Cue Learning in Metamemory: Understanding How People Learn to Judge Memory

Sofia Navarro Báez

Inaugural Dissertation

Submitted in partial fulfillment of the requirements for the degree Doctor of Social
Sciences in the Graduate School of Economic and Social Sciences at
the University of Mannheim

Thesis Defense:

03.09.2024

Dean of the School of Social Sciences:

Prof. Dr. Michael Diehl

Supervisors

Prof. Dr. Arndt Bröder

Prof. Dr. Monika Undorf

Thesis Reviewers:

Prof. Dr. Edgar Erdfelder

Prof. Dr. Beatrice G. Kuhlmann

In loving memory of my father, who showed me how beautiful
it is to wonder about ourselves and the world around us.

Contents

Summary

Metacognition is the remarkable human ability to know and to think about one's own cognition (Flavell, 1979). A large part of metacognition research has been conducted in the domain of memory where people's assessments of their learning and memory are obtained via metamemory judgments. Inferential accounts of metacognition state that metamemory judgments such as judgments of learning (JOLs) – predictions of future memory — are inferences based on cues and heuristics (Koriat, 1997). Since around the last 30 years, researchers have uncovered many cues that underlie metamemory judgments, most of which are valid and lead to accurate metamemory, but some of which are invalid and lead to illusions. Further, researchers have focused on understanding whether cues are directly and/or indirectly used for metamemory judgments through beliefs and/or feelings, respectively. But little is known about how cues informing metamemory judgments are learned or acquired.

The goal of this thesis is to investigate and understand some of the mechanisms through which cues utilized for judging memory are learned. To this end, I used various experimental learning paradigms soliciting metamemory judgments. Across the three studies within this thesis, I found that experience with one's own learning and testing is beneficial for judging the general memorability of pictures by increasing the judgment sensitivity to valid cues (Manuscript I). However, neither experiences across multiple study-test cycles nor cognitive feedback on the recall status and JOL given to each studied item are effective for learning cue validities in judgments for one's own memory, JOLs. Rather, cue validities are learned through informative explanations about metacognition that lead to a deeper understanding of the cues (Manuscript II). Moreover, in a first demonstration of statistical learning influencing metacognition, I showed that regularities are extracted from experience with the environment and then used to inform JOLs. Since regularities in the environment are abundant, demonstrating that cues are learned via statistical learning mechanisms has relevant implications for real world learning (Manuscript III).

With this, I showed when and which experiences are helpful for learning cues, what is effective in alleviating metacognitive illusions, and how to learn cues from the environment. Overall, the three manuscripts in this thesis contribute to our understanding of how cues for metamemory judgments are learned and pave the way for future research on cue learning in metamemory.

Manuscripts

The research in this thesis was conducted at the *Center for Doctoral Studies in Social and Behavioral Sciences* (CDSS) of the *Graduate School of Economic and Social Sciences* (GESS) at the University of Mannheim. It is based on three manuscripts, one of which has been published and two of which have been submitted for publication.

In the three manuscripts, I investigate the question of how cues for metamemory judgments are learned. I found that cue knowledge to judge the generic attribute of memorability in pictures is acquired from experiences with one's own learning and testing (Manuscript I). I also showed that experiences across multiple study-tests cycles and cognitive feedback on the recall status and JOL given to each studied item is not sufficient for learning cue validities, but rather an in-depth explanation of metacognition is needed (Manuscript II). Finally, I demonstrated that cues that inform metamemory judgments are extracted from the environment and learned via statistical learning mechanisms (Manuscript III). In the following chapters, I provide theoretical foundations, manuscript summaries, and an integrative discussion of the manuscripts' results. Details on the experimental paradigms and statistical analyses can be found in the original manuscripts appended to this thesis.

Manuscript I

Navarro-Báez, S., Undorf, M., & Bröder, A. (2024). Predicting the memorability of scene pictures: Improved accuracy through one's own experience. *Quarterly Journal of Experimental Psychology*, *0*(0), 1-20. https://doi.org/10.1177/17470218241239829

Manuscript II

Navarro-Báez, S., Bröder, A., & Undorf, M. (2024). *Mending metacognitive illusions requires metacognitive feedback*. [Manuscript submitted for publication]. Department of Psychology, University of Mannheim. Department of Psychology, Technical University of Darmstadt.

Manuscript III

Navarro-Báez, S., Bröder, A., & Undorf, M. (2024). *Detecting structure: Cues for metacognitive judgments are acquired via statistical learning*. [Manuscript submitted for publication]. Department of Psychology, University of Mannheim. Department of Psychology, Technical University of Darmstadt.

1. Introduction

The human mind has the amazing cognitive ability to reflect upon itself. This allows us to reflect on our own thoughts. Over the centuries, abilities such as self-observation and introspection have been used as methods to gain knowledge about the self and the mind. Since the cognitive revolution of the 1960's, modern psychology has focused on explicitly studying metacognition – not as a methodological tool for studying the mind, but as a mental operation per se.

Flavell (1979) was the first to define *metacognition* as knowledge and cognition about cognitive phenomena. This encompasses our knowledge, thoughts, and evaluations about how we can learn more effectively and remember information better, which are highly relevant in educational and professional settings. Consequently, metacognition research has mainly developed in the domain of memory. *Metamemory* refers to the ability to know and to assess memory (Dunlosky & Thiede, 2013). Researchers typically measure people's assessments of their learning and memory by asking them to make metamemory judgments. Relevant to this thesis, *judgments of learning* (JOLs) are common prospective metamemory judgments about the likelihood of remembering recently studied items on an upcoming test (Dunlosky & Metcalfe, 2009). For example, a student predicts how likely it is that she will remember the definition of epigenetics on an upcoming exam. In Nelson and Narens' (1990) classic conceptual framework of metamemory, metamemory operates at a meta-level, interconnected with object-level cognitive processes like encoding and retrieval, through monitoring and controlling these object-level activities. Thus, in the previous example, the student would not only make a prediction about her memory (monitoring) but also use the output of her prediction to regulate her memory by deciding to allocate further study time to the topic of epigenetics (control) if the prediction was low. People also engage in metamemory in everyday life, for instance, when making a list to prevent forgetting something. These examples illustrate how monitoring is a guide for behavior. Importantly, the causal link between monitoring and control has now been demonstrated by several studies showing that items that participants feel the least confident of remembering are chosen to restudy, memory performance is better if monitoring and regulation are accurate (Metcalfe & Finn, 2008; Rhodes & Castel, 2009; Thiede et al., 2003; Tullis & Benjamin, 2012).

How do people monitor their memory?

Initial theories of metamemory suggested that people have direct access to the varying strengths of the memory traces formed during encoding, and thus, they can transform them into accurate recall probability ratings to predict memory (Cohen et al., 1991; Hart, 1967; Schwartz, 1994). The direct-access view predicts a strong correspondence between metamemory and memory given that both processes are based on the same underlying factor (i.e., memory trace strength). However, this is at odds with what psychology had already demonstrated that people cannot directly observe their cognitive processes as introspection does not produce an accurate picture of the mind (Nisbett & Wilson, 1977). In a seminal study, Koriat (1997) found unequivocal evidence against the direct-access hypothesis: whereas the effect of a factor inherent to the items (i.e., word pair relatedness) was equal on JOLs and the recall criterion, the effect of other extrinsic factor (i.e., study repetition) was stronger on recall and discounted on JOLs. Since then, vast evidence on systematic dissociations between predicted and actual recall have favored inferential accounts of metamemory (see Undorf et al., 2022, for a review). According to the cue-utilization approach (Koriat, 1997), JOLs are inferences based on cues and heuristics about the likelihood of remembering items. A cue can be any stimulus or characteristic that is intrinsic or extrinsic to the materials learned, such as word pair relatedness or number of study trials, respectively. Cues can also be internal to the learner such as the experience of 'ease' during learning (Undorf & Erdfelder, 2011), or idiosyncratic such as the personal significance of to-be-studied materials (Undorf et al., 2022). In this way, people use whichever cues are available to them for informing and making JOLs (Undorf et al., 2018). The accuracy of JOLs will be largely determined by the extent to which the cues used to make JOLs correlate with actual memory performance.

How accurate is metamemory?

Many cues predictive of memory performance have been found to underlie JOLs. For example JOLs are higher for related word pairs such as '*nest – bird*' than for unrelated ones such as '*pineapple – monkey*' (e.g., Arbuckle & Cuddy, 1969), for concrete words such as '*mouse*' than abstract ones such as '*phase*' (e.g., Witherby & Tauber, 2017), and for emotional words such as '*love'* than non-emotional ones such as '*axis*' (e.g., Zimmerman & Kelley, 2010). However, JOLs have also been found to rely on invalid cues like the large font size in which words are displayed (Rhodes & Castel, 2008), or the loudness of words (Frank & Kuhlmann, 2017; Mueller et al., 2014). JOLs can also ignore valid cues like the number of future study opportunities (Kornell & Bjork, 2009). This indicates that although people have a good idea about which cues are predictive of memory and their JOLs are sensitive to those cues, their JOLs also ignore or reflect erroneous ideas about the predictive value of certain qualities, situations, and experiences. Having said that, *relative accuracy,* the degree to which item-by-item JOLs correlate with actual memory performance for each item, is often moderate. But it can suffer when JOLs rely on invalid cues or fail to rely on valid cues (Undorf, et al., 2022).

Methods to improve monitoring

In general, the literature provides two methods to enhance the relative accuracy of JOLs, both of which rely on retrieval practice (Dunlosky & Metcalfe, 2009; Rhodes, 2016). The first method is testing practice, where retrieval success or failure during testing serves as a basis for subsequent memory predictions (Finn & Metcalfe, 2007, 2008). This greatly improves relative accuracy because current retrieval success is often predictive of performance on a future test (Dougherty et al., 2005). The second method that produces significant improvements in relative accuracy is delaying JOLs instead of prompting them immediately (Nelson & Dunlosky, 1991). The delayed-JOL-effect is explained by participants relying on long-term retrieval success or failure, which becomes diagnostic some time after study, rather than immediately when item information is still accessible in working memory (Dunlosky & Nelson, 1992). Another compatible explanation is that successful retrieval boosts final recall performance due to retrieval practice which ensures judgment predictive accuracy (Spellman & Bjork, 1992). In conclusion, both testing practice and delaying JOLs are robust methods that enhance the relative accuracy of JOLs by relying on previous retrieval attempts. One caveat of testing practice is that when JOLs are required for new materials, accuracy remains unaffected because reliance on previous test performance is impossible. Moreover, both methods have the limitation that judgment sensitivity to valid cues, which is a determinant of accuracy, is not enhanced immediately during learning. Thus, ways to enhance monitoring accuracy per se in a first study trial are missing.

The goal of this thesis

Uncovering ways to enhance metacognitive sensitivity to valid cues immediately during the learning of new materials can be a powerful tool for effectively training metacognition. However, little is known about how to increase cue sensitivity of metamemory judgments, and about how people learn or discover new information that they can rely on when making metamemory judgments. Therefore, the goal of this thesis is to understand *how cues that inform metamemory are learned.* Specifically, in Manuscript I, I systematically compared the cue basis and relative accuracy of two different types of metamemory judgments, predictions of one's own later memory performance (JOLs) and predictions of generic item memorability (memorability judgments, MJs).

