# Good Learners Are Poor Monitors: A Negative Relation Between Learning Ability and Monitoring Accuracy

Mengqi Hu[1], Wenbo Zhao[2], Anran Li[1], David R. Shanks[3], Yadi Yu[1], Xiaofang Tian[4,5], Muyi Liu[6], Xiao Hu[6, 7], Liang Luo[1,8], Chunliang Yang[1, 7]

[1] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China.

[2] School of Sociology, Beijing Normal University, Beijing, China.

[3] Division of Psychology and Language Sciences, University College London, London, UK.

[4] College of Education for the Future, Beijing Normal University at Zhuhai, Zhuhai, China.

[5] Department of Basic Courses, Tianjin Vocational Institute, Tianjin, China

[6] Faculty of Psychology, Beijing Normal University, Beijing, China.

[7] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Beijing Normal University, Beijing, China.

[8] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China.

## Author Note

Correspondence concerning this article should be addressed to Chunliang Yang (chunliang.yang@bnu.edu.cn), Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, 19 Xinjiekouwai Street, Haidian District, Beijing 100875, China.

**Abstract**

Effective learning involves not only the ability to quickly acquire knowledge and skills, but also the capacity to accurately monitor one's ongoing learning progress. The present research probed the relation between learning ability and monitoring accuracy. A meta-analysis (Study 1, $N = 2{,}406$) counterintuitively found that individuals with superior learning ability exhibited slightly poorer monitoring accuracy (measured as the resolution of judgments of learning). Study 2 re-analyzed the meta-analysis data and observed that expert learners remembered more items they erroneously believed they would not remember, and this underconfidence in expert learners led to a negative association between learning ability and monitoring accuracy. Studies 3 ($N = 102$, adults aged 18-23) and 4 ($N = 481$, adults aged 18-59) conceptually replicated the findings of Studies 1 and 2 in controlled experiments. These findings challenge the conventional wisdom that good learners are also good monitors, suggesting instead that expert learners are actually the ones with monitoring deficits.

***Keywords:*** Learning ability; Monitoring accuracy; Judgments of learning; Expert underconfidence; Meta-analysis

## Research Transparency Statement

### General Disclosures

**Conflicts of interest**: All authors declare no conflicts of interest. **Funding**: This research was supported by Beijing Philosophy and Social Science Foundation (24DTR067). **Artificial intelligence**: No artificial intelligence assisted technologies were used in this research or the creation of this article. **Ethics**: This research received ethics approval from Faculty of Psychology, Beijing Normal University (Protocol Number: BNU202112300096). **Open Science Framework (OSF)**: To facilitate long-term preservation, all OSF files have been registered at https://osf.io/f7b4r.

### Study One Disclosures

**Preregistration**: No aspects of the study were preregistered. **Materials**: No materials were used in this study. **Data**: All primary data are publicly available (https://osf.io/f7b4r/files/osfstorage). **Analysis scripts**: All analysis scripts are publicly available (https://osf.io/q8hbz). **Computational reproducibility**: The computational reproducibility of the results in the main article (but not the supplementary materials) has been independently confirmed by the journal's STAR team.

### Study Two Disclosures

**Preregistration**: No aspects of the study were preregistered. **Materials**: No materials were used in this study. **Data**: All primary data are publicly available (https://osf.io/f7b4r/files/osfstorage). **Analysis scripts**: All analysis scripts are publicly available (https://osf.io/7w6p4). **Computational reproducibility**: The computational reproducibility of the results in the main article (but not the supplementary materials) has been independently confirmed by the journal's STAR team.

### Study Three Disclosures

**Preregistration**: No aspects of the study were preregistered. **Materials**: All study materials are publicly available (https://osf.io/f7b4r/files/osfstorage). **Data**: All primary data are publicly available (https://osf.io/f7b4r/files/osfstorage). **Analysis scripts**: All analysis scripts are publicly available (https://osf.io/qzfu8). **Computational reproducibility**: The computational reproducibility of the results in the main article (but not the supplementary materials) has been independently confirmed by the journal's STAR team.

**Study Four Disclosures**

**Preregistration**: The research aims/hypotheses, methods (including sample size, inclusion/exclusion criteria), and analysis plan for Study 4 were fully preregistered prior to data collection at the Open Science Framework (https://osf.io/kwbzm). There were minor deviations from the preregistration (we recruited 512 participants instead of the planned 500, and excluded 31 participants for constant JOLs or extreme recall patterns). **Materials**: All study materials are publicly available (https://osf.io/f7b4r/files/osfstorage). **Data**: All primary data are publicly available (https://osf.io/f7b4r/files/osfstorage). **Analysis scripts**: All analysis scripts are publicly available (https://osf.io/ard9g). **Computational reproducibility**: The computational reproducibility of the results in the main article (but not the supplementary materials) has been independently confirmed by the journal's STAR team.

## Statement of Relevance

Successful learning requires not only quickly acquiring knowledge but also accurately monitoring one's own learning progress. Contrary to conventional wisdom, our results reveal that expert learners often struggle to accurately monitor their learning progress, because they typically underestimate their mastery of challenging material. Although teachers, parents, and educational policymakers often operate under the assumption that high-performing students require less support in learning how to learn, our findings instead show that high-performing students are less accurate at evaluating the strength of their learning.

Learning ability—the capacity to acquire knowledge and skills—is widely recognized as a cornerstone of academic achievement and personal growth (Argote, 2013). However, effective learning is not solely about quickly acquiring new information—it also requires the ability to accurately monitor and effectively regulate one's own learning process (Bjork et al., 2013). The capacity for self-assessment, often termed "monitoring ability," enables learners to accurately reflect on their learning progress and adjust their study activities accordingly (Nelson, 1990). Together, learning and monitoring abilities form the backbone of successful self-regulated learning (Zimmerman, 2002), which is increasingly important in a rapidly evolving, information-rich world.

Given the crucial roles of learning and monitoring abilities in successful learning, many studies have investigated the relation between the two since Flavell (1976) first introduced the concept of metacognition (e.g., Brown et al., 1983; Flavell, 1981). It is commonly believed that individuals with stronger cognitive resources—such as memory and attention—are better equipped to monitor their learning progress, resulting in a positive relation between learning and monitoring abilities (Griffin et al., 2008). The expertise-superiority hypothesis further posits that high-ability learners are better at focusing on critical information during encoding, enabling them to make more accurate judgments of learning (JOLs; i.e., subjective estimates about the likelihood of remembering studied information on a future test; Nietfeld & Schraw, 2002).

Supporting this view, many early studies documented a positive relation between absolute JOL accuracy (measured as the discrepancy between JOLs and test performance) and learning ability (measured as test performance) in a variety of learning tasks (e.g., Flavell, 1981). However, it is well-known that absolute JOL accuracy (= $|\text{JOLs} - \text{test performance}|$) is inherently influenced by test performance itself, creating a spurious positive relation between these two variables (Gignac & Zajenkowski, 2020; Nelson, 1984). In the Supplemental Materials (SM), we provide a data simulation to illustrate this spurious positive relation. Consequently, findings based on absolute JOL accuracy cannot inform us about the true relation between learning ability and monitoring capacity (Hasselhorn & Hager, 1989).

To address this issue, recent research has shifted toward investigating relative JOL

accuracy, measured as the inter-item correlation between JOLs and answer accuracy (0 = incorrect; 1 = correct) across trials within an individual (e.g., Hartwig et al., 2012). Relative JOL accuracy, also known as JOL resolution, represents the extent to which a person can accurately distinguish well-learned from poorly learned items, providing a more nuanced measure of monitoring capacity (Dunlosky & Metcalfe, 2009). For the sake of brevity, hereafter we refer to relative JOL accuracy as "JOL accuracy."

Recent studies primarily explored the relation between JOL accuracy and test performance in recognition-based learning tasks, and consistently detected a positive relation between these two measures (e.g., Hartwig et al., 2012; Smith & Was, 2019), leading to the conclusion that "good learners are also good monitors" (Touron et al., 2010). However, this conclusion is clouded by the nature of recognition tests (e.g., multiple-choice tests). Specifically, in recognition-based learning tasks, participants may accurately realize that some challenging items are non-memorable and provide low JOLs to these items, but then correctly guess these items in the final recognition test, thereby leading to an underestimation of JOL accuracy. This underestimation of JOL accuracy is especially pronounced in poor learners because they are more prone to random guessing in recognition tests. As a consequence, greater underestimation of JOL accuracy in poor learners inevitably results in a spurious positive relation between JOL accuracy and recognition performance (Vuorre & Metcalfe, 2022). Vuorre and Metcalfe (2022) reported a set of data simulations to illustrate this spurious positive relation. Therefore, positive relations detected in recognition tasks again cannot justify the conclusion that good learners are also good monitors. As recommended by Vuorre and Metcalfe (2022), a better approach to test this assumption is to investigate the association between JOL accuracy and test performance in recall-based, rather than recognition-based, learning tasks.

Despite advances in research methods, the fundamental question remains unresolved. Does stronger learning ability genuinely align with better monitoring accuracy, as commonly assumed? Or might there be an unexpected disconnection between the two? The current research aims to address this important question. In Study 1, we conducted a meta-analysis, integrating open data across 43 experiments involving recall tests, to determine whether learning ability relates to monitoring accuracy. Surprisingly, it detected a reliable negative

relation, suggesting that more proficient learners tend to be poorer at monitoring their learning status. In Study 2, we proposed and tested three possible explanations for this counterintuitive relation. In Studies 3 and 4, we conducted controlled experiments to conceptually replicate the main findings of Studies 1 and 2.

