current study, we used the classic paradigm, where words are presented in different colors, and the task is to report the color of the font, as quickly and as accurately as possible. The Stroop task is ideal additionally because we could also compare across distractor conditions, from most distracting or time consuming – when the color names mismatch with the color fonts – to the least distracting or time consuming – when the color names and fonts match. And we would be able to see if there are different levels of overconfidence when distraction comes into play, if at all. In other words, we would be able to ask further whether conditions that require more time would lead to a pattern of overconfidence while conditions that require less time would lead to a pattern of relative underconfidence. If it were the case that the time conditions did not matter, in other words, if participants JoRTs, on average, were all lower than actual resolving time across the board, then we could be fairly confident in saying that people were rather overconfident in how quick they were to complete the task.

6. Confidence across culture

A few studies have suggested that time estimations

may be more accurate when using simpler or shorter tasks, like the Stroop task. However, those studies were different in that the time estimations were made prospectively, rather than retrospectively. The judgments of interest here were retrospective, and given the stubborn literature regarding overconfidence and the hindsight bias, our hypothesis tended to remain leaning towards overconfidence. That is, we expected that people's JoRTs, on average, would be shorter than the actual time it took to complete prior tasks (their ARTs). However, as a secondary question, we were also interested to see if there might be potential differences across culture. As previous studies had found varying perceptions of effort in the East and the West (Markus & Kitayama, 1991), we thought the same might be true for time estimations. Specifically, while American students tend to feel more confident with easy learning (Heine et al., 2000), Asian students tend to view ability and effort as being positively related (Stevenson & Stigler, 1994). Indeed, data have shown that East Asian populations are far less overconfident than Westerners, and sometimes even demonstrate striking underconfidence or self-criticism (Kitayama et al., 1997).

Given our primary interests, we decided to conduct our study separately for two sets of participants: Those living in the US and those living in Korea. However, because of the plethora of possible environmental factors that might explain any differences in the time judgments, we did not want to overstate any conclusions regarding the differences. Thus, as a first pass, we thought it would be helpful to simply run the study in the two cultures, and see whether there were any strong patterns. Specifically, we expected that for the American individuals, we would find an overconfidence effect – where JoRTs overall would be faster than ARTs. For the Korean participants, we left our hypothesis to be an open one, keeping in mind that they might be less overconfident to a degree, perhaps even metacognitive accuracy with their JoRTs.

7. The Current Study

In summary, the difference between people's JoRTs, on average, and their ARTs, on average, on a set of prior tasks were used as an indication for over- or underconfidence. Given much of the literature on confidence, and on hindsight bias, we hypothesized that, overall, people would show overconfidence. That is, we expected that people, in retrospect, would report JoRTs to be shorter than their ARTs (illustrated by problem A in **Figure 1**). We were also interested in a few secondary analyses. First, we thought there might be a difference in the types of Stroop trial that were presented: The most time-consuming trials (where the color names were incongruent with the color ink) might lead to a higher level of overconfidence than the least time-consuming trials (where the color names were congruent with the color ink). We were also curious about whether there would be similar differences across culture. In the first experiment, we tested individuals living in the US, and in the second experiment, we tested individuals living in South Korea. Using the classic Stroop task, we examined people's JoRTs and compared those judgments to their ARTs. While we expected to find overconfidence in the US participants -- especially in the trials that were most time consuming -- we had an open hypothesis as to whether the Korean participants would show a similar degree of overconfidence.

8. Experiment 1

Prior to beginning the study, procedures for both experiments were subject to review by the institutions' IRB committees from where the participants were recruited and tested. The full experimental protocol was approved by the 4-year College IRB Committee (approval #2122-0530-054). All methods were performed in accordance with the institution's guidelines and regulation. We confirm that prior to beginning the experiments, informed consent was obtained in writing from all participants. There were no participants under the age of 18.