This systematic comparison demonstrated that both metamemory judgments have similar cue bases and relative accuracy. However, personal learning and testing experiences improve the relative accuracy of MJs by enhancing sensitivity to valid cues. In Manuscript II, in the context of metacognitive illusions, I tested different forms of feedback to enhance JOL sensitivity to valid cues but underweighted by JOLs (i.e., number of future study opportunities) and invalid cues but overweighted by JOLs (i.e., font size, font format). Results showed that people do not learn to appropriately weight cues on their JOLs from cognitive feedback about one's own recall performance and JOL for each studied item. In contrast, additional metacognitive feedback that includes information about actual task performance at the population level, the functioning of metacognition, and ways to improve it increases JOL reliance on the number of future study opportunities, thereby enhancing relative accuracy. In Manuscript III, I tested statistical learning as a mechanism to extract regularities from the environment without explicit instruction, or intention to learn. This demonstrated that such regularities learned through statistical learning are used as cues to inform JOLs even if those are not necessarily predictive of actual memory performance. Overall, this expands our knowledge about metacognition and opens the door for further research on cue learning, a topic that has been largely overlooked in metacognition research.

In Chapter 2, I first give a more extended overview on the historical origins of metamemory research. Second, I expand on the cue-utilization approach to JOLs. Third, relevant to this thesis, I present metamemory accuracy measures, and metacognitive illusions. Finally, I review two existing cue learning methods in the literature.

In Chapter 3, I approach *how people learn cues to inform their metamemory judgments* through three different angles; learning about item memorability from one's own experience (Manuscript I), learning and unlearning cues from feedback (Manuscript II), and learning new cues extracted from the environment (Manuscript III). With this, I contribute to three important specific questions in the field of metamemory: 1) how people learn to judge memorability as general attribute of an item, 2) how to alleviate metacognitive illusions, and 3) how to learn new cues from the environment.

In Chapter 4, I conclude this thesis by integrating the implications of each study manuscript, by clarifying open questions about whether cue learning occurs from experience or feedback, and by examining the nature of the cue content learned. I also outline future research directions regarding the relationship between monitoring and control, as well as the effectiveness of cue learning paradigms

2. Theoretical Foundations

2.1 Historical Origins

The origin of experimental research in metamemory can be traced back to the 1960's with the first systematic investigations on feelings of knowing judgments (see Hart, 1965, 1967). Hart was interested in situations where people cannot directly answer a question, but report the feeling of knowing the answer and having it at the tip of their tongue. Specifically, he wanted to know whether those tips-of-the-tongue experiences are accurate indicators of what is stored in memory. To this end, he developed a paradigm to produce such experiences and evaluate their accuracy. The recall-judgment-recognition paradigm consist of a recall test, and a multiple-choice recognition test. In the recall test, subjects are tested on basic facts (e.g., *Which is the largest planet in our solar system?*) and they are asked to recall the correct answer. If they cannot recall the answer, they are asked to make a binary judgment about correctly recognizing the correct answer among four alternatives in a multiple-choice test (e.g., *Pluto, Venus, Earth, Jupiter*). Results showed that the judgments can discriminate with above-chance accuracy which items will later be recognized and which ones will not. The finding that people can accurately monitor their memory was crucial for the development of metacognition research. Hart interpreted his finding as evidence that people have direct access to the content of their memory. Importantly, however, this study paved the way to develop paradigms that test the validity of people's judgments about their own cognition instead of taking accuracy for granted.

Later, Arbuckle and Cuddy (1969) introduced prospective judgments about memory made during learning (i.e., JOLs). Specifically, their study tested whether people are sensitive to differences in the associative strength of two items. For this, they prompted participants during learning to predict future recall. Results showed that participants' predictions were sensitive to differences in associative strength among items, and therefore, exceeded the criterion of chance performance. In their study, Arbuckle and Cuddy (1969) suggested to look for other possible factors that affect predictions about future recall without affecting actual memory performance. However, this agenda waited for a while since researchers focused on examining the development of metacognition in children (Flavell, 1979).

The next important step in metacognition research was made by Flavell et al. (1970) who did not focus on examining basic metacognitive processes (e.g., direct-access accounts), but rather focused on the development of metacognition across childhood. In their study, they examined children's ability to predict future recall and spontaneous use of learning strategies across different age groups (i.e., nursery school, kindergarden, second graders, and fourth graders). Specifically, children were asked to study a sequence of pictures and make global predictions about the maximum number of pictures that they could correctly recall. In another task, children were asked to study pictures in their own self-paced time to achieve perfect recall while the experimenter was observing their study behavior. Results showed that all groups of children were overconfident in their predictions, but less overconfidence was observed in the two oldest age groups. In addition, more effective learning strategies were applied by the oldest children.

These results were important for Flavell to develop his model of cognitive monitoring with four components (Flavell, 1979): 1) metacognitive knowledge, 2) metacognitive experiences, 3) goals (or tasks), and 4) actions (or strategies). Metacognitive knowledge refers to stored knowledge and beliefs about how our cognition operates (e.g., *I am better at literature than at math*). Metacognitive experiences refer to phenomological cues that arise during learning such as feelings of ease. Goals (or tasks) refer to the goal that one desires to attain with a particular cognitive enterprise, and actions (or strategies) refer to the specific cognitions and self-regulation strategies that one performs to achieve a goal. Flavell's theoretical model provided the first stepping stones for examining the relation between monitoring and regulation of cognition. This opened research on metacognition in educational contexts (Brown et al., 1983). At the same time, the first investigations about memory beliefs were conducted in children (Borkowski et al., 1983; Kurtz & Borkowski, 1987), and questionnaires for evaluating people's knowledge about memory, frequency of monitoring, and control strategies were designed (Gilewski et al., 1990).

The next studies that solicited item-by-item memory predictions made during learning (i.e., JOLs) after Arbuckle and Cuddy (1969) were the studies by Groninger (1976, 1979). Subjects studied a list of words that varied in their word characteristics (concreteness, emotionality, nonsense). Results showed that words which received higher JOLs were more likely to be recognized than words which received lower JOLs. Crucially, Groninger (1976, 1979) interpreted the findings as subjects utilizing word attributes that are predictors of performance in their judgments, thereby moving away from direct-access interpretations. Afterwards, the research on JOLs continued with important study findings such as previous recall performance in a first trial being more predictive of recall performance in a second trial than second-trial-JOLs (King et al., 1980). This study was important for the future line of research on the memory-for-past-test heuristic, which refers to basing memory predictions on previous memory performance when materials are studied repeatedly across trials (Finn & Metcalfe, 2007, 2008).

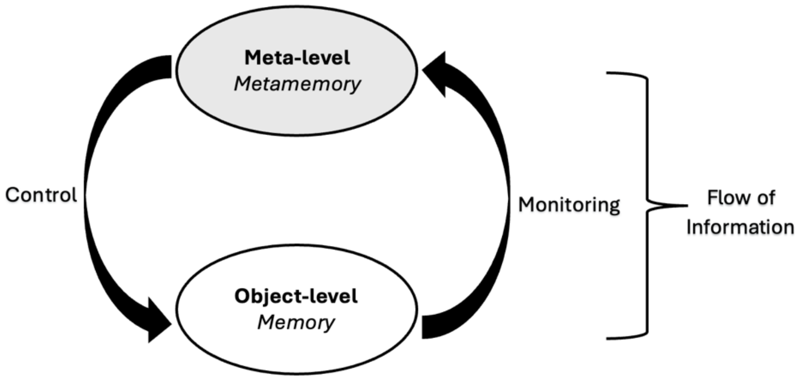
Finally, Nelson and Narens (1990) provided a theoretical framework for metamemory which was an enormous contribution to the metamemory field promoting its rapid development.

The framework is built upon the idea that there are two independent levels: the object-level and the meta-level. Memory corresponds to the object-level and metamemory corresponds to the meta-level (see Figure 1). The two levels are interrelated via the flowing of information to form two processes called *monitoring* and *control.* Monitoring is the information that flows from the object-level to the meta-level. This information changes the state of the meta-level by creating a representation of the object-level, in other words, creating a dynamic model of memory. This model is the content of metamemory. To assess monitoring, participant's introspective reports of their own cognition are solicited. Control is the information that flows from the meta-level to the object-level. This information changes the state of the object-level by producing an action of initiation, continuation, or termination (e.g., a student decides to start, to continue, or to stop studying).

Nelson and Narens also examined the degree to which people have direct (or privileged) access to their own memories. They concluded that feeling-of-knowing judgments do not monitor currently inaccessible items but rather information that is related to the item and speaks about its future retrievability (e.g., *I have recalled this item in many previous occasions*).

Figure 1

Nelson and Narens' (1990) theoretical framework of metamemory

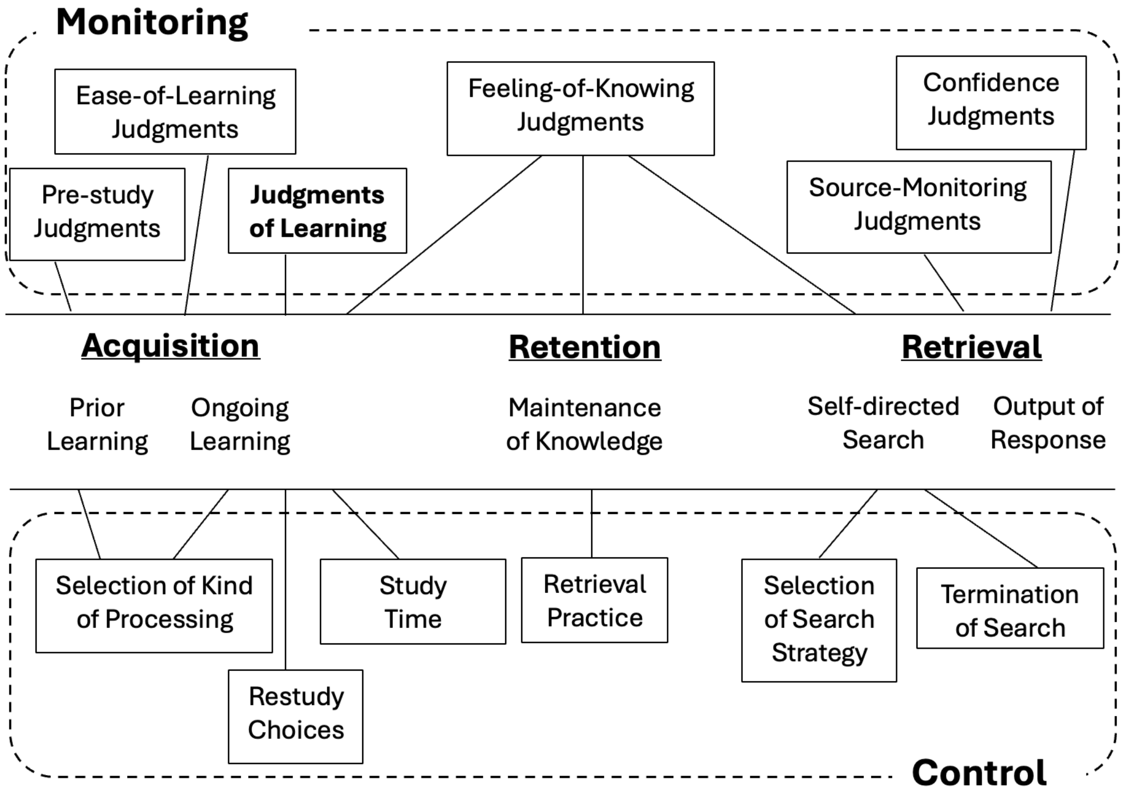


Note. The framework consists of two structures (meta-level and object-level) and two relations (monitoring and control) in terms of the direction of the flow of information between the two levels. The meta-level (metamemory) contains a dynamic model of the object-level (memory).