## Study 1: Meta-analysis

### Method

#### Literature search

We conducted an extensive search for open data at Open Science Framework (OSF). It should be noted that none of the included studies specifically set out to explore the relation between learning ability and monitoring accuracy. Hence, we required the raw data from these studies to calculate target measures, which is why we searched OSF for open data. Our systematic search was initially conducted in July 2022, and then updated in April 2024. The search terms were [JOL* OR judgment* of learning OR judgement* of learning]. In a preliminary search, OSF returned over 10,000 records. Given that the search results were sorted by relevance, we decided to only review the first 1,000 results. We also manually screened the Confidence Database compiled by Rahnev et al. (2020), which contains 145 datasets of metacognition research.

We note that the goal of this meta-analysis was not to provide a comprehensive review of all studies on monitoring accuracy, but rather to utilize publicly available datasets to investigate the specific question of whether there is any relation between learning ability and monitoring accuracy. Using open data is a practical and efficient approach to generate a large-sample dataset to investigate this focused question.

#### Inclusion and exclusion criteria

(a) Only experimental studies were included, which must include both a learning and a testing phase. Furthermore, during the learning phase, participants had to make a JOL after studying each item. Several studies were excluded for methodological reasons. For instance, some studies incorporated a restudy phase between making JOLs and final test (Zimdahl & Undorf, 2021), some required participants to provide JOLs for others rather than for themselves (Tauber & Witherby, 2019), and some introduced a practice test prior to the JOL phase as an intervention to enhance JOL accuracy (Robey et al., 2017).

(b) The final test on learning performance must be in a recall format, such as cued recall or free recall (Mendes et al., 2019). Studies employing recognition tests (e.g., old/new recognition) were excluded for the reasons discussed above (Vuorre & Metcalfe, 2022).

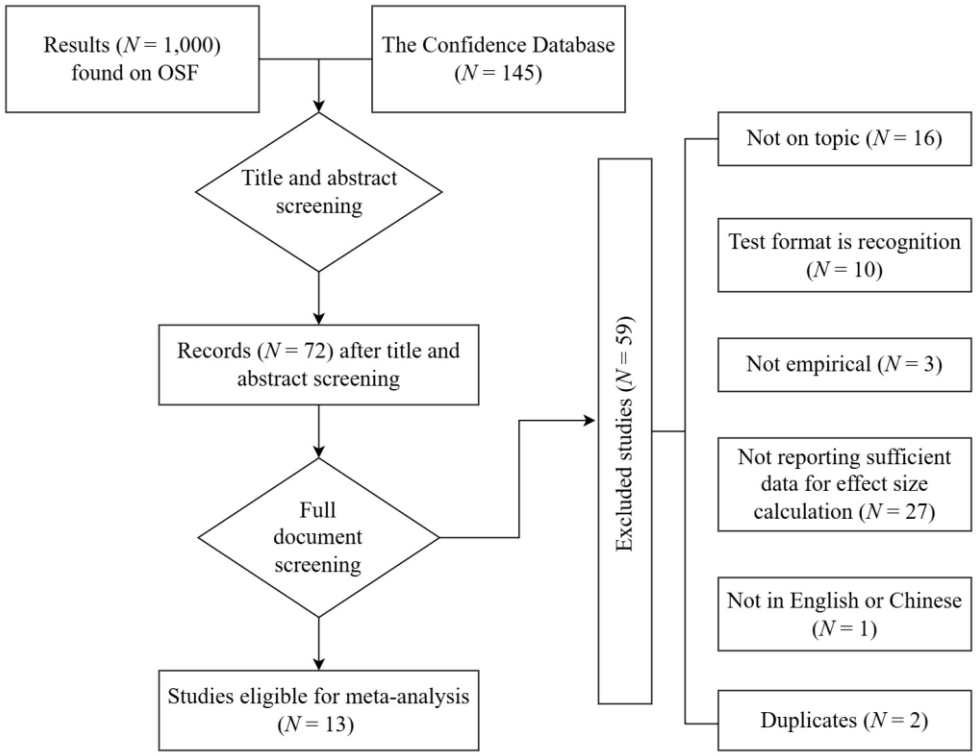
(c) Studies had to provide sufficient details of experimental design and procedure to allow us to re-analyze their raw data and calculate target measures. Those lacking detailed descriptions of experimental methods (e.g., brief conference posters) were excluded.

(d) Studies had to provide item-level data of JOLs and recall accuracy. Those only providing aggregate data were excluded (Myers et al., 2020).

(e) Only studies written in English or Chinese were included.

In total, 13 studies were identified as eligible for meta-analysis, from which 43 effects were extracted, involving data from 2,533 participants. The screening procedure is illustrated in Figure 1.

Figure 1. Flowchart illustrating the study screening process and results




**Moderator coding**

Moderator coding was independently performed by the first and corresponding authors.

In total, five potential moderators were coded, including material type, test format, age, country, and task difficulty.

***Material type*.** The included studies utilized different types of study materials as stimuli, such as (a) images, (b) word pairs, and (c) word lists. It is well-established that learners adopt different strategies to study and utilize varied cues to make JOLs for different types of study materials (Undorf & Bröder, 2021).

***Test format*.** Based on the format of the final test in each experiment, the effects were divided into two categories: (a) cued recall and (b) free recall. There are notable distinctions between cued recall and free recall tests.

***Age*.** Learning ability declines as a function of age across adulthood (Small et al., 1999), but it remains controversial whether monitoring ability declines or not (Siegel & Castel, 2019). Given that this meta-analysis primarily focuses on the relation between these abilities, age was included as a potential moderator. According to participants' mean age in each experiment, the included effects were divided into two categories: (a) young adults ($M_{age}$ ranging from 18.60 to 39.74) and (b) older adults ($M_{age}$ ranging from 67.79 to 72.50).

***Country*.** According to the country from which participants were recruited, the included effects were categorized into three categories: (a) China, (b) Germany, and (c) United States. Coding country as a potential moderator allows us to determine whether the documented findings generalize to different countries and social cultures.

***Task difficulty*.** We computed an average score of test performance across all participants in each experiment to represent the level of task difficulty, and considered task difficulty as a potential moderator.

**Measures of JOL accuracy and learning ability**

The most commonly used measure of JOL accuracy is the Goodman-Kruskall *Gamma* (*G*) correlation, which measures the rank correlation between JOLs and recall accuracy across trials (Nelson, 1984). *G* is computed by calculating $(N_c - N_d)/(N_c + N_d)$, where $N_c$ is the number of pairs where the rankings for both JOLs and recall accuracy align (i.e., the number of concordant pairs), and $N_d$ is the number of pairs where the rankings are reversed (i.e., the number of discordant pairs). For each participant in each experiment, we calculated a *G* as a measure of JOL accuracy. In addition to *G*, the area under the Type 2 receiver operating

characteristic curve (AUROC2) has also been occasionally used to measure JOL accuracy (Fleming & Lau, 2014). Hence, we also employed this method to measure JOL accuracy and performed the same meta-analysis, which showed the same result patterns as those obtained with *G*. Detailed results of the AUROC2 measure are reported in the SM.

Among the included studies, 105 participants provided constant JOLs to all study items, recalled all items correctly, or did not recall any items in the final test. Their data must be excluded because constant values in JOLs or recall accuracy do not permit calculation of *G*. In addition, following Myers et al. (2020), we also removed 22 participants who did not provide JOLs for at least 80% of study items. The final data for the meta-analysis came from 2,406 participants.

In addition to JOL accuracy, we calculated test performance for each participant as a measure of individual learning ability. Test performance was calculated as the proportion of items correctly recalled in the final test, ranging from 0 to 1.

It is important to clarify the distinction between two key measures: test performance (an index of individual learning ability) and mean test performance (an index of group-level task difficulty). For each participant in each experiment, test performance—defined as the proportion of items correctly recalled by that participant—was used as a measure of learning ability. In contrast, mean test performance—calculated as the average of test performance across all participants in a given experiment—was used as a measure of task difficulty. Since all participants within an experiment completed the same learning task, individual differences in test performance should primarily reflect variations in learning ability. By contrast, differences in mean test performance across experiments (or across different learning tasks) should mainly reflect variations in task difficulty.

**Methods for meta-analysis**

In each of the included experiments, we first computed a Pearson's *r* correlation between test performance and JOL accuracy (indexed by *G* values) across participants to determine the relation between learning ability and monitoring accuracy. These *r* scores were then transformed to Fisher's *Z*s for meta-analysis. To address non-independence issues arising from multiple effects extracted from the same study, we conducted three-level random-effects meta-analyses, which consider three variance components, including sampling variance at

level 1, variance between effect sizes extracted from the same study at level 2, and variance between studies at level 3 (Cheung, 2014). Heterogeneity was assessed via *Q* tests and $I^2$, with $I^2$ further split into between- and within-study components, denoted as $I^2_{\text{between}}$ and $I^2_{\text{within}}$, respectively. Univariate meta-regression analyses were conducted to detect potential moderators. Notably, none of the coded factors exhibited significant moderating effects, alleviating concerns about potential confounding effects among the included moderators.

**Results**

The weighted mean relationship between test performance and JOL accuracy was significantly negative, Fisher's $Z$ = -0.10 [-0.15, -0.05], $r$ = -.10, $p$ < .001 (Figure 2), indicating that high-ability learners are actually poorer at gauging their learning status. This negative correlation ($r$ = -.10) means that an increase of 1 *SD* in learning ability (measured as test performance) was associated with a reduction of 0.1 *SD* in monitoring accuracy (measured as the *G*). There was some heterogeneity among the included effects, $Q(42)$ = 65.49, $p$ = .01. Specifically, within-study heterogeneity was at a low-to-moderate level, $I^2_{\text{within}}$ = 34.6%, and between-study heterogeneity was minimal, $I^2_{\text{between}}$ < 0.1%. Among the 43 effects, 31 exhibited a negative relation, with only 12 showing the converse pattern. The proportion of effects showing a negative relation was substantially greater than the proportion showing the converse pattern, $\chi^2(1) = 8.40$, $p$ = .004. Three analyses of publication bias detection consistently found no evidence of publication bias (see the SM for the detailed results).