9. Methods

Participants

Seventy-five students attending college in the US participated in the experiment for course credit. We conducted a sensitivity power analysis to examine whether our results would provide enough power. The sensitivity analysis indicated that, with the sample size of 75, the effect size (f) was expected to be $\geqq 0.14$, or $\eta 2=0.27$.

Design and Procedure

The procedure was conducted in the laboratory, where all participants arrived and filled out consent

forms before beginning the experiment. The entire experiment was presented on a computer monitor, with a typical grayish black background. All stimuli were presented by the E-prime 3.0 version, and responses were recorded using Chronos. Three versions of the classic Stroop task were used: Congruent, incongruent, and neutral. In the congruent version, the color name (either *blue*, *red*, *green*, or *yellow*) and the ink matched (e.g. the word *blue* was presented on the screen in blue ink). In the incongruent version, the color name and the ink mismatched (e.g. the word *blue* was presented on the screen in one of the 3 other ink colors, red, green, or yellow). In the neutral version, a non-color word (e.g. *dictionary*) was presented in one of 4 colors: blue, red, green, or yellow.

After signing consent forms, participants were given instructions on the Stroop task, where they were told clearly that the goal was to respond with the color that matched the ink color of the word presented on the screen. Before beginning the experiment, they were presented with practice trials where each of four colors—red, yellow, blue, and green – were presented on the screen. Once they understood the task and completed the practice trials, the main experiment began.

There were 36 trials in all, 12 from each of the congruent, incongruent, and neutral versions, blocked and presented in a crossover design. For the analyses, we were able to drop Order as there were no differences for that variable. We included a fixation stimulus for 500ms before each trial as a way to orient participant's attention. For each trial, there was a time limit of 2000ms. For all versions of the task, participants had to respond, as quickly and as accurately as possible, with the color of the ink on a button box. Actual resolving times (ARTs) were recorded, which indicated the time starting from the moment the word appeared on the screen up until the moment a color button was pressed. Button responses were made by pressing one of 4 buttons, labelled with 4 color stickers: blue, red, green, and yellow).

After every 3 trials, JoRTs were recorded. Participants were instructed to make a judgment of how long, on average over the 3 previous trials, it took them to respond to the color button after the word appeared on the screen. They made their JoRTs by typing in the number of milliseconds using a keyboard prompt. We acknowledge that having participants make average JoRTs across 3 trials would not allow us to measure precise judgments, but we wanted to avoid any distractions that might arise from having to switch back and forth between tasks on each trial.

Our main interest was to see how accurate (or how inaccurate) their JoRTs were overall. We were also curious to see whether the amount of time on task -- which varied across condition type (congruent, incongruent, neutral) would change the accuracy of people's JoRTs. Thus, we primarily calculated whether JoRTs were shorter or longer than their ARTs across trial condition. As a result, this study consisted of a (Stroop: Congruent vs. incongruent vs. neutral) within-subjects design, with the difference between ARTs and JoRTs as the main dependent measure. However, we also were curious about our secondary question, and, therefore, conducted additional analyses regarding culture.

10. Experiment 2

One hundred sixty-four students attending a 4-year college in South Korea participated for course credit. We conducted a sensitivity analysis for Experiment 2. The expected effect size (f) with a sample size of 164 was ≥ 0.1, which is equivalent to $\eta 2=0.42$. Our effect size was greater than the expected effect size. All participants received course credit for their participation. The design and procedure for Experiment 2 were exactly the same as the procedure for Experiment 1, except that the instructions were given in Korean, and the procedure was carried out in a later semester.

11. Results

Experiment 1 – US participants

The data-trimming procedure (La Heij et al., 2001) was used, as this procedure had been used typically for Stroop task trials, where we omitted times faster than 99ms for ARTs and JoRTs as well as trials with a correct rate below 80%, resulting, here, a remaining 95.28% of the data. See Table 1 for descriptive statistics for ARTs and JoRTs across the congruent, incongruent, and neutral conditions in the Stroop task. We also explored what types of judgment people made. For instance, it would be worrisome if participants only used very specific round numbers, such as 1 second or

500ms. As can be seen in the **Figure 2**, the histograms show that for the different conditions, participants did not tend to use only specific whole numbers when making the judgments.