Furthermore, Nelson and Narens' theoretical framework categorized some examples of monitoring and control judgments based on the time in which they are made during the memory process; acquisition, retention, and retrieval (see Figure 2). Research has now shown that monitoring judgments rely on different qualitative information depending on the time in which they are solicited (Busey et al., 2000; Dougherty et al., 2005), and that monitoring may be composed of two distinct abilities, one occurring during learning and one during retrieval (McDonough et al., 2021). As previously mentioned, this thesis focuses on monitoring judgments of learning (JOLs) made during the acquisition phase to predict future test performance of recently acquired items. Importantly, this thesis also expands the metamemory framework by examining judgments about the generic memorability of items, not made during learning but tapping into people's metacognitive models of item memorability.

Figure 2

Categorization of metamemory judgments in the theoretical metamemory framework by Nelson and Narens (1990)



Note. Main stages of memory (acquisition, retention, retrieval) and some examples of monitoring judgments (top) and control judgments (bottom) linked to the stages of memory in which the judgments are made. This thesis focuses on judgments of learning monitoring ongoing learning.

2.2 Cue-utilization Approach to JOLs

Koriat (1997) put forward the cue-utilization approach to JOLs. This approach explained several findings in the literature regarding the effect of different variables (e.g., word concreteness, word frequency, encoding strategies, number of study trials) on both the judgment and the memory criterion or just one (Begg et al., 1989; Mazzoni & Cornoldi, 1993; Zechmeister & Shaughnessy, 1980). The cue-utilization approach assumes that JOLs as other metamemory judgments are inferential in nature. This means that participants do not directly monitor the memory traces of items being learned, but rather use heuristics or cue information to infer future recall. Specifically, Koriat (1997) tested whether cue effects on predicted and actual recall would be similar or different as well as the sensitivity of JOLs to different cues. Results showed that JOLs followed the same pattern as recall performance for word pair relatedness (i.e., higher for related pairs than unrelated ones). However, JOLs differed from recall performance as they did not predict better memory for word pairs learned in two trials than in one trial only. This resulted in underconfident JOLs for word pairs learned across two trials, an effect called the underconfidence-with-practice (UWP) effect. The finding that there are variables affecting predicted and actual recall differently provided unequivocal evidence for the cue-utilization approach. Furthermore, Koriat's cue-utilization approach made two other important contributions: 1) providing a taxonomy for cues, and 2) suggesting two inferential routes by which cues influence JOLs.

First, the taxonomy differentiates among three main classes of cues: extrinsic, intrinsic, and mnemonic. *Extrinsic cues* are related to the conditions in which items are learned or the encoding strategies employed for their learning, such as the number of study repetitions (e.g., Kornell & Bjork, 2009). *Intrinsic cues* are features inherent to the study items, such as word pair relatedness (e.g., Undorf & Erdfelder, 2015), or word concreteness (e.g., Witherby & Tauber, 2017). *Mnemonic cues* refer to internal personal cues such as the subjective experience of ease that arises during learning, i.e., fluency (Alter & Oppenheimer, 2009). Importantly, mnemonic cues are sensitive to intrinsic and extrinsic cues. This is because experiences of ease during learning can arise from other underlying cues such as word pair relatedness.

Second, the cue-utilization approach specifies two inferential routes how cues influence judgments. One route is theory-based or analytic: people deliberately apply a rule or a theory about a cue. For example, people believe that concrete words are more likely to be remembered than abstract words, and thus, make higher JOLs for concrete words than abstract ones (Witherby & Tauber, 2017). It is suggested that extrinsic and intrinsic cues affect JOLs directly through the theory-based route. Another route is experience-based or non-analytic: people implicitly use intrinsic or extrinsic cues via mnemonic cues (i.e., fluency) to inform their JOLs, without necessarily being aware of the actual cue that is producing fluency feelings (Koriat & Levy-Sadot, 1999). For example, people make higher JOLs for items written with their dominant hand than non-dominant hand because writing with the dominant hand elicit fluency experiences even when there is no belief that the writing hand impact memory performance (Susser et al., 2017; Susser & Mulligan, 2015).

Many metamemory studies have investigated the relative contributions of theory- and experience-based cues to JOLs. Thus, several methods have been developed to assess whether cues influence JOLs directly via beliefs, for instance; pre-study JOLs in which only cue information is provided (i.e., *the word that you are about to study is concrete*) (Castel, 2008; Witherby & Tauber, 2017), global predictions made before studying in combination with global poststudy predictions made after studying (Frank & Kuhlmann, 2017), and post-experimental questionnaires (Undorf et al., 2017). Methods that have been used to measure whether the cues influence JOLs indirectly via fluency are: response times in a lexical decision task (Mueller et al., 2013), self-paced study times (Undorf & Erdfelder, 2015), trials to acquisition (Undorf & Erdfelder, 2015), and mental image latency (Hertzog et al., 2003), among others. Importantly, recent theoretical developments suggest that fluency can also influence JOLs via beliefs (i.e., analytic way) such as "*it is quicker to learn, it must be memorable*" (Undorf, 2020). Therefore, independent beliefs and fluency measures must be obtained to establish claims about the analytic or non-analytic basis of JOLs.

2.3 Metamemory Accuracy Measures

A relevant question in metacognition research is the accuracy of metacognitive judgments (Koriat, 2007). Overall, metacognitive judgments are considered accurate if they closely correspond to the criterion (memory). However, this correspondence can be measured in either *relative* or *absolute* terms, which are the two central aspects of accuracy to differentiate.

Relative accuracy (or resolution) refers to the degree to which the judgments discriminate between remembered and non-remembered items. This is the psychological ability to detect which items are more likely to be remembered than others. The within-subjects Goodman-Kruskal gamma correlation has become the standard measure of resolution as recommended by Nelson (1984). Goodman-Kruskal gamma is a rank correlation that measures at the ordinal level the association strength between the two observable variables: JOLs (continuous, from 0% to 100%) and memory performance (dichotomous, remembered = 1, or non-remembered = 0), ignoring ties. The recommendation is to calculate a gamma correlation for each participant's data, JOLs and memory performance. These gamma correlations by participant are then submitted to statistical tests such as *t*-tests or Anovas to make group inferences about the effects of different variables. Gamma is calculated by counting the number of concordant and discordant pairs. A concordant pair consists of two items, one of which has a higher JOL than the other, and it is recalled while the other is not. A discordant pair consists of two items, one of which has a higher JOL than the other, but it is not recalled while the other is, and vice versa. A tie pair consists of two items that have the same value either in the predictor (JOL) or in the criterion (memory), these pairs are ignored in the gamma computation, as mentioned above.

The gamma correlation is computed as follows:

$$\text{Gamma} = (\text{Concordances} - \text{Discordances}) / (\text{Concordances} + \text{Discordances})$$

Gamma ranges from +1 to -1, with higher values indicating better resolution, and a gamma of zero indicating that resolution is at chance level. Nelson (1989) noticed that the relation between Gamma and the probability measure V (V = Concordances) / (Concordances + Discordances) is one-to-one and linear:

$$V = .5\ (\text{Gamma}) + .5$$

Calculating a gamma correlation for each participant's data and then submitting it to an inferential test such as *t*-test or ANOVA is considered a by-participants analysis. This is because only the random sampling variation of participants is considered. Recently, Gamma has been criticized due to inflated Type 1 error in by-participant analysis because the random sampling variation of the items is ignored (Murayama et al., 2014). A solution proposed to this problem is the use of logistic mixed models that consider both the random participant effect and the random item effect, with the memory performance as the outcome variable and JOLs as the predictor variable. Another criticism of gamma is that it deviates from its actual value with liberal or conservative response bias (Masson & Rotello, 2009). The authors showed that this occurs despite gamma not making any distributional assumptions beyond the ordinal level, but because the gamma computation ignores ties. They suggested to use measures grounded on signal detection models that make distributional assumptions to correct for ties. However, to date, despite the criticism, the most widely used measure of relative accuracy among metamemory researchers is still the Goodman-Kruskal gamma correlation. This is probably the case because easy comparisons of relative accuracy as measured by gamma can be made across studies.

The other aspect of accuracy is *absolute accuracy* (or calibration) which refers to the absolute difference of the average JOL and average memory performance. For example, if participants study a list of items and correctly recall 40% of them, while their average JOL is also 40%, this indicates perfect calibration. Bias is the calibration measure most widely used (Dunlosky & Metcalfe, 2009). Bias is easily computed by subtracting the average memory performance from the average JOL of each participant (Nelson & Dunlosky, 1991). Positive values of bias indicate overconfidence while negative values of bias indicate underconfidence. A bias equal to zero indicates perfect calibration. Importantly, to calculate bias, both JOLs and memory performance need to be in the same numerical scale. This is the reason for researchers to solicit JOL on a 0-100% probability scale rather than on a 7-point or 4-point Likert scale.

Relative and absolute accuracy have different implications as suggested by Koriat (2007). Specifically, relative accuracy is important at the item-level, while absolute accuracy is important at the task-level. For example, in the scenario where a student is preparing for an exam, relative accuracy helps the student to determine how much study time to allocate to each topic of the exam. Absolute accuracy is important for assessing the overall level of preparedness, it helps the student to make an informed decision regarding whether to stop or continue studying for the exam as whole. Importantly, both measures of accuracy, resolution, and calibration, are independent. For example, this is demonstrated by the underconfidence-with-practice-effect (UWP), in which there is a shift from overconfidence to underconfidence from the first study-test cycle on, but accompanied by an improvement in resolution (Koriat et al., 2002). This is presumably because JOLs fail to consider the benefit of a retrieval experience and a second learning trial, while increased reliance on mnemonic subjective experience or reliance on the memory-for-past-test heuristic improves resolution (Koriat, 1997).

2.4 Metacognitive Illusions

Systematic dissociations between predicted and actual memory (i.e., metacognitive illusions) have attracted great attention due to the opportunity to investigate the inferential basis of the judgments (i.e., theory-, experience-based). Relevant to this thesis (Manuscript II), I will now focus on three metacognitive illusions: the font size illusion, the stability bias, and the font format illusion.

The font size illusion occurs when large-font words elicit higher JOLs compared to small-font words, despite not resulting in better memory performance (Rhodes & Castel, 2008). This illusion has been widely replicated (see meta-analysis, Chang & Brainerd, 2022). In the original study by Rhodes & Castel (2008), the illusion persisted despite experience across multiple study-test cycles and warnings but was reduced by manipulating the fluency of words via the font format (i.e., standard vs. AlTeRnAtInG). This implied that the font size illusion was non-analytic, i.e., experience based. However, later evidence suggested that fluency does not contribute as much as previously assumed: Mueller et al. (2014) reported that large-font words were not more fluent than small-font words as measured by reaction times in a lexical decision task or self-paced study time. Further, Luna et al. (2019) found evidence for the role of beliefs by showing that people believe large-font words are more important than small-font words. Similarly, Undorf & Zimdahl (2019) showed that JOLs increase monotonically with the font size of words to the point that words are not fluent anymore due to be presented in a very large font size. To date, the debate regarding the inferential basis of the font size illusion continues but belief (analytic) explanations are favored.