Figure 2. Forest plot of the meta-analysis in Study 1

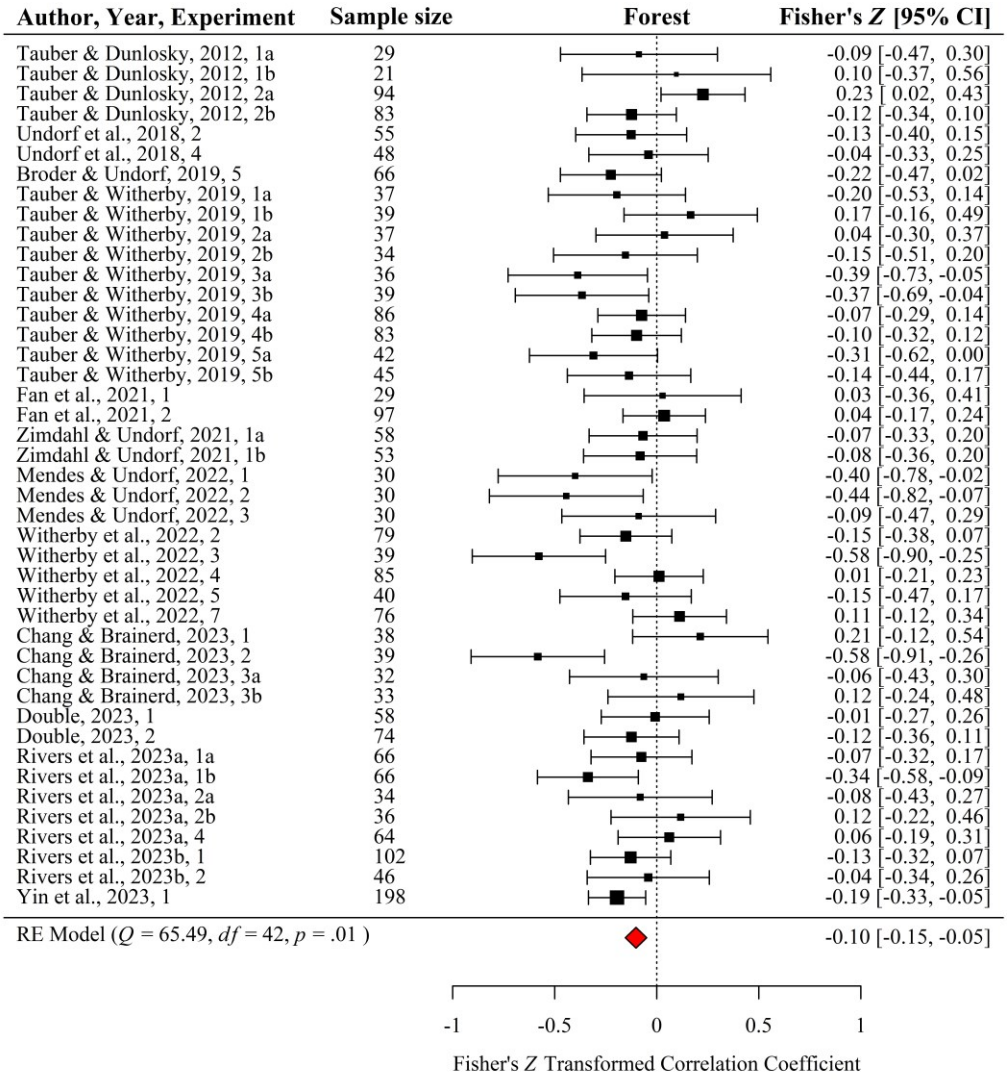



Table 1 lists detailed results of the moderator analyses. None of the included moderators showed a detectable moderating effect, *p*s $\geq$ .16. Critically, as shown in Table 1, the negative relation between test performance and JOL accuracy generalizes to various types of study materials, test formats, age groups, countries (or social cultures), and tasks with varying levels of difficulty.

Table 1. Moderator analysis results

| Moderators | *k* | Fisher's *Z* | 95% CI | *r* | $Q_M$ | *p* |
|---|---|---|---|---|---|---|
| Material type | | | | | 3.70 | .157 |
| Image | 2 | -0.31 | [-0.55, -0.08] | -.30 | | .009 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Word list | 16 | -0.07 | [-0.15, 0.01] | -.07 | | .098 |
| Word pair | 25 | -0.10 | [-0.17, -0.04] | -.10 | | .003 |
| Test format | | | | | 0.52 | .472 |
| Cued recall | 23 | -0.12 | [-0.19, -0.05] | -.12 | | .001 |
| Free recall | 20 | -0.08 | [-0.16, -0.01] | -.08 | | .035 |
| Age | | | | | 0.39 | .531 |
| Old adults | 7 | -0.14 | [-0.28, -0.01] | -.14 | | .042 |
| Young adults | 36 | -0.09 | [-0.15, -0.04] | -.09 | | .001 |
| Country | | | | | 1.14 | .565 |
| China | 3 | -0.08 | [-0.25, 0.10] | -.08 | | .397 |
| Germany | 8 | -0.17 | [-0.30, -0.04] | -.16 | | .013 |
| United States | 32 | -0.09 | [-0.15, -0.03] | -.09 | | .005 |
| Task difficulty | 43 | | | | 0.34 | .562 |

## Study 2: Assessment of Three Possible Explanations

To our knowledge, no existing theories are available to explain the negative relation between learning ability and monitoring accuracy observed in Study 1. Here we propose three possible explanations for this counterintuitive relation and test them in Study 2.

### Statistical artifact

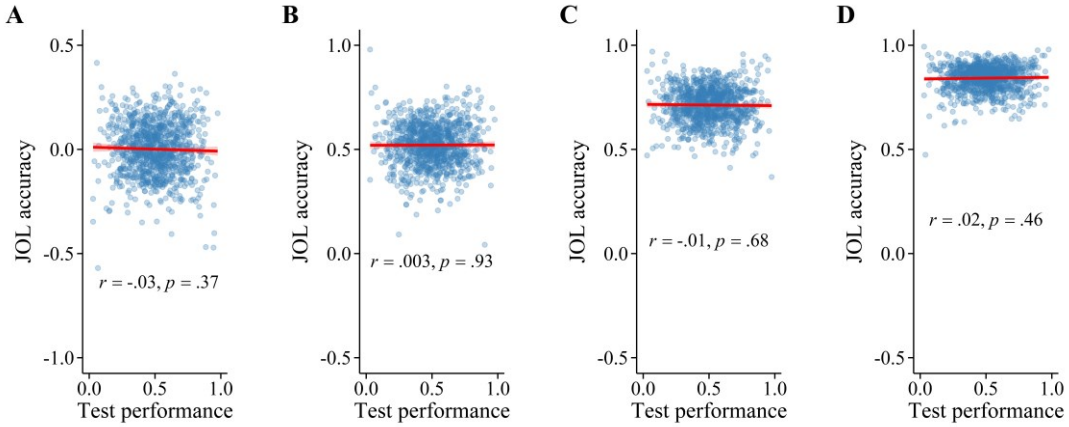
The first possible explanation is that the negative relation between test performance and JOL accuracy is merely a statistical artifact resulting from our data analysis methods. Specifically, in the meta-analysis of Study 1, recall accuracy in the final test was used to calculate both JOL accuracy (indexed by the *G* between JOLs and recall accuracy across trials) and test performance (indexed by average recall accuracy across all trials). This repeated use of recall accuracy may introduce statistical dependency between JOL accuracy and test performance, leading to a spurious relation between the two. To test this hypothesis, we followed Vuorre and Metcalfe (2022) and performed a data simulation.

Details of the simulation procedure are reported in the SM. In brief, data were generated for 1,000 simulated participants, each studying 100 items and making a JOL for each item

before taking a final recall test. Participants' learning abilities were sampled from a normal distribution, with variability reflecting individual differences in the proportion of items they could remember. These learning abilities were then used to simulate trial-level recall outcomes. In four separate simulations, we varied the difference between participants' mean JOLs for remembered and forgotten items to reflect different levels of monitoring accuracy.

As shown in Figure 3, across varying levels of JOL accuracy, repeated use of recall accuracy does not yield any relation between JOL accuracy and test performance in recall tests (for related findings, see Vuorre & Metcalfe, 2022). These results do not support the statistical artifact explanation.

Figure 3. Scatter plots illustrating the null relation between JOL accuracy and test performance in the data simulation




*Note*: In Panels A-D, the average *G*s are 0 (poor JOL accuracy), 0.52 (modest JOL accuracy), 0.71 (good JOL accuracy), and 0.84 (excellent JOL accuracy), respectively. The red line is the regression line, with error bars depicting 95% CI.

**Scale usage**

The second possible explanation concerns a potential difference in JOL scale usage between high- and low-ability learners. Specifically, expert learners generally have high confidence in their learning performance and tend to provide high JOLs for most or even all study items. This tendency can lead to low variance in their JOLs, resulting in poor JOL accuracy. Put differently, individuals with high learning ability may only use a narrow range

of the JOL scale (e.g., 80-100 on a 0-100 scale) to make JOLs, reducing JOL variance and resolution. Supporting this explanation, Witherby et al. (2023) found that the lowest JOLs reported by students with high prior knowledge were at 50 or even 80 on a 0-100 JOL scale, substantially higher than the lowest JOLs (e.g., 0) given by students with low prior knowledge.