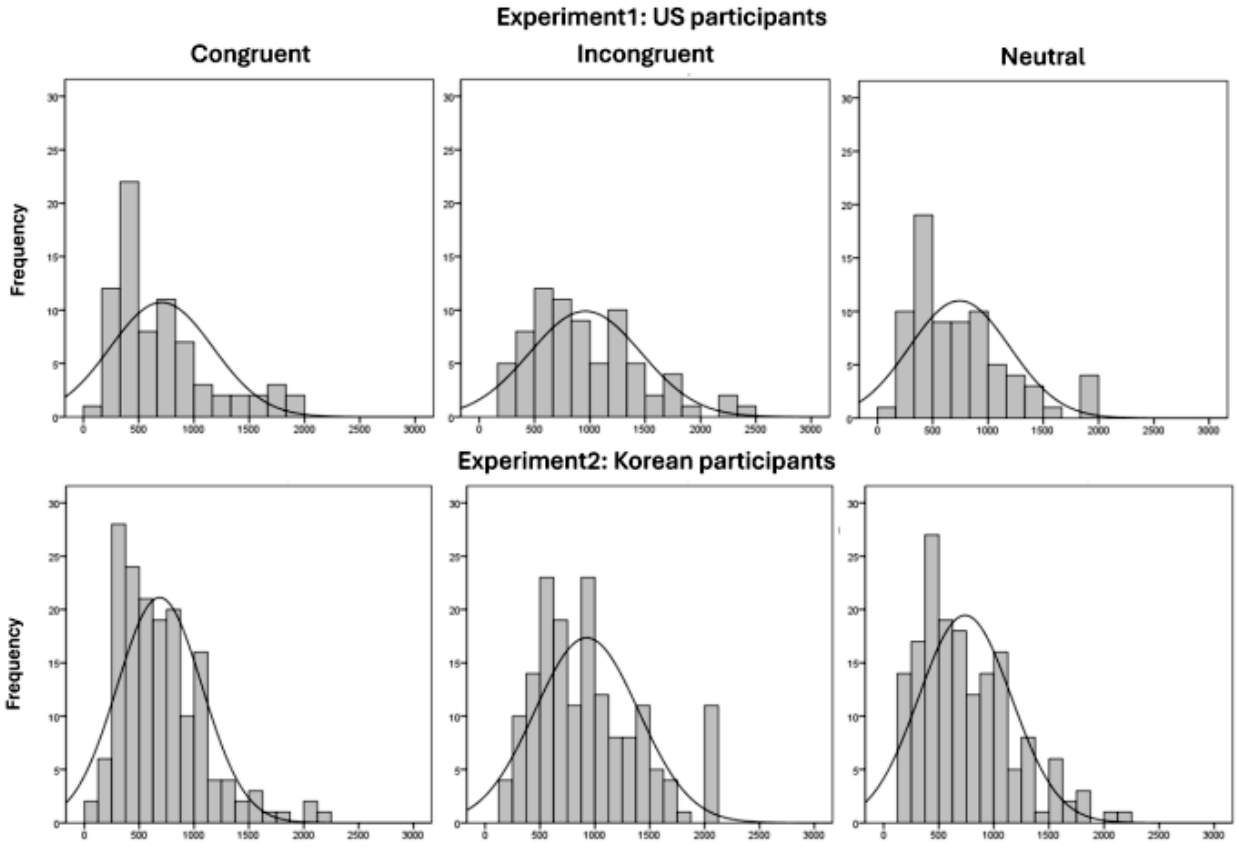

Figure 2. Caption: Histograms for making the JoRTs, separated by the Congruent (left panels), Incongruent (middle panels), and Neutral (right panels) conditions. The top row represents the data from Experiment 1 (US participants) and the bottom row represents the data from Experiment 2 (Korean participants).