Furthermore, two meta-analyses have now shown that there is a small but significant font size effect on memory, suggesting that the font size illusion is a gross overestimation of a small effect (Chang & Brainerd, 2022; Luna et al., 2018).

The stability bias illusion occurs when people assume that memories are stable over time, and thus, they fail to consider the benefit of future learning or the impact of forgetting (Kornell & Bjork, 2009). Specifically, Kornell & Bjork (2009) asked participants to predict their recall performance in a later test scheduled to take place after one, two, three, or four study trials (i.e., ST, SST, SSST, SSST). Results showed that, when the number of study trial was manipulated between-subjects, recall performance increased from Test 1 to Test 4 by an average of 33% points while predicted recall increased only by an average of 3%. The same was found when using aggregate predictions on a between-subjects basis, participants did not predict better recall with more study trials (i.e., average predicted recall for Test 1 was 54% and average predicted recall for Test 4 was 57%). Manipulating number of study trials on a within-subjects basis made JOLs slightly more sensitive to the number of study trials but participants still largely underestimated how much learning occurs (i.e., actual recall from Test 1 to Test 4 increased by an average of 43% points while predicted recall increased only by an average of 8%). In addition, warning participants about the beneficial effect of learning opportunities still did not fully reduce underestimations of learning (i.e., actual recall from Test 1 to Test 4 increased by an average of 48% points while predicted recall increased only by an average of 15%). Overall, participants underestimated how much they can learn with more study. This was despite they held the belief that studying results in better memory as measured by making two predictions for the same item, one for recalling the item in Test 1 and another one for recalling the item in Test 2 or 4. The authors concluded that people failed to apply their metacognitive beliefs that studying results in better memory.

The font format illusion occurs when standard-font words elicit higher JOLs compared to AlTeRnAtInG-font words, despite not resulting in better recall performance (Rhodes & Castel, 2008, but see Mueller et al., 2013). This illusion has been used as a manipulation of perceptual fluency by two studies. In the study by Rhodes & Castel (2008), there was a significant interaction between font size and font format, indicating that the font size of words does not affect JOLs when words are presented disfluently in the alternating format compared to the standard format. However, Mueller et al. (2013) found no significant interaction between word pair relatedness and font format, indicating that the effect of relatedness is not disrupted by presenting words in alternating format. The fact that both studies have used font format as a manipulation of fluency does not indicate that beliefs do not contribute to the font format illusion. To sum up, the font size illusion, stability bias, and font format illusion are systematic dissociations between JOLs and actual memory that occur due to erroneous beliefs about memory, the failure in the application of beliefs, and/or fluency.

2.5 Cue Learning Methods

In general, the topic of cue learning has not been addressed in metamemory research as such. On the one hand, there are methods in the literature to enhance monitoring accuracy. On the other hand, researchers have tried to mend metacognitive illusions on JOLs. As mentioned in the introduction, testing practice and delaying JOLs are two robust methods that improve monitoring accuracy by relying on retrieval attempts of previously learned items (Dougherty et al., 2005; Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991).

In this thesis, I focus on methods for learning new cues to judge memory and for enhancing judgment cue sensitivity during the learning of new items. This is important because control judgments such as the allocation of study time, strategy selection, or restudy choices happen during the acquisition of information (see bottom part of Figure 2). At the same time, these decisions do not only happen for previously learned items but also happen for new items currently being learned. For example, during an initial study session, a student determines which topics need more focus and which can be covered with a single review. Given that accurate monitoring leads to effective regulation of learning (e.g., Metcalfe & Finn, 2008; Rhodes & Castel, 2009; Tullis & Benjamin, 2012), it is crucial to know more about how to enhance judgment sensitivity to valid cues during learning. However, currently little is known about how to enhance cue sensitivity of immediate JOLs and how cues are generally learned. In this subsection, I will provide an overview of methods that can be considered for cue learning. These methods were originally investigated in the contexts of learning about strategy effectiveness from experience (i.e., knowledge updating research) and mending metacognitive illusions through direct instructions or warnings.

Knowledge updating refers to how individuals learn from experience about the effects of study strategies (Dunlosky & Hertzog, 2000). In knowledge updating studies, participants complete two study-test cycles in which they learn word pairs under a high effective study strategy (e.g. imagery,) or under a low effective study strategy (e.g., repetition) instruction. However, JOLs in the first cycle are not sufficiently sensitive to the relative effectiveness of both strategies. Because pairs studied under imagery are better recalled than pairs learned under repetition (mean difference of around 20%), JOLs in the second cycle are expected to accurately predict this difference in magnitude between the two strategies. The knowledge updating framework delineates four assumptions in which experiences from the first study-test cycle would lead to improved JOLs in a second cycle. The first assumption is that the strategies must be differentially effective. The second assumption is that monitoring must occur during learning and testing. The third assumption is that recall performance must not only be monitored but also attributed to the specific strategy. The last assumption is that the newly acquired knowledge about strategy effectiveness must be utilized in JOLs. However, results from knowledge updating studies have shown that JOLs in a second cycle are not fully adjusted to reflect the superiority of the imagery strategy over repetition (Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Mueller et al., 2015). This incomplete JOL update occurs despite global predictions about the number of items learned under each study strategy being better adjusted, though not entirely.

One explanation for why participants do not fully update their JOLs is that the utilization assumption fails because other more prominent cues during learning overshadow the strategy effectiveness cue, this is called the *encoding-disrupts-updating* hypothesis (Mueller et al., 2015). Another explanation is that there is a failure in the attribution of item recall performance to the strategy, this is called the *inferential-deficit* hypothesis (Dunlosky & Hertzog, 2000; Matvey et al., 2002). Both hypotheses have received partial support. Mueller et al. (2015) used pre-study JOLs, which are not influenced by experiences during learning (e.g., *You are about to study a pair using imagery (or repetition), please rate how likely you are to remember it*), to test the encoding-disrupts-updating hypothesis. They found that pre-study JOLs are almost equivalent to immediate JOLs in that both show a utilization deficit by discounting newly acquired knowledge about strategy effectiveness in JOLs. Directly testing the inferential-deficit hypothesis by providing feedback on how many pairs are correctly recalled did not lead to more updating (Mueller et al., 2015). But, another study found that testing pairs in strategy blocks and presenting the strategy during test helped participants to monitor their test performance and make correct study strategy attributions which in turn was effective for JOL updating (Price et al., 2008). In sum, the knowledge updating framework for JOLs outlines assumptions that might support cue learning. This research shows that the second assumption (i.e., monitoring of study and test performance), the third assumption (i.e., test performance attributions to cues), and the fourth assumption (i.e., utilization of newly acquired knowledge) can fail and prevent adaptive cue learning.

In the context of metacognitive illusions, researchers have tried to correct the metamemory biases by warning participants. Such warnings have been used to prevent the font size illusion (Rhodes & Castel, 2008, Experiment 4), the stability bias (Kornell & Bjork, 2009, Experiment 8), the interleaving illusion (Yan et al., 2016, Experiment 3, 4 , and 5), and the foresight bias (Koriat & Bjork, 2006a). All these studies used immediate JOLs except the study by Yan et al. (2016) which used aggregate memory predictions. Warnings are often provided in the form of a short information about the effect of the cue manipulated on memory (e.g., *the size of the word should not affect what you will be able to remember later so do not pay attention to size*). In the studies on the font size illusion (Rhodes & Castel, 2008) and the stability bias (Kornell & Bjork, 2009), warnings were provided before participants started with the experimental tasks. In the study by Yan et al. (2016), warnings were provided after the study phase but before collecting aggregate memory predictions. In the study by Koriat and Bjork (2006a), warnings were provided between the first and the second study-test cycle. In most studies, warnings were not successful at mending the metamemory illusions except for the study by Koriat & Bjork (2006a). One critical difference of Koriat's and Bjork study is that warnings were not provided in the form of text but were given by the researcher who carefully explained the illusion and illustrated it by administering a short exercise to the participants.

Yan et al. (2016) suggested that metacognitive judgments are hard to change because of a) pre-existing beliefs about learning and memory, b) experiences of fluency during learning, and c) the belief of being unique as a learner which implies that any feedback about how others perform in the task is not informative to oneself. It is likely that all or a combination of these factors contribute to the resistance to change of metamemory judgments. If warnings do not change pre-existing beliefs, participants may disregard warnings as they consider themselves better experts on their cognition. Alternatively, if warnings change beliefs, participants' judgments may not improve because experiences during learning overshadow the use of beliefs (i.e., the utilization assumption from the knowledge updating framework fails). The latter possibility is supported by studies showing that beliefs need to be activated to impact JOLs (Ariel et al., 2014; Undorf & Erdfelder, 2015). But it is not supported by studies showing that people can integrate multiple cues on their JOLs and do not focus on single unified feeling of 'ease' that overshadows the use of individual cues (Undorf et al., 2018; Undorf & Bröder, 2020).

In sum, knowledge-updating research tells us that JOLs are not fully updated to reflect the beneficial effects of a study strategy on memory from experience across cycles. This can happen either because participants fail to monitor their study and test performance, fail to attribute their test performance to the effective strategy, or fail to use the recently acquired knowledge when making item-by-item JOLs. Further, research on mending metacognitive illusions tells us that warnings are not effective for people learning to adequately use cues for their JOLs because the warnings may not change pre-existing beliefs, or experiences during learning overshadow the knowledge gained from the warning.

Studies highlight the challenges of learning and applying cues in metamemory judgments. This further underscores that it is crucial to understand how to increase monitoring sensitivity to cues predictive of memory and how to learn new cues that inform metamemory judgments. In the following chapter, I test and discuss cue learning methods from three different angles: 1) learning about item memorability from one's own experience, 2) learning and unlearning cues from feedback, and 3) learning new cues extracted from the environment. The next chapter summarizes the three manuscripts and present their core results (the full manuscripts can be found appended to this thesis).

3. Cue Learning in Metamemory

In the three manuscripts presented in this chapter, I approach the question of *how people learn cues to inform their metamemory judgments* through three different angles; learning about item memorability from one's own experience (Manuscript I), learning and unlearning cues from feedback (Manuscript II), and learning new cues extracted from the environment (Manuscript III). Specifically, Manuscript I tests the cue basis and resolution of two types of metamemory judgments made for different memory criteria (own memory [JOLs] vs. generic item memorability) and show that learning and testing experiences in a JOL task led to learning cues predictive of general item memorability, but not vice versa. Manuscript II tests the effectiveness of different forms of feedback to mend metacognitive illusions in JOLs and demonstrate that cognitive feedback about one's own recall performance and JOL for each studied item is not effective while additional metacognitive feedback with an in-depth explanation about biased metacognition is. In Manuscript III, I take a novel approach by testing and demonstrating statistical learning as a mechanism for learning cues from the environment that subsequently inform JOLs.