To test the scale-usage explanation, we re-analyzed the data from the meta-analysis in Study 1. Specifically, we first calculated the standard deviation (*SD*) of JOLs for each participant and took JOL *SD* as a measure of JOL scale usage. Next, a Pearson *r* correlation between JOL *SD* and test performance was calculated in each experiment. These *r* scores were then transformed into Fisher's *Z*s and submitted to a multilevel random-effects meta-analysis. The results revealed no significant relationship between JOL *SD* and test performance, Fisher's $Z$ = -0.04 [-0.11, 0.03], $r$ = -.04, $p$ = .32, suggesting little systematic difference in JOL scale usage between high- and low-ability learners. These results do not support the scale usage explanation.

**Expert underconfidence**

The third possible explanation was developed based on the findings of Witherby et al. (2023). In their study, Witherby et al. observed a negative relation between prior knowledge and JOL accuracy in recall-based learning tasks. Furthermore, they found that this negative relation primarily derived from the fact that participants with high prior knowledge could remember more items they previously judged as non-memorable. In other words, the negative relation between prior knowledge and JOL accuracy arose from the fact that participants with high prior knowledge underestimated their ability to remember challenging items (i.e., the items they predicted they would not remember but actually successfully remembered).

Building on Witherby et al.'s explanation, we propose an expert underconfidence hypothesis to account for the negative relation between learning ability and monitoring accuracy. Specifically, we hypothesize that high-ability learners tend to remember many items they mistakenly believe they will not remember, and this underconfidence in expert learners amplifies the gap between their JOLs and recall accuracy, in turn leading to a negative relationship between learning ability and monitoring accuracy across individuals.

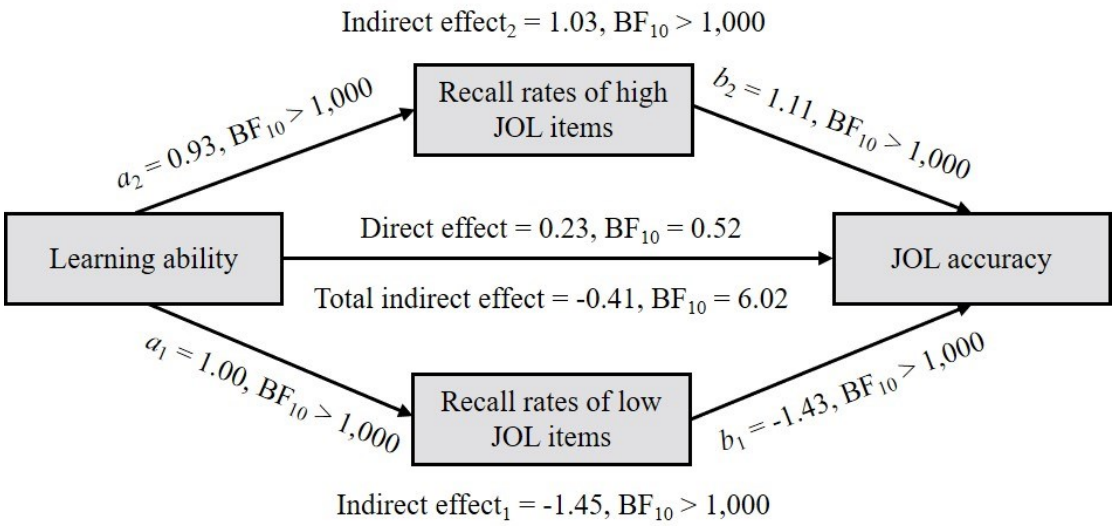
To test this explanation, we re-analyzed the data from the meta-analysis in Study 1 and

conducted a Bayesian mixed-effects mediation analysis using the R *brms* package, with all parameters set as default. Specifically, for each participant in each experiment, we divided study items into three terciles based on a tri-partite ranking of JOLs, including a high JOL set (comprising one-third of items with the highest JOLs), a medium JOL set (comprising one-third of items with mid-range JOLs), and a low JOL set (comprising one-third of items with the lowest JOLs). When study items were associated with tied JOLs at the separation boundaries, they were randomly assigned to maintain balanced set sizes. Low JOL items were those perceived as forgettable and high JOL items were those perceived as memorable. We then calculated recall rates of both low and high JOL items for each participant.

In the mediation analysis, learning ability (indexed by test performance) was treated as the independent variable, recall rates of low and high JOL items were included as two parallel mediators, and JOL accuracy was treated as the dependent variable, with random effects added at both the study and experiment levels. As shown in Figure 4, the results revealed a negative indirect effect of learning ability on JOL accuracy through improving recall of low JOL items, $a_1*b_1 + \sigma_{ab} = -1.45$, $BF_{10} > 1{,}000$, suggesting that the negative relation between learning ability and monitoring accuracy is at least partially due to the fact that high-ability learners can remember many items they erroneously believe they will forget (Goodman, 1960; Kenny et al., 2003). This tendency reduces JOL accuracy because these items are perceived as forgettable but are in fact remembered.

Meanwhile, there was also a positive indirect effect of learning ability on JOL accuracy through improving recall of high JOL items, $a_2*b_2 + \sigma_{ab} = 1.03$, $BF_{10} > 1{,}000$, suggesting that high-ability learners can successfully remember more items they believe they will remember, which in turn improves their JOL accuracy. This tendency increases JOL accuracy because these items are perceived as memorable and are indeed remembered. This positive indirect effect (1.03) partially cancelled out the negative indirect effect (-1.45). However, the negative indirect effect was stronger than the positive one, leading to an overall negative relation between learning ability and JOL accuracy, total indirect effect = -0.41, $BF_{10} = 6.02$. Overall, these findings support the expert underconfidence hypothesis.

Figure 4. Mediation results in Study 2




## Study 3: Experimental Investigation

Given that the findings of Studies 1 and 2 are novel, the first aim of Study 3 was to test their replicability in a controlled experiment. Study 3 also aimed to address several limitations of Studies 1 and 2. In those studies, recall accuracy was used for calculating both test performance and JOL accuracy. To avoid this issue, in Study 3, participants' learning ability and monitoring accuracy were measured in two separate tasks. Furthermore, in the mediation analysis of Study 2, JOLs were made on a continuous scale (e.g., 0-100), but we arbitrarily classified high JOL items as the ones that participants thought they would remember and low JOL items as the ones that they thought they would forget. To measure these categorical judgments more directly, in Study 3, we followed Witherby et al. (2023) and asked participants to make JOLs on a binary scale (*0 = I will not remember it*; *1 = I will remember it*), rather than on a continuous scale.

### Method

### Participants

A pilot study with 30 participants found a moderate negative relation between test performance and JOL accuracy, $r$ = -.34. A power analysis, conducted via G*Power (Faul et al., 2007), indicated that 71 participants were needed to detect a significant (one-tailed, α = .05) negative relation with .90 power. We pre-planned to use a one-tailed correlation analysis because we already had an *a priori* hypothesis about the direction of this relation

according to Study 1's meta-analysis and our pilot results. Considering potential participant exclusion due to constant values in JOLs or recall accuracy, we decided to conservatively increase the sample size to 120.

Accordingly, 120 participants were recruited from Tianjin Vocational Institute. Eighteen participants provided constant JOL values across all trials, and their data were excluded, leaving final data from 102 participants ($M_{age}$ = 19.55, $SD$ = 0.79; 77.5% female). All participants were native Chinese speakers and had no prior learning experience of the Swahili language. They provided informed consent and received monetary compensation. This research received ethics approval from Faculty of Psychology, Beijing Normal University (Protocol Number: BNU202112300096).

**Materials**

A learning task, featuring 30 paintings developed by Soares and Storm (2022), was employed to assess participants' learning ability. Each image (1351 × 736 pixels) displayed a painting along with its title and the artist's name. Considering that some participants might have prior knowledge about certain artists, we replaced all artist names with common Chinese names.

Participants' monitoring ability was assessed in a monitoring task which consisted of 23 Swahili-Chinese word pairs selected from the Swahili-Chinese database developed by Fan et al. (2025). The difficulty levels of the selected word pairs, defined as mean recall rates in Fan et al.'s database, ranged from 0.27 to 0.67. Three pairs were used for practice, with the remaining 20 pairs used in the formal experiment. Data from practice trials were excluded from analyses.

**Procedure**

Each participant completed both a learning and a monitoring task, with task order counterbalanced across participants. In the learning task, participants were instructed to study 30 paintings and remember as many visual details as possible. During the learning phase, the 30 paintings were presented one-by-one in a random order, with each painting displayed for 20 s. After studying all paintings, participants completed a 10-min distractor task, in which they solved various algebra problems. After the distractor task, they took a final test on all paintings. The test included 60 multiple-choice questions (e.g., *What type of building is*

*depicted in Li Mingwei's 'Rain'?*), with two questions on each painting. Each question included one correct option and three lures. Test questions were presented one-by-one in a random order, but with a constraint that the two questions for each painting were always presented consecutively. The final test was untimed, and no feedback was provided. The experiment was programmed using *jsPsych* 7.2.3 (de Leeuw, 2015).

In the monitoring task, participants studied 20 Swahili-Chinese word pairs. They were informed that after studying each word pair, they would need to predict whether they would remember the Chinese translation of the Swahili word in a test 1-min later. The word pairs were presented one-by-one in a random order, with each word pair shown for 10 s. After studying each pair, participants made a binary JOL (*0 = I will not remember it*; *1 = I will remember it*). JOL ratings were self-paced. After studying all word pairs, participants solved a set of algebra problems for 60 s. Then they proceeded to the final test, where the 20 Swahili words were presented one-by-one in a random order and participants were asked to recall the Chinese translation to each Swahili word. The final test was untimed, and no feedback was provided. The experiment was programmed using *PsychoPy* 2023.2.3 (Peirce et al., 2019).