First, we found that participants spent the longest time when resolving the incongruent trials as compared to either of the other condition trials: Incongruent vs. congruent ($M_{difference}$= 174.70, *SD*=17.39, *p*<.001)and incongruent vs. neutral ($M_{difference}$= 151.94, *SD*=15.03, *p*<.001). There was no difference between congruent and neutral trials ($M_{difference}$= -22.75, *SD*=9.83, *p*=.07). Somewhat as expected, participants gave longer estimates on the incongruent trials: Incongruent vs. congruent ($M_{difference}$= 252.30, *SD*=41.63, *p*<.001) and incongruent vs. neutral ($M_{difference}$= 216.36, *SD*=42.78, *p*<.001) trials led to differences, while there were no differences between congruent and neutral trials ($M_{difference}$= -35.92, *SD*=31.67, *p*=.78). These results were basically as expected given the nature of the Stroop task.

Most importantly, we found that people's JoRTs were numerically longer than their ARTs -- see **Table 1** for mean RTs scores across condition. As can be seen in the Table, the difference scores, as calculated by subtracting JoRTs from ARTs, were all negative, in the direction of under-, not overconfidence. We conducted one-sample t-tests (test value = 0) using the difference scores (Also see **Table 1** for the mean difference scores). This was a simple way in which to see if people's judgments were different than the actual time spent resolving the Stroop trials. Results showed that for all three conditions, participants consistently estimated their resolving times -- their JoRTs -- to be longer than the actual amount of time it took to solve the task -- their ARTs: Neutral (*t*=2.99, *df*=74 *p*<.01), incongruent (*t*=2.74, *df*=74 *p*<.01), congruent (*t*=3.93, *df*=74 *p*<.001). Overall, regardless of the Stroop condition, JoRTs were consistently longer than ARTs. This finding went against our hypothesis.

Table 1. Descriptive ART and JoRT data for the congruent, incongruent, and neutral conditions on the Stroop task (Experiment 1, US participants). Negative difference scores represent underconfidence.

N= 75	Type	M	SD
Congruent	ART	560.23	103.15
	JoRT	707.10	465.37
	M difference	-146.87**	53.53
Incongruent	ART	734.92	158.76
	JoRT	959.39	504.12
	M difference	-224.47***	57.06
Neutral	ART	582.98	85.63
	JoRT	743.02	453.45
	M difference	-160.05**	53.57

* $p<.05$, ** $p<.01$, *** $p<.001$

Finally, to confirm any differences between Condition (Congruent, Incongruent, Neutral) and the difference scores -- between JoRTs and ARTs, we conducted a repeated measure ANOVA. While not quite reaching statistical significance ($F(2,148)=2.814$) $p=.06$, $\eta^2=.04$), results showed that there was a difference between the congruent condition ($M_{difference}$=145.87, SD=463.55) and the incongruent condition ($M_{difference}$=224.46, SD=494.19), $t(74)=-2.1$, $p<.05$). This might provide a hint -- in terms of numerical patterns -- that if a task was more time consuming, people had a greater tendency to acknowledge that time.

Overall, our primary result from Experiment 1 went directly against the hypothesis we outlined at the outset of the study. These findings also went against some of the literature that has shown a systematic overconfidence effect, particularly when looking at hindsight bias. Those data have had researchers conclude that after gaining new knowledge, people erroneously judge that they had "resolved [or knew the solution to] that task all along." Using JoRT as a measure for confidence, the data here suggest the opposite. Not only did participants fail to exhibit that they had "known it all along," but they presented the opposite. Their judgments indicated that they thought they had come up with the resolutions more slowly than they had actually come up with the resolutions. Results from Experiment 2, testing a different population, are presented below.