3.1. Learning About Item Memorability From One's Own Experience

Navarro-Báez, S., Undorf, M., & Bröder, A. (2024). Predicting the memorability of scene pictures: Improved accuracy through one's own experience. *Quarterly Journal of Experimental Psychology*, *0*(0), 1-20. https://doi.org/10.1177/17470218241239829

All data, materials, and analyses from the present manuscript are available at https://osf.io/hpy6q/. Experiments 1 and 2 were not preregistered. Experiment 3 was preregistered at https://osf.io/3fujm.

The human visual memory capacity for pictures of scenes is extremely good (Nickerson, 1965; Shepard, 1967; Standing, 1973; Standing et al., 1970). However, studies investigating metamemory accuracy of naturalistic scene pictures have yielded conflicting evidence: It is not clear how accurate people are at predicting which pictures will be remembered and which will not.

The conflicting findings on metamemory accuracy for scene pictures stem from studies using either *memorability judgments* (MJs) — judgments of stimulus memorability in general — or *judgments of learning* (JOLs) — predictions of one's own later memory performance for recently studied items. While MJs have been found to be unpredictive of actual picture memorability (Isola et al., 2011a, 2011b, 2014), JOLs have been found to be moderately predictive of participant's actual recognition memory for pictures (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). In this manuscript, we systematically compared the cue basis and relative accuracy of JOLs and MJs to test whether accuracy differences are due to MJs referring to memorability as a generic item attributed versus JOLs referring to one's own future memory performance. In doing so, we found that MJs become more accurate after a JOL task due to increased sensitivity to cues predictive of picture memorability. This finding was replicated in Experiment 2. Experiment 3 was designed to disentangle which specific component of the JOL task drives the improvement in MJ cue basis and accuracy. I will first provide a short theoretical background on the two types of metamemory judgments: MJs and JOLs. I will then present the three experiments of this manuscript.

In a large-scale series of studies, Isola et al. (2011a, 2011b, 2014) quantified the memorability of more than 2,000 images of real-world scenes from the SUN database (Xiao et al., 2010) using a repeat detection task. In this task, a total of 665 participants saw sequences of 120 images and were asked to detect whenever there was a repetition of an image. Image memorability was measured as the percentage of correct detections by participants. Further, MJs were obtained in two tasks from different samples of 30 participants each. In the first task, every time a participant identified a repetition of an image, the question "*Is this a memorable image? Yes/No*" had to be answered for 36 additional images. In the second task, the question "*If you were to come across this image in the morning, and then happen to see it again at the end of the day, do you think you would realize that you have seen this image earlier in the day? Yes/No*" had to be answered. Results showed that MJs did not predict image memorability: correlations between MJs and image memorability were $\rho = -0.19$ in the first task and $\rho = -0.02$ in the second task. This result suggests that people lack insight into picture memorability.

In contrast, JOL studies using pictures of scenes have found that JOLs are relatively accurate in terms of relative accuracy and track cue effects on actual memory performance (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). This is illustrated in the study by Undorf and Bröder (2021) using also pictures from the SUN database (Xiao et al., 2010), in which JOLs were higher for all the fives cues that helped memory (i.e., contextually distinctiveness, color, telling a story, two repetitions, containing persons) and only failed to reflect that peacefulness hindered memory. This result suggests that people know which image features are predictive of picture memorability and can accurately predict their recognition memory performance on this basis.

The conflicting findings on the accuracy of MJs and JOLs may arise from genuine differences in the metamemory processes underlying each judgment. Specifically, the judgments are made for distinct aspects of memorability (i.e., one's own memory vs. generic picture memorability). There is evidence that people use different cues when the metamemory judgment is about personal memory than when it is about others' memory (Tullis & Fraundorf, 2017). The metamemory processes might also be different because of the task in which the metamemory judgments are solicited (i.e., during intentional learning vs. judgment task only). Cues vary in their availability depending on the time in the learning process that the judgment is solicited. For instance, pre-study JOLs ("*You are about to study an emotional item")* are less accurate than immediate JOLs because they cannot rely on experiential cues such as familiarity or encoding fluency (Price & Harrison, 2017; Undorf & Bröder, 2020). The same is true for ease-of-learning judgments ("*How easy is it to learn this item?*") made prior to learning (Kelemen et al., 2000; Leonesio & Nelson, 1990; Pieger et al., 2016).

It is also possible, however, that differences in accuracy between the two judgments do not reflect genuine metamemory differences but rather result from methodological variations between the studies. JOL studies with pictures typically use an old/new recognition memory test in which participants intentionally learn pictures prior to the test (Caplan et al., 2019; Hourihan, 2020; Hourihan & Bursey, 2017; Kao et al., 2005; Undorf & Bröder, 2021). In contrast, MJs were related to memory performance measured in a repeat detection task in which participants simultaneously encoded and recognized pictures (Isola et al., 2011a, 2011b, 2014). Moreover, the memory criterion is different, JOLs are related to participant's individual memory performance (i.e., correlation of JOLs with item recognition memory by participant), whereas MJs are related to image memorability at the item level (i.e., correlation of MJs with item recognition memory aggregated across participants). Image memorability is highly consistent across participants (Bainbridge et al., 2013; Bylinskii et al., 2015; Isola, et al. 2011a; Isola, et al., 2011b, 2014), but there is also idiosyncratic information contributing to judgment predictive accuracy that get lost when recognition memory performance is aggregated across participants (Bröder & Undorf, 2019; Undorf et al., 2022).

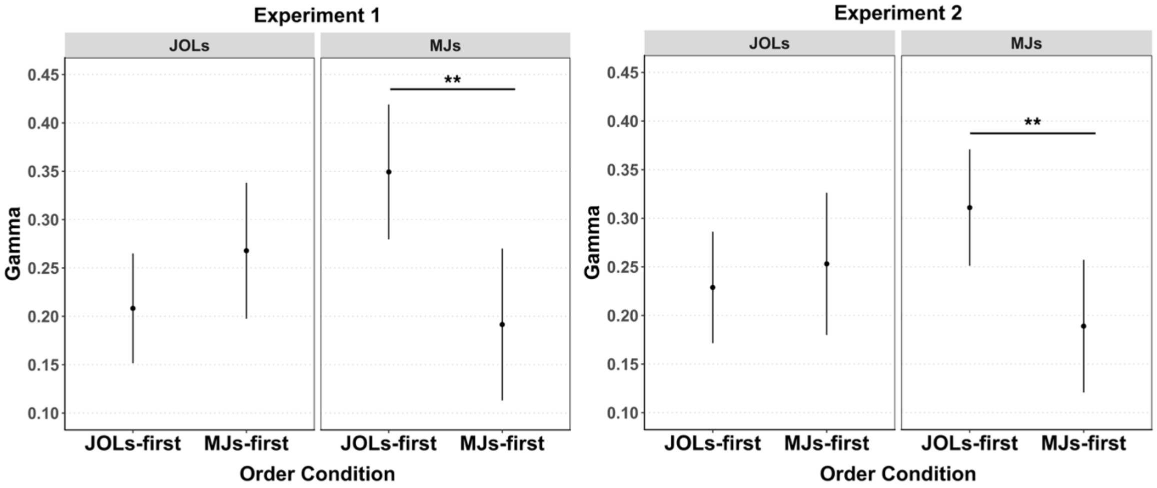
As previously mentioned, in this manuscript, we directly compared the cue basis and relative accuracy of JOLs and MJs for pictures of scenes. For this, participants made JOLs and MJs which differ on the aspect of picture memorability judged (one's own future memory vs. memorability as a generic item attribute) and the task (during learning vs. judgment-only task), respectively. At the same time, we ensured that the JOL and MJ methods were as similar as possible in all other respects (i.e., same 0-100 judgment scale, same materials with identical cues manipulated within picture sets, same memory criterion value). *Population scene memorability* was the memory criterion value used which was defined as the proportion of recognition hits minus the proportion of false alarms per scene in each experiment's recognition memory task. This measure corresponds to the corrected hit rate, also known as *Pr* (Snodgrass & Corwin, 1988). If there were true metamemory processing differences between the judgments, we expected to see clear differences in cue use and judgment accuracy. In contrast, if discrepant findings were largely due to methodological differences across studies, we expected to see similar cue basis and accuracy of JOLs and MJs.

In Experiments 1 and 2, we used a within-subjects design by presenting a JOL task and an MJ task to the same participants, with the order of tasks counterbalanced between participants (order condition: JOLs-first, MJs-first). In the JOL task, 52 participants (Experiment 1) and 50 participants (Experiment 2) provided a JOL after studying each picture from a set of 120 pictures, and following a 3-min filler task, completed a recognition memory test with 240 pictures. In the MJ task, participants made an MJ for each picture from another set of 120 pictures. They were explicitly told not to study the pictures, but only to judge their general memorability.

In Experiment 1 we orthogonally varied the cues aesthetics and interestingness in two levels (high vs. low) within picture sets. Aesthetics and interestingness were the image attributes identified as negative predictors of image memorability, but positively affecting MJs in Isola et al. (2011a, 2011b, 2014). Results showed that both JOLs and MJ were unaffected by aesthetics but were higher for pictures high rather than low in interestingness. Recognition memory performance was better for pictures low in aesthetics than high, and for picture high in interestingness than low. The metamemory results suggest a similar cue basis of JOLs and MJs. Furthermore, reliable resolution showed that both metamemory judgments captured differences in the relative population memorability of scenes. However, the accuracy of MJs improved substantially after completing a JOL task, whereas completing an MJ task first did not affect JOL accuracy (see Figure 3 and Table 1). A potential reason for this improvement in MJs is that participants gained knowledge about the abstract image feature of memorability in the JOL task.

Figure 3

Gamma Correlations Between Population Scene Memorability (Hit Rate Corrected per Scene) and Judgments of Learning (JOLs) or Memorability Judgments (MJs) in Each Task Order Condition of Experiment 1 and Experiment 2



Note. Error bars represent one standard error of the mean. *JOLs-first* = the JOL task was completed first and the MJ task second. *MJs-first* = the MJ task was completed first and the JOL task second.

Experiment 2 aimed to replicate the partly unexpected findings of Experiment 1. Because JOLs and MJs were based on similar cues in Experiment 1, we did not manipulate individual cues in Experiment 2, but instead used scenes that varied widely in scene memorability. Results showed that both JOLs and MJs were predictive of differences in population scene memorability. We replicated the finding that the relative accuracy of MJs improved after a JOL task, whereas prior experiences with making MJs did not improve JOL accuracy (see Figure 3 and Table 1). Importantly, MJs increased more strongly with scene memorability in the JOLs-first than in the MJs-first condition, indicating that MJs become more sensitive to scene memorability effects after a JOL task. In contrast, scene memorability effects on JOLs were unaffected by the task order condition. This finding supports our hypothesis that participants learn about the general memorability of scenes by completing a JOL task and, thus, make MJs for new set of pictures on an updated basis.