**Results**

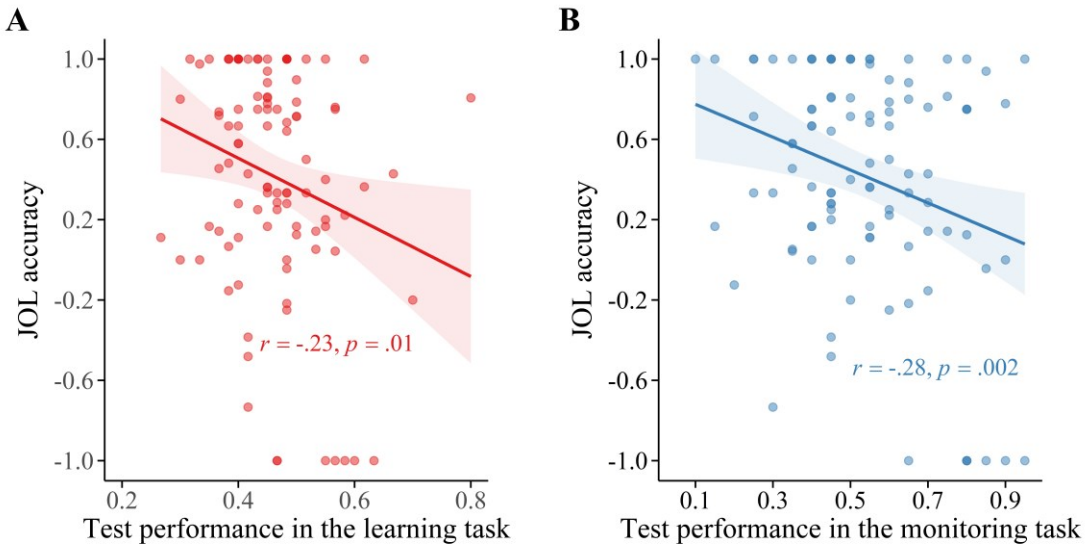
All Bayesian analyses reported in Studies 3 and 4 were performed via the R *BayesFactor* package with all parameters set as default, except for the Bayesian mediation analyses, which were performed via the R *brms* package with all parameters again set as default.

For each participant, we calculated a *G* between JOLs and recall accuracy in the monitoring task as a measure of JOL accuracy. Because it is impossible to construct the full Type 2 ROC for binary JOLs (Fleming & Lau, 2014), AUROC2 scores were not calculated. Test performance in the learning task was calculated as a measure of learning ability. Test performance in the monitoring task was also calculated, serving as a secondary measure of learning ability. A two-tailed correlation analysis showed a positive relation between test performance in the learning and monitoring tasks, $r = .24$, $p = .02$, $BF_{10} = 3.45$, suggesting that good learners in one task are also good learners in the other one.

A one-tailed correlation analysis revealed a negative relation between JOL accuracy in the monitoring task and test performance in the learning task, $r = -.23$, $p = .01$, $BF_{10} = 5.75$ (Figure 5A). There was also a negative correlation between JOL accuracy and test

performance in the monitoring task, $r$ = -.28, $p$ = .002, $BF_{10}$ = 19.77 (Figure 5B). Together, these results indicate that, regardless of whether learning ability and JOL accuracy are measured within the same task or across different tasks, there is always a negative relation between these two variables.

Figure 5. Scatter plot depicting the relation between learning ability and monitoring accuracy in Study 3

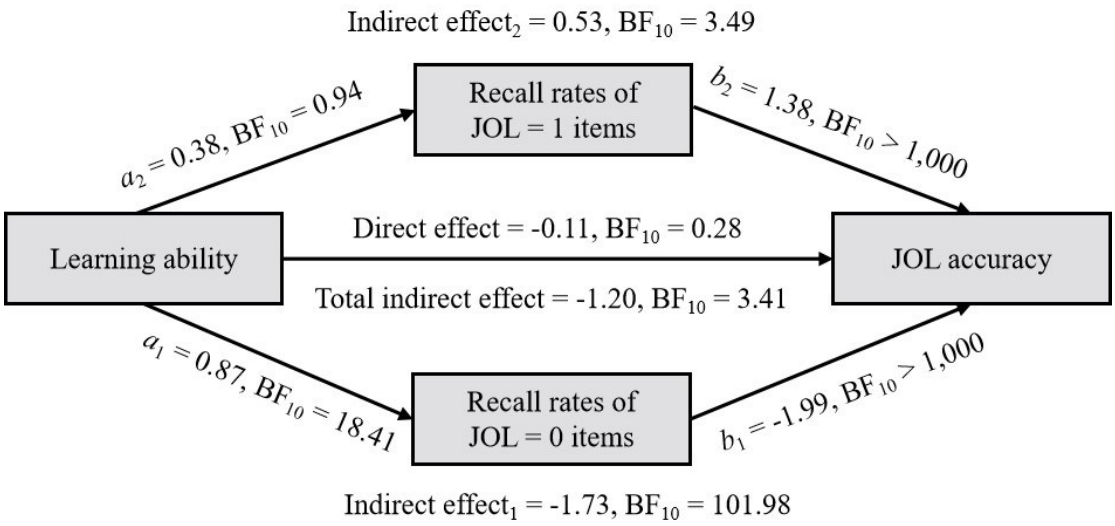



*Note*: Panel A depicts the relation between JOL accuracy in the monitoring task and test performance in the learning task; Panel B illustrates the relation between JOL accuracy and test performance in the monitoring task. Error bars represent 95% CI.

To test the expert underconfidence hypothesis, we conducted a Bayesian mediation analysis, in which learning ability (indexed by test performance in the learning task) was treated as the independent variable, monitoring accuracy (indexed by *G* values in the monitoring task) served as the dependent variable, and recall rates of JOL = 0 and JOL = 1 items were included as two parallel mediators. As shown in Figure 6, there was a negative indirect effect of learning ability on JOL accuracy via improving recall of JOL = 0 items, $a_1*b_1$ = -1.73, $BF_{10}$ = 101.98, reflecting that expert learners are more likely to remember the items perceived as non-memorable, but doing so reduces their monitoring accuracy. Meanwhile, there was also a positive indirect effect of learning ability on JOL accuracy via

improving recall of JOL = 1 items, $a_2*b_2 = 0.53$, $BF_{10} = 3.49$, reflecting that expert learners are more likely to remember the items perceived as memorable, and doing so increases their monitoring accuracy. Critically, the total indirect effect was negative, $a_T*b_T = -1.20$, $BF_{10} = 3.41$. This composite mediation effect suggests that, while the negative mediation effect through improving recall of JOL = 0 items and the positive mediation effect through improving recall of JOL = 1 items counteracted each other, the former (-1.73) was much stronger than the latter (0.53), leading to an overall negative relation between learning ability and monitoring accuracy. Lastly, we performed another mediation analysis, substituting the independent variable with test performance in the monitoring task. This analysis showed the exact same result pattern (see the SM).

Figure 6. Mediation results in Study 3

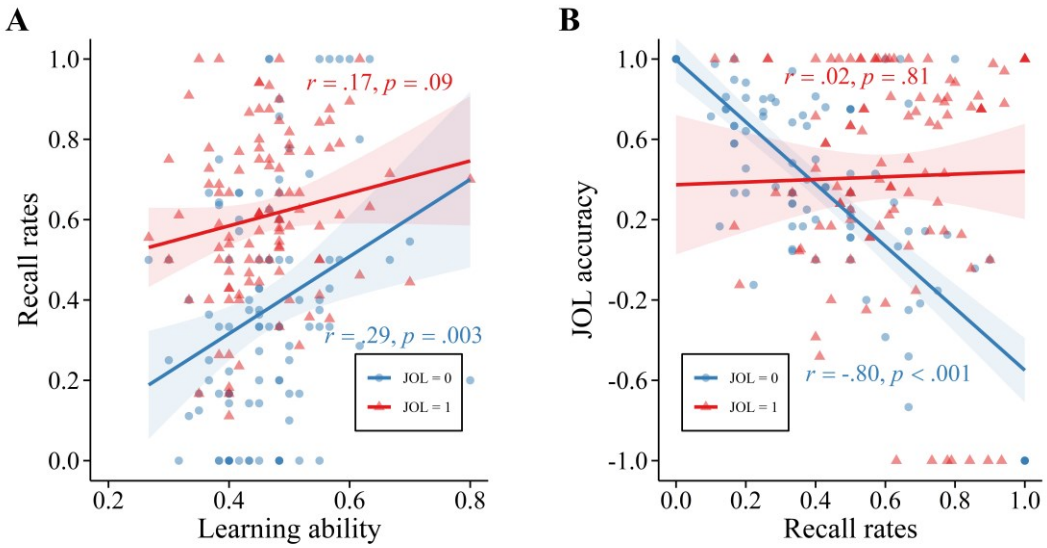



*Note*: In this mediation model, learning ability was indexed by test performance in the learning task.

Figure 7 provides a visual representation of the mediation effects, illustrating the relations among the independent variable, the two mediators, and the dependent variable. As depicted in Figure 7A, learning ability (indexed by test performance in the learning task) positively predicted recall rates of both JOL = 0 and JOL = 1 items, and these prediction effects did not differ greatly. However, as shown in Figure 7B, successfully remembering JOL

= 0 items substantially reduced JOL accuracy. By contrast, successfully remembering JOL = 1 items only slightly improved JOL accuracy. Therefore, the total relation between learning ability and JOL accuracy was negative.