Experiment 2 – Korean participants

Subsequent to the same data-trimming procedure (La Heij et al., 2001) that had been used in Experiment 1, where we omitted times faster than 99ms for ARTs and JoRTs as well as trials with a correct rate below 80%, we were able to conduct analyses on the remaining 96.23% of the data. The general results of Experiment 2 are presented in **Table 2**. As can be seen, the numerical patterns we obtained were similar to the ones that were found in Experiment 1. First, we found that, like with the American participants, the Korean participants also spent the longest time when resolving the incongruent trials than either of the other condition trials: Incongruent vs. congruent ($M_{difference}$= 183.70, SD=10.74, $p<.001$) and incongruent vs. neutral ($M_{difference}$= 149.72, SD=10.49, $p<.001$). And again, mimicking the pattern above, participants judged that they took longest on the incongruent trials: Incongruent vs. congruent ($M_{difference}$= 240.17, SD=25.90, $p<.001$) as well as incongruent vs. neutral trials ($M_{difference}$= 187.62, SD=24.62, $p<.001$) differed from one other.

Table 2. Descriptive ART and JoRT data for the congruent, incongruent, and neutral conditions on the Stroop task (Experiment 2, Korean participants). Negative difference scores represent underconfidence.

N=164	Type	M	SD
Congruent	ART	559.47	99.90
	JoRT	685.80	387.14
	M difference	-126.33***	31.29
Incongruent	ART	743.16	154.78
	JoRT	925.97	471.48
	M difference	-182.81***	38.29
Neutral	ART	593.44	95.76
	JoRT	738.35	420.34
	M difference	-144.91***	33.23

* $p<.05$, ** $p<.01$, *** $p<.001$

The mean JoRTs and ARTs are presented in **Table 2** for each of the conditions. As before, the mean differences turned out all to be negative, again going against our hypothesis that people's JoRTs would be shorter than people's ARTs. And as in the analysis for Experiment 1, we asked whether people's difference scores (between ARTs and JoRTs) were significantly different from zero. Using difference scores between ARTs and JoRTs, we proceeded to conduct one-sample t-tests (test value = 0). In support of the data above, we found that the congruent (t=4.04, df=163 $p<.001$),

incongruent (t=4.78, df=163 p<.001), and neutral (t=4.36, df=163 p<.001), difference scores were all significantly different from zero, suggesting a mis-targeting of JoRTs in the direction of underconfidence.

We also conducted an ANOVA on Condition to see any effect on differences between JoRTs and ARTs, which resulted in a significant main effect (F(2,326)=3.38, p<.05, η^2=.02). Paired t-tested showed significant differences between congruent trials ($M_{difference}$=126.33, SD=400.66) and incongruent trials ($M_{difference}$=182.8, SD=490.28) (t(163)=-2.37, p<.05), strengthening the numerical results of Experiment 1, where a more time-consuming task was less likely ignored, and, on the contrary, perhaps amplified.

Overall, the data from Experiment 2 replicated the data from Experiment 1 in that people were generally underconfident, going against the original hypothesis regarding cultural differences. We did expect a bit less overconfidence for the Korean participants, but, with the current data itself, did not found anything telling. However, interestingly, comparing across Experiment 1 and Experiment 2, we actually found, numerically, a smaller bias for the Korean participants than for the US participants. This went somewhat against our expectation - that Korean individuals might be more underconfident than American individuals. To determine if we could find anything statistically, we tried two analyses. One was to conduct a mixed ANOVA with difference scores from the three conditions (congruent, incongruent, neutral-within subject variable) and group (US vs Korean). When we did so, there was no interaction between difference score and group (F(2,474)=0.241, p=.786). The second method was to compare the difference scores between congruent and incongruent conditions (the two interesting conditions). However, similar to the previous method, there was no interaction effect (F(1,237)=0.24, p=.625). Thus, regarding the question of culture, a more direct procedure should be explored in the future, with a sufficient sample size.