Table 1

Means (SDs) of the Gamma Correlation Between Population Scene Memorability (Hit Rate Corrected per Scene) or Participant's Own Memory Performance and JOLs or MJs in Each Task Order Condition of Experiments 1, 2, and 3

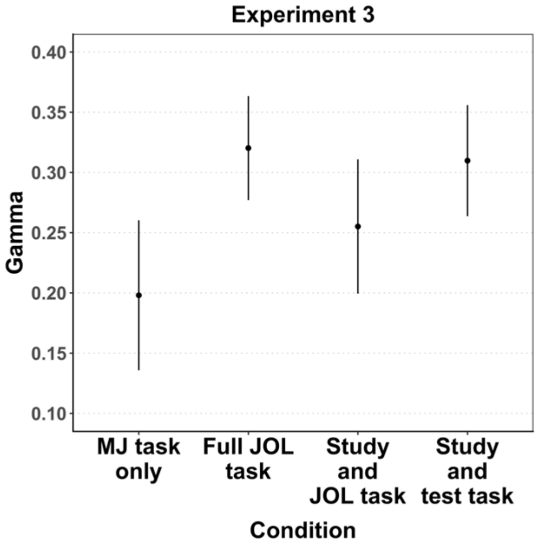
Experiment and condition	Accuracy criterion		
	Population scene memorability		Own memory performance
Experiment 1	JOLs	MJs	JOLs
JOLs first	.21 (.14)	.35 (.17)	.36 (.15)
MJs first	.27 (.17)	.19 (.19)	.40 (.16)
Experiment 2			
JOLs first	.23 (.14)	.31 (.15)	.33 (.25)
MJs first	.25 (.18)	.19 (.17)	.45 (.18)
Experiment 3			
MJ-task-only	-	.20 (.22)	-
Full-JOL-task	.26 (.14)	.32 (.15)	.38 (.22)
Study-and-JOL-task	.28 (.18)	.26 (.20)	-
Study-and-test-task	-	.31 (.17)	-

Note. JOLs = judgments of learning, MJs = memorability judgments, Population scene memorability = hit rate minus false alarm rate per scene across participants in each experiment

Experiment 3 aimed at disentangling which component of the JOL task drives the improvement in MJ accuracy. For this, in the first part of the experiment, different groups of participants completed either the full JOL task (*full-JOL-task* condition), a learning phase with JOLs but without test (*study-and-JOL-task* condition), a learning without JOLs but with test (*study-and-test-task* condition), or no component of the JOL task (*MJ-task-only* condition). In the second part of the experiment, all 205 participants completed an MJ task. Results showed that MJ accuracy improved in all experimental conditions in comparison to the control condition (MJ-task only), see Table 1 and Figure 4. As the learning phase is the common factor in all experimental conditions, our result suggests that a learning phase by itself provides experiences that are beneficial for subsequently assessing the memorability of pictures. We did not find evidence for additive effects of making JOLs and taking a test on MJ accuracy. Regarding the individual effects of making JOLs and taking a test on MJ accuracy, gamma correlations suggested that neither making JOLs nor taking a test improves MJ accuracy, as did the mixed-effects model analysis. In contrast, the analysis of Pearson correlations suggested that a recognition memory test improves MJ accuracy more than making JOLs. Importantly, MJs were influenced more strongly by scene memorability in the study-and-test-task condition than in the study-and-JOL-task condition. This result suggests that completing a recognition memory test is more beneficial for enhancing the MJ sensitivity to valid cues than making JOLs.

Figure 4

Gamma Correlations Between Population Scene Memorability (Hit Rate Corrected per Scene) and Memorability Judgments (MJs) in Each Condition of Experiment 3



Note. Error bars represent one standard error of the mean.

In conclusion, the three experiments in this manuscript revealed that both MJs and JOLs are moderately accurate at predicting differences in the population memorability of scenes. The experiments also revealed that MJs and JOLs have a similar cue basis when pictures differed in aesthetics and interestingness (Experiment 1) or represented a broad range of scene memorability (Experiments 2 and 3). This implies that discrepant findings on the accuracy of JOLs and MJs reported in prior work were largely due to methodological differences across studies. At the same time, we did find a notable difference between JOLs and MJs: MJ accuracy improved with prior learning and testing experience, whereas JOL accuracy was independent of prior assessments of general memorability. This shows that reflections about and experiences with one's own learning and testing contribute to people's understanding and knowledge about memorability as a generic attribute of an item.

3.2. Learning and Unlearning Cues From Feedback

Navarro-Báez, S., Bröder, A., & Undorf, M. (2024). *Mending metacognitive illusions requires metacognitive feedback*. [Manuscript submitted for publication]. Department of Psychology, University of Mannheim. Department of Psychology, Technical University of Darmstadt.

All data, materials, and analyses from the present manuscript are available at https://osf.io/vgy7d/?view_only=5d02381f12754484ad1c422cd1d4f44b. Designs and analyses of all experiments were not preregistered.

Accurate metacognitive monitoring — the real-time assessment of cognitive processes — is important because it guides behavior (Nelson & Narens, 1990; also see Thiede et al., 2003; Tullis & Benjamin, 2012). This implies that the regulation of behavior such as self-regulated learning cannot be successful if metacognitive monitoring judgments are incorrect. Metacognitive illusions occur when the metacognitive judgments rely on invalid cues or fail to rely on valid cues predictive of performance (see Undorf et al., 2022, for a review). In this manuscript, we addressed the practically relevant question of *how to improve metacognitive awareness of cue validity*. To this end, we tested the effectiveness of two forms of feedback, *cognitive feedback* (Balzer et al., 1989) alone and with additional *metacognitive feedback* (Fiedler et al., 2020). We specifically tested whether these forms of feedback can improve metacognitive awareness of cue validity in JOLs and mend metacognitive illusions. Finding effective ways to train metacognitive monitoring is important because prior research has not yet discovered how to achieve robust and large improvements (e.g., Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Kornell & Bjork, 2009; Mueller et al., 2015; Pan & Rivers, 2023; Yan et al., 2016).

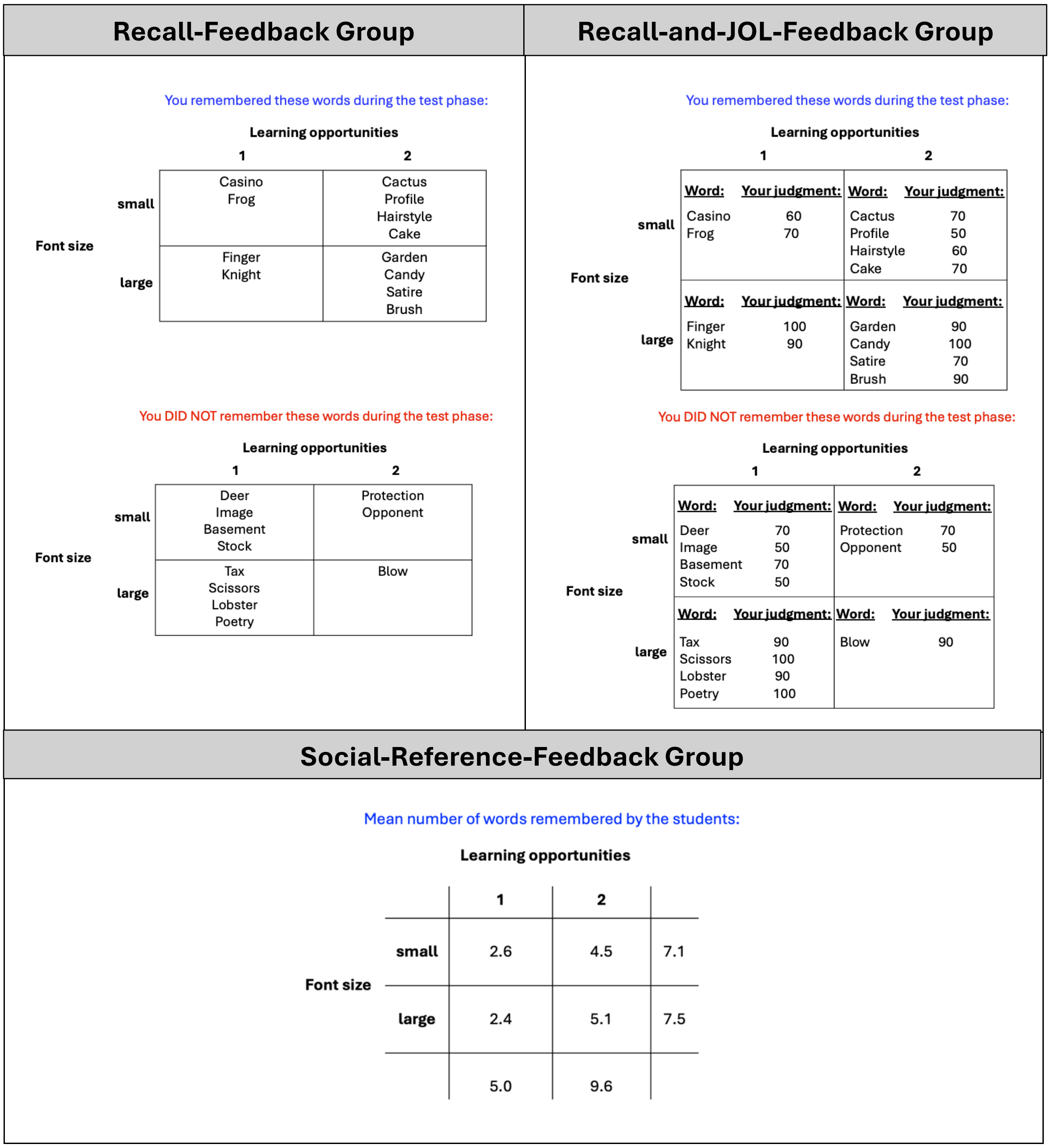
Different methods have and have not been successful in promoting metacognitive awareness of the validity (or invalidity) of cues. These methods include: experience across multiple study-test cycles (Castel, 2008; Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Mueller et al., 2015; Pan & Rivers, 2023; Sungkhasettee et al., 2011), warnings (Koriat & Bjork, 2006a; Kornell & Bjork, 2009; Rhodes & Castel, 2008; Yan et al., 2016), increasing the salience of relevant aspects of the study (Castel, 2008; Price et al., 2008; Yan et al., 2016), and performance feedback (Mueller et al., 2015; Pan & Rivers, 2023; Tullis et al., 2013). The method that has proven the most successful is increasing the salience of relevant aspects of the study. For instance, participants learned to predict the primacy and recency effect in memory when the serial position of items was presented before studying each item (Castel, 2008). In contrast, experience across multiple study-test cycles, warnings, and performance feedback have been often ineffective for improving the cue basis of item-by-item JOLs.

In another line of research, namely judgment and decision-making research, a famous paper by Brehmer (1980) reviewed studies on the ability to learn from experience in probabilistic situations such as clinical inference. Brehmer (1980) came to the pessimistic conclusion that people do not learn from experience with mere outcome feedback (i.e., knowing the outcome) in complex and uncertain tasks because of the number of biases that prevent them from learning (e.g., confirmation bias, positivity bias, belief that the world is deterministic rather than probabilistic). In contrast, a review by Balzer et al. (1989) revealed that so-called *cognitive feedback* is effective for improving the cue basis of judgments compared to mere outcome feedback. Cognitive feedback consists of task information, cognitive information, and functional validity information. Task information refers to the relations between cues and criterion (i.e., task system) – "*Which cues are valid?*". Cognitive information refers to the subject's cognitive system – "*How do people use the cues for their judgments?*". Functional validity information refers to the relation of the cognitive system to the task system – "*What is the difference between the former two?*". Several studies have demonstrated that cognitive feedback improves the cue basis and accuracy of judgments about the external world (e.g., Karlsson et al., 2004; Newell et al., 2009; Seong & Bisantz, 2008). Since judgments about external criteria are like judgments about one's own cognition in that both judgments rely on probabilistic cues, cognitive feedback may be beneficial for learning to distinguish the different predictive validities of cues in metacognition as well.