Figure 7. Scatter plots depicting the relations among different variables in Study 3




*Note*: Panel A depicts the relations between learning ability (indexed by test performance in the learning task) and recall rates of JOL = 0 and JOL = 1 items; Panel B illustrates the relations between JOL accuracy and recall rates of JOL = 0 and JOL = 1 items. Error bars represent 95% CI.

### *Dunning–Kruger effect*

Intuitively, readers may consider that the negative relation between JOL accuracy (i.e., relative JOL accuracy) and test performance runs counter to the well-known *Dunning–Kruger* (DK) effect (Dunning, 2011; Kruger & Dunning, 1999), which refers to the phenomenon that people with limited ability in a given domain substantially overestimate their task performance, whereas those with high ability estimate their performance more accurately (or slightly underestimate it). However, it should be highlighted that our findings do not directly challenge the DK effect. Specifically, the current research probed the relation between relative JOL accuracy (i.e., JOL resolution) and test performance. By contrast, the DK effect concerns the relation between absolute accuracy of metacognitive judgments (e.g., absolute JOL

accuracy or JOL calibration) and objective task performance (e.g., test performance).

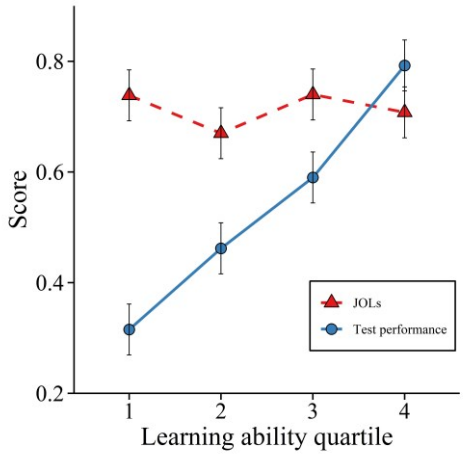
We re-analyzed the data from the monitoring task of Study 3 to demonstrate that the DK effect was also observed in the current data. Specifically, we first calculated an average JOL across trials for each participant to represent the proportion of items the participant predicted he or she would remember (i.e., subjective estimate of learning ability). We also calculated each participant's test performance in the monitoring task as a measure of objective learning ability. Next, we divided the 102 participants into quartiles (1 vs. 2 vs. 3 vs. 4) according to their test performance in the monitoring task, with Group 1 comprising the 26 participants with lowest test performance (very poor learning ability) and Group 4 comprising the 26 participants with highest test performance (excellent learning ability). Each of Groups 2 and 3 contained 25 participants.

A Bayesian mixed analysis of variance (ANOVA) was conducted, with measurement type (JOLs vs. test performance) as a within-subjects variable and group as a between-subjects variable. As shown in Figure 8, the results reveal a substantial interaction between measurement type and group, $F(3, 98) = 41.64$, $p < .001$, $\eta_p^2 = .56$, $BF_{10} > 1{,}000$. These results perfectly replicate the DK effect, with Group 1 showing substantial overestimation of test performance, difference between JOLs and test performance = .42, 95% CI = [.35, .49], $t(25) = 12.57$, $p < .001$, Cohen's $d = 2.47$, $BF_{10} > 1{,}000$, and Group 4 exhibiting underestimation of test performance, difference between JOLs and test performance = -.08, 95% CI = [-.15, -.02], $t(25) = -2.79$, $p = .01$, Cohen's $d = -0.55$, $BF_{10} = 4.72$.

Many researchers have argued that the DK effect may be merely a statistical artifact (Gignac & Zajenkowski, 2020; Jansen et al., 2021; Krueger & Mueller, 2002), induced by *regression-toward-the-mean* (Stigler, 1997) and the *better-than-average* effect (Alicke et al., 1995). A data simulation is provided in the SM to demonstrate the statistical issues associated with the DK effect, as illustrated in the present data. Whatever the merits of these arguments against the DK effect as a general phenomenon, it is clear that when trial-by-trial metacognitive judgments are obtained, as was done here, the standard DK analysis as shown in Figure 8 encourages an incorrect inference. Far from indicating worse metacognitive monitoring in poor than good learners, poor learners in fact have better metacognitive insight into their learning. Poor learners (relative to good ones) may overestimate their learning and

provide average JOLs that are objectively too high (the DK effect), while at the same time showing a tighter correlation between their trial-by-trial JOLs and later recall accuracy (i.e., greater JOL resolution).

Figure 8. Line plot depicting the DK effect in Study 3




*Note*: Participants were divided into quartiles according to their test performance in the monitoring task. Error bars represent 95% CI.

## Study 4: Pre-registered Replication

In Study 4, we conducted a pre-registered, large-sample experiment to further test the replicability of the findings of Studies 1-3. It also aimed to address a limitation of Study 3, in which JOLs were made on a binary scale and there were only 20 trials in the monitoring task. The binary JOLs and small number of trials might jointly induce frequent occurrence of extreme *G* values at 1 or -1, as can be seen in Figure 5. To mitigate this problem, Study 4 employed a continuous (i.e., 0-100) JOL scale and increased the number of trials to 30.

### Method

### Participants

We pre-registered to recruit 500 participants to run a large-sample replication experiment (https://osf.io/kwbzm). Accordingly, 512 participants were recruited from NAODAO (an online behavioral research platform; https://www.naodao.com/). Data from 31 participants were excluded because they provided constant JOLs across all trials, recalled all items

correctly, or did not recall any items in the final test, leaving final data from 481 participants ($M_{age}$ = 24.85, *SD* = 5.94; 50.9% female). According to the negative relation (*r* = -.28) detected in the monitoring task of Study 3, this sample size had a statistical power greater than .99 to detect a significant (one-tailed, α = .05) negative relation between test performance and monitoring accuracy. All participants were native Chinese speakers and had no prior learning experience of the Swahili language. They provided online informed consent and received monetary compensation. This research received ethics approval from Faculty of Psychology, Beijing Normal University (Protocol Number: BNU202112300096).

**Materials and procedure**

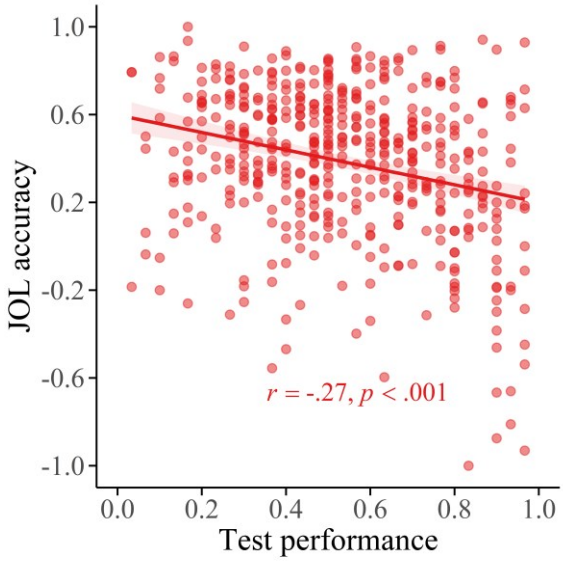
The stimuli were 33 Swahili-Chinese word pairs selected from Fan et al. (2025), with their difficulty levels ranging from 0.12 to 0.87. Three pairs were used for practice, with the remaining 30 pairs used in the formal experiment. Data from practice trials were excluded from analyses.

The procedure was the same as that in the monitoring task of Study 3, except that participants studied 30 (rather than 20) Swahili-Chinese word pairs, and made their JOLs on a continuous (rather than binary) scale ranging from 0 (*Sure I will not remember it*) to 100 (*Sure I will remember it*). The experiment was programmed using *jsPsych* 7.2.3 (de Leeuw, 2015).

**Results**

For each participant, we calculated a *G* correlation between JOLs and recall accuracy as a measure of JOL accuracy. An AUROC2 score was also calculated as a second measure of JOL accuracy, which showed the same result patterns as those of *G* (see the SM). Test performance was calculated as a measure of learning ability. A one-tailed correlation analysis revealed a negative relation between JOL accuracy (indexed by *G*) and test performance, *r* = -.27, $p < .001$, $BF_{10} > 1{,}000$ (Figure 9).

Figure 9. Scatter plot depicting the relation between learning ability and monitoring accuracy in Study 4




*Note*: Error bars represent 95% CI.

To test the expert underconfidence explanation, for each participant we divided the 30 study items into three terciles based on a tri-partite ranking of JOLs, including a high JOL set (comprising 10 items with the highest JOLs), a medium JOL set (comprising 10 items with mid-range JOLs), and a low JOL set (comprising 10 items with the lowest JOLs). If some items had tied JOLs at the separation boundaries, they were randomly assigned to ensure each set contained 10 items. Then, we computed recall rates for the low and the high JOL items, respectively. Next, a Bayesian mediation analysis was conducted, in which learning ability (indexed by test performance) was treated as the independent variable, monitoring accuracy (indexed by *G*) served as the dependent variable, and recall rates of low and high JOL items were included as two parallel mediators.

As shown in Figure 10, there was a negative indirect effect of learning ability on JOL accuracy via improving recall of low JOL items, $a_1*b_1 = -1.26$, $BF_{10} > 1{,}000$. Meanwhile, there was also a positive indirect effect of learning ability on JOL accuracy via improving recall of high JOL items, $a_2*b_2 = 0.76$, $BF_{10} > 1{,}000$. Critically, the total indirect effect was negative, $a_T*b_T = -0.50$, $BF_{10} = 130.10$. These results again support the expert underconfidence explanation.