12. Discussion

Overall, and surprisingly, our results suggested underconfidence, not overconfidence. That is, when referring to **Figure 1**, participants' JoRTs were systematically *longer* than their ARTs – people thought that they had taken more time to resolve the Stroop problems than they truly had taken. We also found that this result occurred in two different cultures. In Experiment 1, we found that participants attending college in the US displayed consistent underconfidence; In Experiment 2, we found that those attending a college in Korea did so as well. The lack of difference across culture was a bit surprising given that, in the past, there had been evidence of a divergence in confidence across cultures of the East and West (Kitayama et al., 2004). However, we acknowledge the limitations in our participant pool size, and that further investigation where we look at culture as a within-experiment variable with sufficient n would be required to make any strong conclusions here. Based on the literature exhibiting systematic and stubborn hindsight bias (Son et al., 2021), we had thought that, in the least, the US participants would show some amount of overconfidence. That is, we expected that after having resolved the Stroop problems, individuals would have thought that "they had known it all along," or more accurately, "to some particular degree faster."

We did find, as expected, that both the US and Korean participants took longer to resolve the Stroop problems in the incongruent condition. If a color word was written in a different color ink, the interference of the color word from reading was, as expected, unavoidable. However, the main question in the research – how people *judged* the speed in which they were able to utter the color of the ink – pointed to underconfidence. Even when the actual solving times were relatively longer, their judgments of those times were inflated even beyond that, suggesting, in line with our interpretation, that people were not confident in their resolving ability. On the contrary, they felt consistently slower.

One issue that needed some mulling over was the notion that people simply aren't able to estimate such short time periods. The difference between the ARTs and the JoRTs ranged from a period of approximately 125ms to 224ms, with the actual JoRTs starting from 680ms to 960ms. People's time estimations are never perfect, but this range across the fastest estimations – for the congruent trials – to the slowest estimations – the incongruent trials – suggest that there was some ability to calculate their ability to resolve even tasks as quick as the Stroop task. As mentioned in the introduction, there remains to be conclusive evidence

as to whether the "underconfident" JoRTs were due to extreme *decomposition* of short processes (Connolly & Dean, 1997; Forsyth & Burt, 2008), or whether people felt that the Stroop task took more time to resolve than it actually did. On the other hand, given the quick response times, another limitation in our study was that we sacrificed precision, in that participants were asked to make JoRTs after every 3 trials, rather than every trial. In future studies, particularly for materials that take longer to learn, it would be good to seek a replication of the current data where JoRTs are made for each learning session.

Interestingly, in both Experiment 1 and Experiment 2, we did find that what we are calling the underconfidence effect -- as measured by the difference between ARTs and JoRT -- turned out the be exacerbated in the incongruent conditions as compared to the congruent conditions. This finding allows us to think that the cues and clues that participants are using to make the JoRT is not only a process of "direct access" of putting time into the prior task (considering the older metacognitive literature on *direct access* vs *cue-driven* mechanisms, e.g. Koriat, 1993; Metcalfe et al., 1993), but rather, a process that includes an assessment of the external cue. In this case, a cue-driven rule-of thumb, something like "when the stimulus word is incongruent to the ink color, I was probably fairly slow in responding" may be affecting the level of the JoRT. This type of rule-of-thumb may appear to be strategy of alleviating the hindsight bias, but, on the flipside, it may be a feigned alleviation, akin to covering a wound with a band-aid. In other words, if the accuracy of the JoRT were related to directly returning to, and recalling, the past time that was exerted on a task, or, a set of tasks, then we could be more confident that the bias could be avoided using the simple strategy of having people make a JoRT. One prior study (Son et al., 2021), in fact, concluded that the hindsight bias might be made to disappear when participants are required to put themselves in the shoes of a naive child, but data pointed more strongly to the idea that people were using a rule-of-thumb strategy -- "kids are not as smart as me, so what I (now) know probably will take them a very long time to learn." Further research on the notion of whether the hindsight bias can be alleviated by various "rule-of-thumb" strategies is warranted and would be an important research paradigm. In the least, the model of hindsight bias, or the "I knew it all along effect," depicted in **Figure 1** here with the short arrow, is too simple.