In all three experiments of this manuscript, participants completed three study-test cycles in which they studied three different lists of single words. Participants made JOLs and received either feedback or no feedback after each cycle. Experiment 1 aimed to test the effectiveness of cognitive feedback for mending the font size illusion (Rhodes & Castel, 2008) and the stability bias (Kornell & Bjork, 2009). Experiment 1 entailed four between-subjects groups. In the control group, participants received no feedback, so they only had their own memory of test performance as feedback on cue validity. In the outcome feedback group (recall-feedback group), participants saw the words they had recalled and not recalled, organized by the two cues (see Figure 5). In the cognitive feedback group (recall-and-JOL-feedback group), the list was accompanied by the JOL participants had given to each word during study. This enables to compare the actual recall (i.e., task information) with their prediction (i.e., cognitive information). Finally, the social-reference-feedback group was informed about the cues manipulated in the experiment and their cue validity, accompanied by a table showing the average performance of other participants doing this task.

Figure 5

Example Feedback Presented to Participants in the Recall-Feedback, Recall-and-JOL-Feedback, and Social-Reference-Feedback Group in Experiment 1



Recall-Feedback Group

You remembered these words during the test phase:

		Learning opportunities	
		1	2
Font size	small	Casino Frog	Cactus Profile Hairstyle Cake
	large	Finger Knight	Garden Candy Satire Brush

You DID NOT remember these words during the test phase:

		Learning opportunities	
		1	2
Font size	small	Deer Image Basement Stock	Protection Opponent
	large	Tax Scissors Lobster Poetry	Blow

Recall-and-JOL-Feedback Group

You remembered these words during the test phase:

		Learning opportunities			
		1		2	
		Word:	Your judgment:	Word:	Your judgment:
Font size	small	Casino Frog	60 70	Cactus Profile Hairstyle Cake	70 50 60 70
	large	Finger Knight	100 90	Garden Candy Satire Brush	90 100 70 90

You DID NOT remember these words during the test phase:

		Learning opportunities			
		1		2	
		Word:	Your judgment:	Word:	Your judgment:
Font size	small	Deer Image Basement Stock	70 50 70 50	Protection Opponent	70 50
	large	Tax Scissors Lobster Poetry	90 100 90 100	Blow	90

Social-Reference-Feedback Group

Mean number of words remembered by the students:

		Learning opportunities		
		1	2	
Font size	small	2.6	4.5	7.1
	large	2.4	5.1	7.5
		5.0	9.6	

We hypothesized that the cognitive information in the recall-and-JOL-feedback group would increase JOL reliance on number of future study opportunities (i.e., valid cue) and decrease JOL reliance on font size (i.e., invalid cue), and in turn, increase relative accuracy. At the same time, it was an open question whether the outcome feedback in the recall-feedback group would lead to better cue weighting and accuracy as suggested by the inferential deficit hypothesis

(Dunlosky & Hertzog, 2000). This hypothesis states that JOLs do not improve from experience because there are limited cognitive resources to monitor test performance and make inferences about valid cues. Hence, improvements should occur when participants are presented with feedback on their recall performance. It was also an open question whether social reference information would be sufficient for improvements or even go beyond the cognitive feedback. If, in contrast, improvements occur in all groups, this would indicate that study-test experience is beneficial for learning cue validities and implementing them in JOLs.

Table 2

Means (SDs) of the Gamma Correlation between JOLs and Recall Performance and Bias Measure in Each Cycle and Group of Experiments 1, 2, and 3

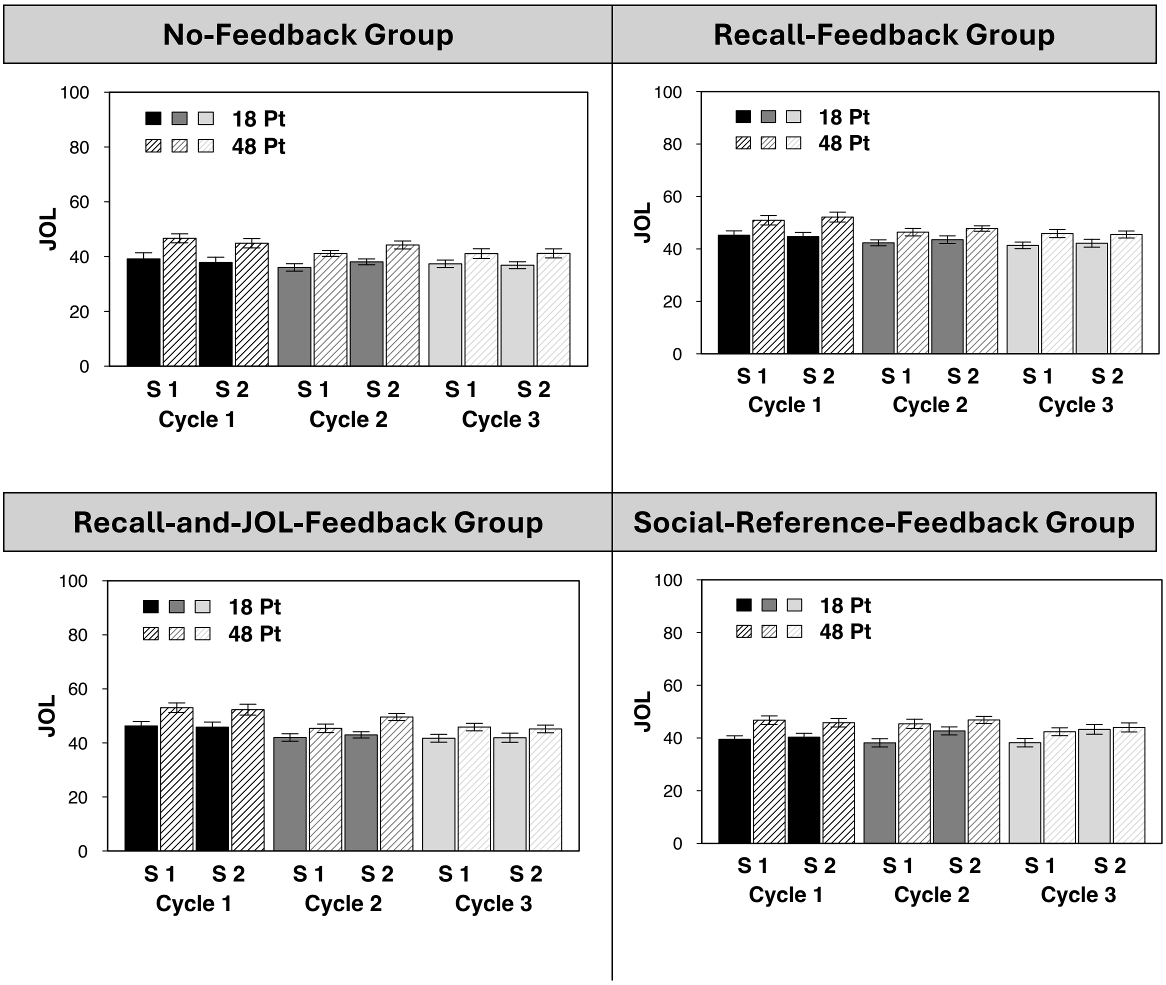
	Cycle					
Experiment and group	Gamma			Bias		
	1	2	3	1	2	3
Experiment 1						
No feedback	.33 (.23)	.30 (.34)	.26 (.37)	8.28 (18.64)	1.43 (20.15)	-0.83 (18.64)
Recall feedback	.21 (.27)	.36 (.26)	.29 (.28)	7.12 (18.41)	-1.75 (18.86)	-1.67 (18.25)
Recall and JOL feedback	.20 (.27)	.22 (.33)	.25 (.30)	7.49 (26.25)	-3.97 (23.18)	-1.40 (18.99)
Social reference feedback	.29 (.27)	.31 (.29)	.31 (.37)	-2.60 (21.68)	-7.98 (20.35)	-7.16 (14.86)
Experiment 2						
No feedback	.31 (.32)	.31 (.41)	.41 (.26)	10.26 (27.65)	-5.53 (21.40)	-5.03 (20.49)
Catch-all cognitive feedback	.26 (.26)	.30 (.29)	.23 (.37)	8.24 (19.19)	-0.69 (14.27)	-1.52 (14.03)
Experiment 3						
No feedback	.21 (.25)	.36 (.23)	.30 (.28)	8.08 (21.30)	-5.94 (16.45)	-7.54 (18.68)
Metacognitive feedback	.28 (.22)	.30 (.31)	.44 (.26)	5.60 (16.08)	1.21 (22.71)	0.12 (21.16)

Experiment 1 (N = 156) results showed that neither type of feedback improved the cue basis of JOLs or JOL accuracy. Participants continued overweighting font size and underweighting number of future study opportunities in their JOLs even with the opportunity to relate individual JOLs to actual memory performance in the recall-and-JOL-feedback group (see Figure 6). At the same time, there were neither improvements in relative nor absolute accuracy (see Table 2). JOLs switched from overconfidence to underconfidence after the first cycle (see Table 2), a pattern that it is well established for repeated study-test cycles using the same materials and known as the underconfidence-with-practice effect (Koriat et al., 2002).

Figure 6

Mean Judgments of Learning (JOL) in Each Cycle for Words Presented Once (S1) or Twice (S2) in a Small (18 pt) or a Large (48 pt) Font Size in Each Group of Experiment 1