Figure 10. Mediation results in Study 4

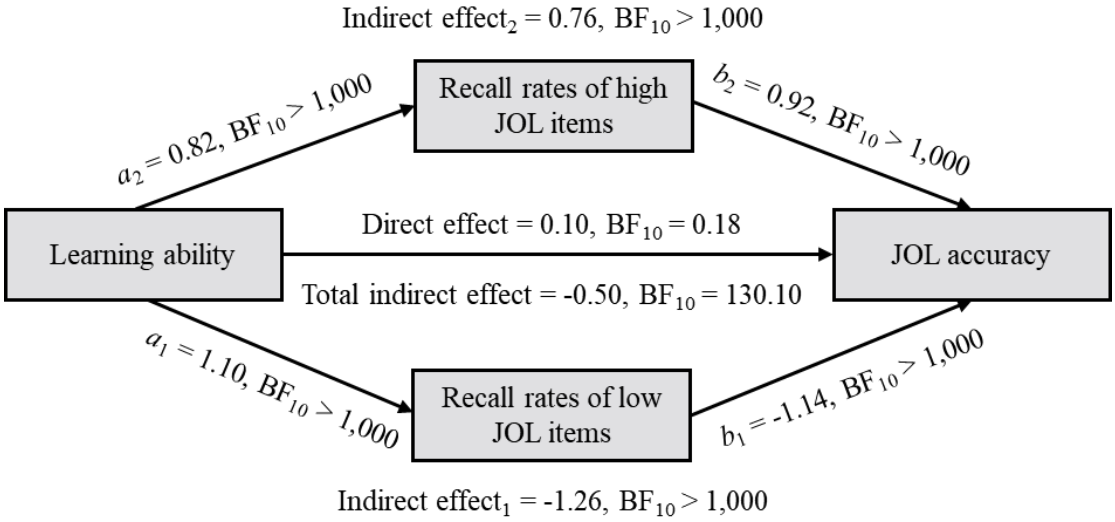
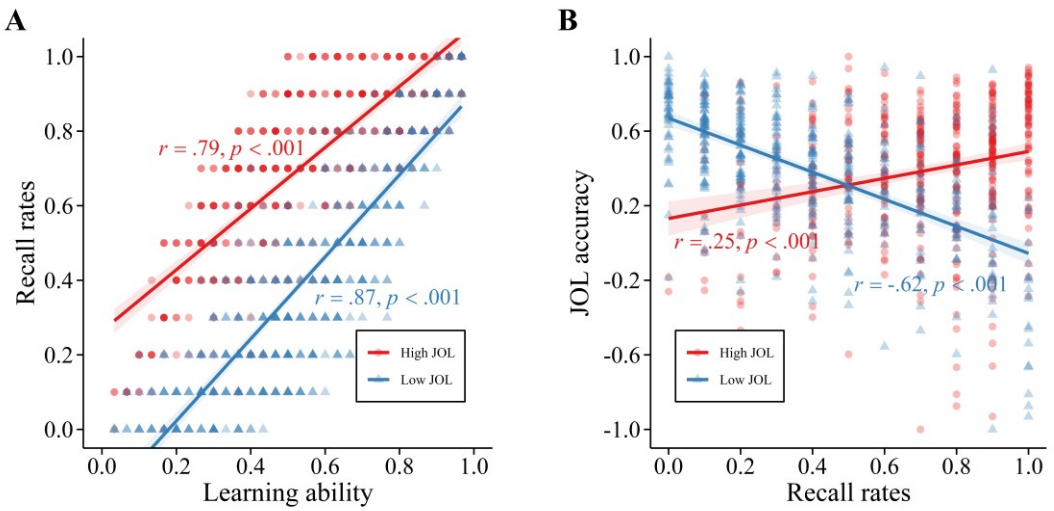



Figure 11 provides a visual representation of the mediation effects. As shown in Figure 11A, learning ability positively predicted recall rates of both low and high JOL items, and these prediction effects did not differ greatly. However, as shown in Figure 11B, successfully remembering low JOL items substantially reduced JOL accuracy. By contrast, successfully remembering high JOL only slightly improved JOL accuracy. Therefore, the total relation between learning ability and JOL accuracy was negative.

Figure 11. Scatter plot depicting the relations among different variables in Study 4

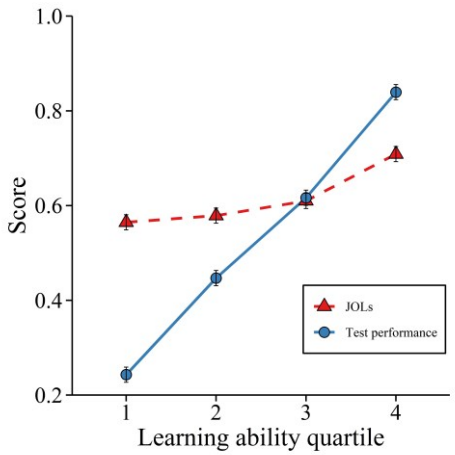



*Note*: Panel A depicts the relations between learning ability (indexed by test performance) and recall rates of low and high JOL items; Panel B illustrates the relations between JOL accuracy

(indexed by *G*) and recall rates of low and high JOL items. Error bars represent 95% CI.

Finally, we re-analyzed the data using the same methods as in Study 3 to investigate the DK effect. As shown in Figure 12, a Bayesian mixed ANOVA revealed a substantial interaction between measurement type (JOLs vs. test performance) and group (1 vs. 2 vs. 3 vs. 4), $F(3, 477) = 267.87$, $p < .001$, $\eta_p^2 = .63$, $BF_{10} > 1{,}000$, with Group 1 (i.e., participants with lowest test performance) showing substantial overestimation of test performance, difference between JOLs and test performance = 0.32, 95% CI = [0.29, 0.35], $t(119) = 23.69$, $p < .001$, Cohen's $d = 2.16$, $BF_{10} > 1{,}000$, and Group 4 (i.e., participants with highest test performance) exhibiting underestimation, difference between JOLs and test performance = -0.13, 95% CI = [-0.15, -0.11], $t(119) = -11.87$, $p < .001$, Cohen's $d = -1.08$, $BF_{10} > 1{,}000$. Importantly, the degree of overestimation in Group 1 was much stronger than the degree of underestimation in Group 4. These results again successfully replicate the DK effect.

Figure 12. Line plot depicting the DK effect in Study 4




*Note*: Participants were divided into quartiles according to their test performance. Error bars represent 95% CI.

## General Discussion

The present research documents a small size of negative relation between learning ability and (relative) JOL accuracy. Furthermore, the meta-analysis in Study 1 shows that this negative relation holds across different types of study materials, test formats, age groups,

countries/social cultures, and learning tasks of varying difficulty, although it is worth noting that this meta-analysis was not comprehensive. These findings unveil a paradox that challenges a long-standing assumption in the fields of learning and metacognition: that stronger learners are naturally more adept at accurately assessing their learning progress (Griffin et al., 2008; Nietfeld & Schraw, 2002).

What causes this negative relation and misalignment between learning ability and monitoring accuracy? One plausible explanation, which we term "expert underconfidence," posits that high-ability learners tend to underestimate their retention capabilities, particularly for challenging material they judge as non-memorable (Witherby et al., 2023). This underconfidence introduces a bias in their JOLs, causing them to underestimate what they are likely to remember and overestimate what they might forget. These findings suggest that high learning ability and accurate self-monitoring are distinct skills that do not necessarily develop together, and being a good learner does not necessarily translate into having high monitoring accuracy (Vlach et al., 2019).

The expert underconfidence hypothesis offers a compelling explanation for the present findings, but it may not be the sole factor at play. It is also possible that expert learners rely on different cues to assess their learning progress (Koriat, 1997). Rather than focusing on simple or superficial cues, they may engage in deeper processing, which, while beneficial for learning, could disrupt the straightforward assessment of learning progress (Van Gog et al., 2011). Their cognitive resources may be focused on encoding the information rather than on monitoring their learning progress (Bryce et al., 2023). Effectively this is a problem of resource allocation. Suggestive evidence supporting this explanation comes from Bryce et al. (2023) and Li et al. (2024), which showed that monitoring accuracy is compromised when cognitive resources are limited, and multitasking impairs JOL accuracy. Direct tests of this explanation are called for.

While teachers, parents, and educational policymakers often operate under the assumption that high-performing students require less support in learning how to learn (Grünke, 2006), our findings suggest that such students, although excelling in acquiring knowledge, may still benefit from metacognitive support. Specifically, they may need assistance in enhancing their awareness of what they know and do not know. However, the

current data also suggest a potential trade-off: the underconfidence exhibited by expert learners may serve an adaptive strategy. Specifically, their underconfidence in challenging material may prompt them to allocate additional resources for encoding (Yang et al., 2017). Therefore, while metacognitive support (e.g., metacognitive training) could help raise their monitoring accuracy (Handel et al., 2020), it may concurrently disrupt their adaptive learning behaviors. Further research is needed to determine whether enhancing expert learners' monitoring accuracy would truly improve, or hinder, their learning performance. These findings, which were derived primarily from Chinese participants, should be generalized to other populations with caution. Future research could also profitably explore the neural mechanisms underlying the dissociation between learning ability and monitoring capacity. For instance, neuroimaging studies could examine whether specific brain regions associated with metacognition, such as the anterior prefrontal cortex (Fleming et al., 2014), function differently in high- and low-ability learners.

In summary, our findings reveal a novel, important, yet overlooked aspect of self-regulated learning: high-ability learners are not always accurate judges of the strength of their learning. The negative relation between learning ability and monitoring accuracy challenges conventional wisdom, suggesting that high learning ability does not guarantee accurate self-monitoring. The expert underconfidence hypothesis offers a potential explanation for this misalignment.