While still an early question, the current data allow us to be optimistic about using simple methods that might be useful for counteracting the hindsight bias. Not much has been talked about in terms of the degree to which the people are myopic in hindsight. Perhaps when people make a general hindsight judgment typically found in the literature, they are not really thinking about the time that it took with any specificity. Perhaps, as written in the introduction above, people are simply avoiding going back to a time when they felt naïve, when they had to exert a whole lot of time, or even worse, when time was still filled with errors. It may be that thinking about past learning, especially the detours, is rather aversive, and, therefore, learners will not re-visit it, especially when not required. The current data required people to consciously think about (at least some of) the time it took to resolve a series of prior problems. And seeing the surprising result, we might conclude that this requirement – of having individuals make the JoRT – was enough to keep people from falling into the hindsight bias.

When we began this experiment, our thought was that an obstacle to persistence on some current task may be because people would underestimate the time it took to resolve similar tasks previously. In order to approach this problem, our method was to uncover the degree to which people erred in the JoRTs. The simple mechanism was that if people consistently gave shorter JoRTs (as compared to their ARTs), then the metacognitive mis-judgment would spark further decisions that might be harmful to persistence. For instance, as in the examples described in the introduction, JoRTs that miss the mark on the shorter side would allow the individual to believe, erroneously, that the subsequent task is taking too long. What we found was the opposite, and what we might even call a possible anti-dote to hindsight bias (with the acknowledgment that underconfidence is a metacognitive error as well) -- referring to the idea that obligatory judgments may help alleviate the bias, or even make it vanish, as they did here. Indeed, the notion of requiring people to travel back (mentally) to previous learning sessions seems to be gaining traction in the literature on hindsight bias.

Ackerman and colleagues (2020), for instance, had participants make judgments about prior answer as well as prior confidence judgments, and they found that, interestingly, people were able to update their current judgments when they had learned the materials successfully, as compared to when they did not learn the material successfully. These results seem in line with the results in the current study, where having people explicitly think about their past learning improves current judgment accuracy (see also Groß et al., 2023, for explicit bias countering strategies).

The question of how long it takes to resolve a problem, in relation to one's actual resolving time, is a crucial one. More and more, learners are at risk of believing that they must be quick and errorless when it comes to learning, and one wonders if this belief is exacerbated by the illusion of "knowing it all along." In the data presented here, we found, perhaps on a more positive note, that there is no evidence of the illusion of being "a natural." When explicitly asked to make JoRTs, participants in two cultures not only knew that they took time to resolve the Stroop trials, but also overestimated that time. We interpret the data as supportive in countering the hindsight bias, and look forward to examining the consequences, good or bad, of the unexpected underconfidence we found. In the least, here, we have developed a new metacognitive judgment that seems to encourage individuals to think explicitly about their past time investments, allowing them to acknowledge that there was, indeed, a learning process.

Data Availability Statement

The datasets generated during and/or analyzed during the current study are available in the data repository under the same name as the title of this manuscript, https://osf.io/28vb4/.

Competing interest statement

The authors declare no competing interests.

References

[1] Ackerman, R., Bernstein, D. M., & Kumar, R. (2020). Metacognitive hindsight bias. *Memory & Cognition*, 48, 731-744.

[2] Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger III, H. L., & McDermott, K. B. (2008). Examining the testing effect with open and closed book tests. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22(7), 861-876.

[3] Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, 4(5), 316-321.

[4] Bernstein, D. M., Atance, C., Loftus, G. R., & Meltzoff, A. (2004). We saw it all along: Visual hindsight bias in children and adults. *Psychological Science*, 15(4), 264-267.

[5] Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4), 475.

[6] Connolly, T., & Dean, D. (1997). Decomposed versus holistic estimates of effort required for software writing tasks. *Management Science*, 43(7), 1029-1045.

[7] Craik, F. I., & Simon, E. (1980). Age differences in memory: The roles of attention and depth of processing. *New directions in memory and aging*, 95-112.

[8] Fiedler, K., Ackerman, R., & Scarampi, C. (2019). Metacognition: Monitoring and controlling one's own knowledge, reasoning and decisions. *The psychology of human thought: An introduction*, 89-111.