## References

*References marked with * indicate studies included in Study 1's meta-analysis*

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, *68*(5), 804–825. https://doi.org/10.1037/0022-3514.68.5.804

Argote, L. (2013). *Organizational learning: Creating, retaining, and transferring knowledge*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-5251-5

*Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta Psychologica*, *197*, 153–165. https://doi.org/10.1016/j.actpsy.2019.04.011

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markham (Eds.), *Handbook of child psychology: Vol. 3. Cognitive development* (pp. 77–166). Wiley.

Bryce, D., Kattner, F., Birngruber, T., & Wellingerhof, P. (2023). Monitoring accuracy suffers when working memory demands increase: Evidence of a dependent relationship. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(12), 1909–1922. https://doi.org/10.1037/xlm0001262

*Chang, M., & Brainerd, C. J. (2023). Changed-goal or cue-strengthening? Examining the reactivity of judgments of learning with the dual-retrieval model. *Metacognition and Learning*, *18*(1), 183–217. https://doi.org/10.1007/s11409-022-09321-y

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. https://doi.org/10.1037/a0032968

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1-12. https://doi.org/10.3758/s13428-014-0458-y

*Double, K. S. (2023). Do judgments of learning impair recall when uninformative cues are

salient? *Journal of Intelligence*, *11*(10), 203. https://doi.org/10.3390/jintelligence11100203

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage Publications.

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology*, *44*, 247–296. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

*Fan, T., Zheng, J., Hu, X., Su, N., Yin, Y., Yang, C., & Luo, L. (2021). The contribution of metamemory beliefs to the font size effect on judgments of learning: Is word frequency a moderating factor? *PLOS ONE*, *16*(9), e0257547. https://doi.org/10.1371/journal.pone.0257547

Fan, T., Zhao, W., Sun, B., Liu, S., Yin, Y., Xu, M., Hu, X., Yang, C., & Luo, L. (2025). A normative database of Swahili–Chinese paired associates. *Behavior Research Methods*, *57*(1), 1-14. https://doi.org/10.3758/s13428-024-02531-z

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Flavell, J. (1976). Metacognitive aspects of problem solving. *The Nature of Intelligence*, *12*, 231–235.

Flavell, J. H. (1981). Cognitive monitoring. In W. P. Dickson (Ed.), *Children's oral communication skills* (pp. 35–60). Academic Press.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137*(10), 2811–2822. https://doi.org/10.1093/brain/awu221

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence*, *80*, 101449. https://doi.org/10.1016/j.intell.2020.101449

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, *55*(292), 708-713.

https://doi.org/10.1080/01621459.1960.10483369

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*(1), 93–103. https://doi.org/10.3758/mc.36.1.93

Grünke, M. (2006). Zur Effektivität von Fördermethoden bei Kindern und Jugendlichen mit Lernstörungen: Eine Synopse vorliegender Metaanalysen. *Kindheit und Entwicklung*, *15*(4), 239–254. https://doi.org/10.1026/0942-5403.15.4.239

Handel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing. *Learning and Instruction*, *65*, Article 101245. https://doi.org/10.1016/j.learninstruc.2019.101245

Hartwig, M. K., Was, C. A., Isaacson, R. M., & Dunlosky, J. (2012). General knowledge monitoring as a predictor of in-class exam performance. *British Journal of Educational Psychology*, *82*(3), 456–468. https://doi.org/10.1111/j.2044-8279.2011.02038.x

Hasselhorn, M., & Hager, W. (1989). Prediction accuracy and memory performance: Correlational and experimental tests of a metamemory hypothesis. *Psychological Research*, *51*(3), 147–152. https://doi.org/10.1007/Bf00309310

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*(6), 756–763. https://doi.org/10.1038/s41562-021-01057-0

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, *8*(2), 115. https://doi.org/10.1037/1082-989x.8.2.115

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance.

*Journal of Personality and Social Psychology*, *82*(2), 180–188. https://doi.org/10.1037/0022-3514.82.2.180

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Li, S., Liu, Y., Jing, A., & Wang, Y. (2024). The effects of multitasking on metacognitive monitoring in primary and secondary school students. *Journal of Experimental Child Psychology*, *242*, 105908. https://doi.org/10.1016/j.jecp.2024.105908

Mendes, P. S., Luna, K., & Albuquerque, P. B. (2019). Word frequency effects on judgments of learning: More than just beliefs. *The Journal of General Psychology*, *148*(2), 124–148. https://doi.org/10.1080/00221309.2019.1706073

*Mendes, P. S., & Undorf, M. (2022). On the pervasive effect of word frequency in metamemory. *Quarterly Journal of Experimental Psychology*, *75*(8), 1411–1427. https://doi.org/10.1177/17470218211053329

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition*, *48*(5), 745–758. https://doi.org/10.3758/s13421-020-01025-5

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, *26*, 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research*, *95*(3), 131–142. https://doi.org/10.1080/00220670209596583

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51(1)*, 195-203. https://doi.org/10.3758/s13428-018-01193-y

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdogan, B., Arbuzova, P., Atlas, L. Y., Balci, F., Bang, J. W., Begue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., . . . Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

*Rivers, M. L., Janes, J. L., Dunlosky, J., Witherby, A. E., & Tauber, S. K. (2023a). Exploring the role of attentional reorienting in the reactive effects of judgments of learning on memory performance. *Journal of Intelligence*, *11*(8), 164. https://doi.org/10.3390/jintelligence11080164

*Rivers, M. L., Dunlosky, J., Janes, J. L., Witherby, A. E., & Tauber, S. K. (2023b). Judgments of learning enhance recall for category-cued but not letter-cued items. *Memory & Cognition*, *51*(7), 1547–1561. https://doi.org/10.3758/s13421-023-01417-3

Robey, A. M., Dougherty, M. R., & Buttaccio, D. R. (2017). Making retrospective confidence judgments improves learners' ability to decide what not to study. *Psychological Science*, *28*(11), 1683–1693. https://doi.org/10.1177/0956797617718800

Siegel, A. L. M., & Castel, A. D. (2019). Age-related differences in metacognition for memory capacity and selectivity. *Memory*, *27*(9), 1236-1249. https://doi.org/10.1080/09658211.2019.1645859

Small, S. A., Stern, Y., Tang, M., & Mayeux, R. (1999). Selective decline in memory function among healthy elderly. *Neurology*, *52*(7), 1392. https://doi.org/10.1212/WNL.52.7.1392

Smith, F. X., & Was, C. A. (2019). Knowledge monitoring calibration: Individual differences in sensitivity and specificity as predictors of academic achievement. *Educational Sciences: Theory and Practice*, *19*(4), 80–87. https://doi.org/10.12738/estp.2019.4.006

Soares, J. S., & Storm, B. C. (2022). Does taking multiple photos lead to a photo-taking-impairment effect? *Psychonomic Bulletin & Review*, *29*(6), 2211–2218. https://doi.org/10.3758/s13423-022-02149-2

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical*

*Methods in Medical Research*, *6*(2), 103–114. https://doi.org/10.1177/096228029700600202

*Tauber, S. K., & Dunlosky, J. (2012). Can older adults accurately judge their learning of emotional information? *Psychology and Aging*, *27*(4), 924–933. https://doi.org/10.1037/a0028447

*Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology and Aging*, *34*(6), 836–847. https://doi.org/10.1037/pag0000376

Touron, D. R., Oransky, N., Meier, M. E., & Hines, J. C. (2010). Metacognitive monitoring and strategic behaviour in working memory performance. *Quarterly Journal of Experimental Psychology*, *63*(8), 1533–1551. https://doi.org/10.1080/17470210903418937

*Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition*, *46*(4), 507–519. https://doi.org/10.3758/s13421-017-0780-6

Undorf, M., & Bröder, A. (2021). Metamemory for pictures of naturalistic scenes: Assessment of accuracy and cue utilization. *Memory & Cognition*, *49*(7), 1405–1422. https://doi.org/10.3758/s13421-021-01170-5

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Applied Cognitive Psychology*, *25*(4), 584–587. https://doi.org/10.1002/acp.1726

Vlach, H. A., Bredemann, C. A., & Kraft, C. (2019). To mass or space? Young children do not possess adults' incorrect biases about spaced learning. *Journal of Experimental Child Psychology*, *183*, 115–133. https://doi.org/10.1016/j.jecp.2019.02.003

Vuorre, M., & Metcalfe, J. (2022). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning*, *17*(2), 269–291. https://doi.org/10.1007/s11409-020-09257-1

*Witherby, A. E., Tauber, S. K., & Goodrich, M. (2022). People hold mood-congruent beliefs about memory but do not use these beliefs when monitoring their learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(4), 499–519.

https://doi.org/10.1037/xlm0001096

Witherby, A. E., Carpenter, S. K., & Smith, A. M. (2023). Exploring the relationship between prior knowledge and metacognitive monitoring accuracy. *Metacognition and Learning*, 1–31. https://doi.org/10.1007/s11409-023-09344-z

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1073. https://doi.org/10.1037/xlm0000363

*Yin, Y., Li, B., Hu, X., Guo, X., Yang, C., & Luo, L. (2023). The relationship between dispositional mindfulness and relative accuracy of judgments of learning: The moderating role of test anxiety. *Journal of Intelligence*, *11*(7), 132. https://doi.org/10.3390/jintelligence11070132

*Zimdahl, M. F., & Undorf, M. (2021). Hindsight bias in metamemory: Outcome knowledge influences the recollection of judgments of learning. *Memory*, *29*(5), 559–572. https://doi.org/10.1080/09658211.2021.1919144

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2