[9] Forsyth, D. K., & Burt, C. D. (2008). Allocating time to future tasks: The effect of task segmentation on planning fallacy bias. *Memory & cognition*, 36, 791-798.

[10] Greene, C. M., Levine, L. J., Loftus, E. F., & Murphy, G. (2023). Just as I expected? Hindsight bias for the outcome of a national referendum is moderated by outcome valence and surprise. *Applied Cognitive Psychology*, 37(5), 1016-1026.

[11] Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on psychological science*, 7(5), 464-481.

[12] Halkjelsvik, T., & Jørgensen, M. (2012). From origami to software development: A review of studies on judgment-based predictions of performance time. *Psychological bulletin*, 138(2),

238.

[13] Harley, E. M., Carlsen, K. A., & Loftus, G. R. (2004). The "saw-it-all-along" effect: demonstrations of visual hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(5), 960.

[14] Heine, S. J., Takata, T., & Lehman, D. R. (2000). Beyond self-presentation: Evidence for self-criticism among Japanese. *Personality and Social Psychology Bulletin*, 26(1), 71-78.

[15] Hom Jr, H. L. (2023). Perspective-taking and hindsight bias: When the target is oneself and/or a peer. *Current Psychology*, 42(16), 13987-13998.

[16] Hong, Y.-y., Chiu, C.-y., Dweck, C. S., Lin, D. M.-S., & Wan, W. (1999). Implicit theories, attributions, and coping: a meaning system approach. *Journal of Personality and Social psychology*, 77(3), 588.

[17] Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of verbal learning and verbal behavior*, 17(6), 649-667.

[18] Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory*, 17(4), 471-479.

[19] Kitayama, S., Markus, H. R., Matsumoto, H., & Norasakkunkit, V. (1997). Individual and collective processes in the construction of the self: self-enhancement in the United States and self-criticism in Japan. *Journal of Personality and Social psychology*, 72(6), 1245.

[20] Kitayama, S., Snibbe, A. C., Markus, H. R., & Suzuki, T. (2004). Is there any "free" choice? Self and dissonance in two cultures. *Psychological Science*, 15(8), 527-533.

[21] Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6(2), 107.

[22] Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 23(9), 1297-1317.

[23] Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, 17(5), 493-501.

[24] La Heij, W., Van der Heijden, A., & Plooij, P. (2001). A paradoxical exposure-duration effect in the Stroop task: Temporal segregation between stimulus attributes facilitates selection. *Journal of Experimental Psychology: Human Perception and Performance*, 27(3), 622.

[25] Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological review*, 98(2), 224.

[26] Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, 15(1), 174-179.

[27] Musch, J., & Wagner, T. (2007). Did everybody know it all along? A review of individual differences in hindsight bias. *Social cognition*, 25(1), 64-82.

[28] OnlineFischhoff, B. (1975). Hindsight≠ Foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 288-299.

[29] Putnam, A. L., Deng, W., & DeSoto, K. A. (2022). Confidence ratings are better predictors of future performance than delayed judgments of learning. *Memory*, 30(5), 537-553.

[30] Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on psychological science*, 7(5), 411-426.

[31] Schill, H. M., Gray, S. M., & Brady, T. F. (2023). Visual hindsight bias for abnormal mammograms in radiologists. *Journal of Medical Imaging*, 10(S1), S11910-S11910.

[32] Son, L. K., Hong, S. S., Han, L., Lee, Y., & Kim, T. H. (2021). Taking a naïve other's perspective to debias the hindsight bias: Did it backfire? *New Ideas in Psychology*, 62, 100867.

[33] Stevenson, H., & Stigler, J. W. (1994). *Learning gap: Why our schools are failing and what we can learn from Japanese and Chinese educ*. Simon and Schuster.

[34] Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6), 643.

[35] Thomas, K. E., Newstead, S. E., & Handley, S. J.

(2003). Exploring the time prediction process: The effects of task experience and complexity on prediction accuracy. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 17(6), 655-673.

[36] Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of memory and language*, 64(2), 109-118.

[37] Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, 15(1), 41-44.