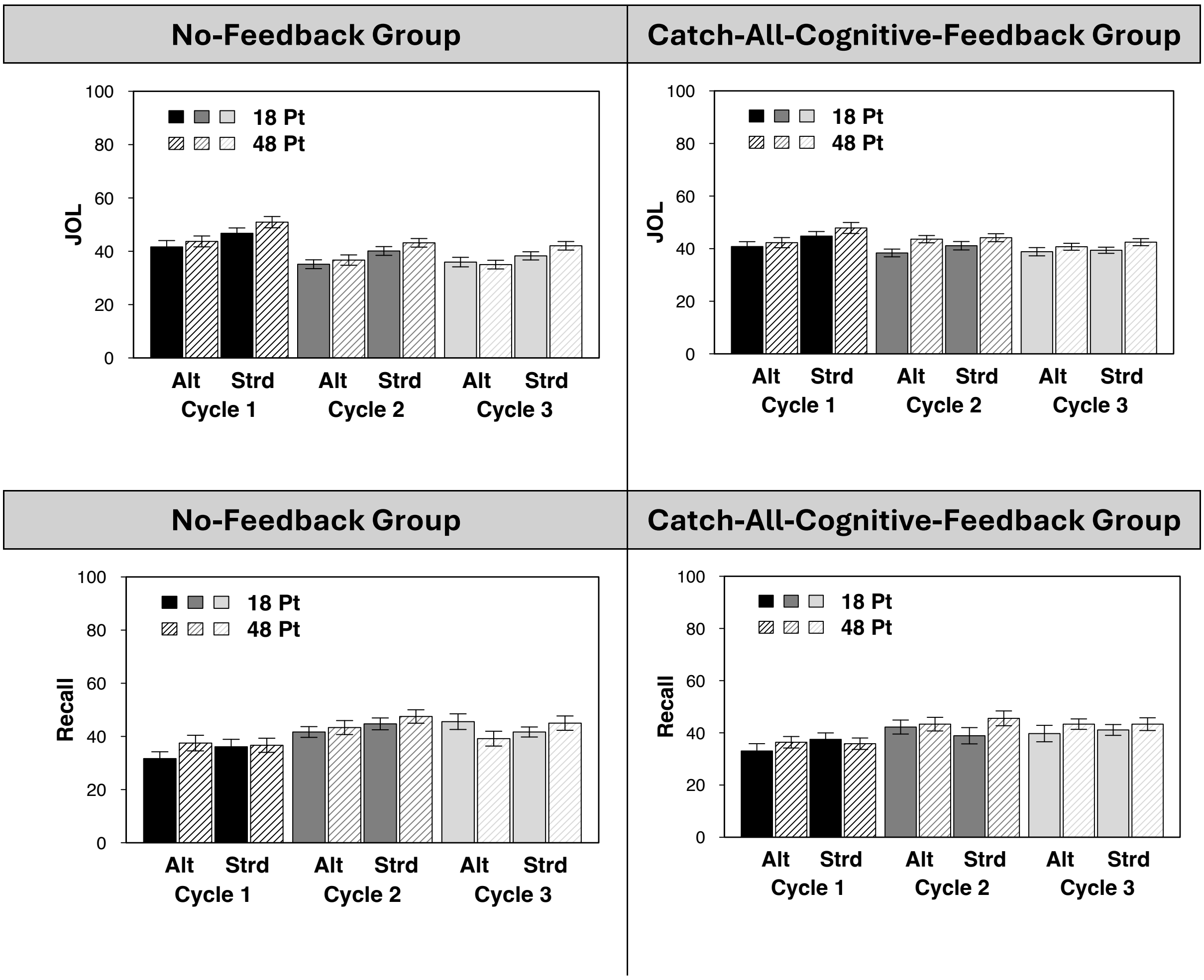
Note. Error bars represent one standard error of the mean.



Experiment 2 (N = 80) combined all forms of feedback into a single ‘catch-all cognitive feedback’ group to provide participants with maximum information. Participants in the catch-all-cognitive-feedback group saw the same information as recall-and-JOL-feedback group (see Figure 5) followed by the same information as the social-reference-feedback group (see Figure 5). In Experiment 2, we manipulated invalid cue font format, standard vs. aLtErnAtiNg words (Mueller et al., 2013; Rhodes & Castel, 2008), in addition to font size with the aim of facilitating cue learning when the two cues are perceptual. Results showed no improvements in cue use or JOL accuracy in the catch-all-cognitive-feedback group compared to the control group which received no feedback. Participants in this group continued overweighting font size and font format despite receiving maximum information (see Figure 7). Relative accuracy did not improve across cycles, and JOLs switched again from overconfidence to underconfidence after Cycle 1 (see Table 2).

Figure 7

Mean Judgments of Learning (JOL) and Percentage of Correctly Recalled Words (Recall) in Each Cycle for Words Presented in Alternating (Alt) or Standard (Strd) Font and in a Small (18 pt) or a Large (48 pt) Font Size in Each Group of Experiment 2

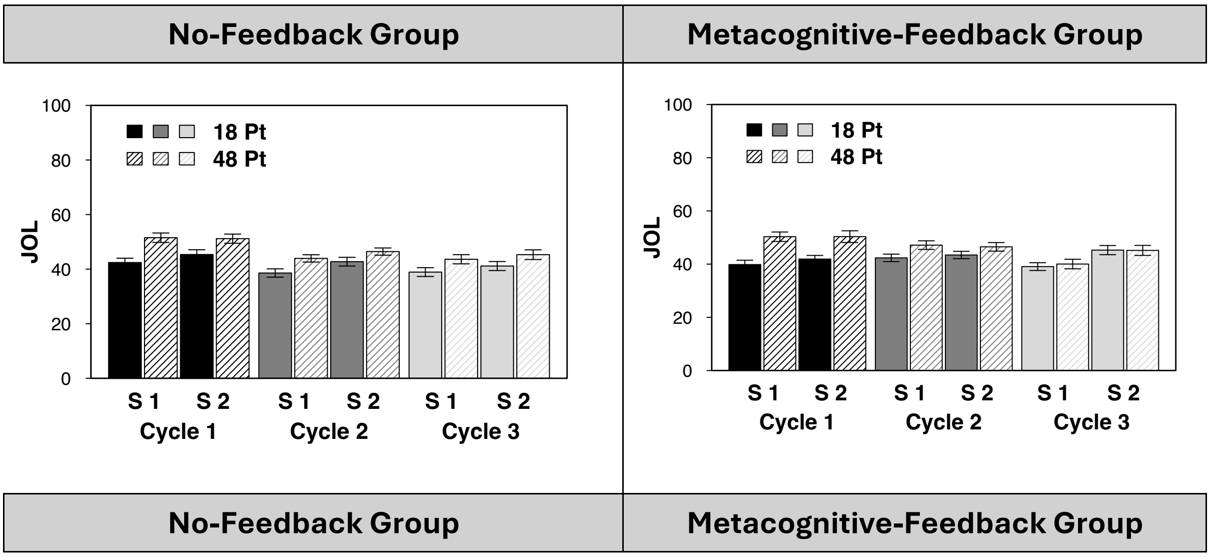


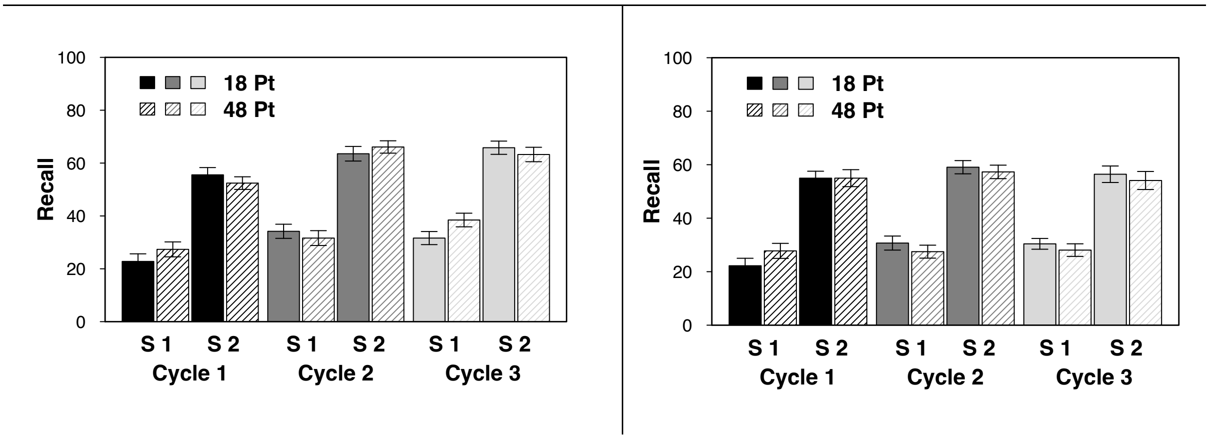
Note. Error bars represent one standard error of the mean.

Experiment 2 showed that even a very blunt form of cognitive feedback was not effective for improving JOLs. This might be because pre-existing beliefs influence how the cognitive feedback is interpreted and stored (Yan et al., 2016). Thus, in Experiment 3 (N = 77), we designed a new form of feedback to remedy that participants might be misperceiving the cognitive feedback due to pre-existing beliefs. We followed Fiedler's et. al. (2020) recommendations of an effective form of feedback to design an informative 'metacognitive feedback'. After being presented with the average memory performance of other participants, participants read textual information from a first-person perspective about possible perceptions during learning (e.g., *words written in a large font size are particularly conspicuous during learning and are perceived as particularly easy to read and learn*), the validities of the cues (e.g., *the font size of words does not usually affect memory performance*), and recommendations on which factors to focus when making JOLs (e.g., *an additional learning opportunity has a stronger influence on memory that the font size*). Further, Experiment 3 used a similar approach as Pan and Rivers (2023) to ensure that participants fully attended and understood the feedback. This approach was asking participants to describe how each of the cues affected their memory and their JOLs after receiving feedback. Experiment 3 manipulated font size and the number of future study repetitions as in Experiment 1.

Figure 8

Mean Judgments of Learning (JOL) and Percentage of Correctly Recalled Words (Recall) in Each Cycle for Words Presented Once (S1) or Twice (S2) in a Small (18 pt) or a Large (48 pt) Font Size in Each Group of Experiment 3





Note. Error bars represent one standard error of the mean.

Experiment 3 results showed that the metacognitive feedback was successful at increasing JOL reliance on number of future study opportunities (see Figure 8), improving resolution, and descriptively, improving calibration (see Table 2). The questionnaire data clearly showed that most participants correctly identified better memory performance for words learned twice than once, but fewer participants correctly identified their memory performance for large- and small-font words. Although font size descriptively affected JOLs less strongly in Cycle 2 and even less so in Cycle 3 in the metacognitive-feedback group than in the no-feedback group (see Figure 8), the three-way interaction (Group x Cycle x Font Size) was not significant. This might be because JOLs are easier to correct for those cues with robust predictive validity, and not for those invalid cues. Specifically, experiences of high fluency produced by invalid cues may still contribute despite metacognitive beliefs being correct. Alternatively, individual differences in the effect of font size on memory performance (i.e., some participants showing better recall for large than small words) did not facilitate that all participants updated their JOLs which was reflected in the non-significant three-way interaction. Alternatively, given that we found the font size illusion decreased across study-test cycles in all conditions (Cycle x Font Size interaction), beneficial effects of metacognitive feedback were harder to detect, and doing so would have required more statistical power.

To sum up, we found that cognitive feedback presenting individual task performance (recall only, recall and JOL) for each studied item and/or aggregated recall performance from previous participants was not effective at correcting either illusion (Experiments 1 and 2). In contrast, additional metacognitive feedback informing participants about possible metacognitions during the task, their biased nature, and ways to enhance the accuracy of their JOLs was effective for remedying the stability bias and improving the relative accuracy of JOLs (Experiment 3).

This is the first demonstration that providing participants with the JOL and recall status of each studied item organized by cues is not effective for improving the cue basis and resolution of JOLs in a subsequent cycle. A compelling reason presumably is that more in depth-knowledge about metacognition than acquired when receiving cognitive feedback about one's JOLs and recall performance is needed to improve the judgment cue basis. This is demonstrated by our finding that JOLs relied on the number of future study opportunities after metacognitive feedback in Experiment 3. In conclusion, this study shows that cognitive feedback alone is not enough for improving the cue basis and resolution of JOLs and rather additional metacognitive feedback with an in-depth explanation of biased metacognition is needed. This is a very promising direction to mending metacognitive illusions, which has proved challenging in most prior studies.

3.3. Learning New Cues Extracted From The Environment

Navarro-Báez, S., Bröder, A., & Undorf, M. (2024). *Detecting structure: Cues for metacognitive judgments are acquired via statistical learning*. [Manuscript submitted for publication]. Department of Psychology, University of Mannheim. Department of Psychology, Technical University of Darmstadt.

All data, materials, and analyses from the present manuscript are available at https://osf.io/pze9b/?view_only=c37f0bb0877742848971d646fb77b2d6. Experiment 1 was preregistered at https://osf.io/836y2. Experiment 2 was preregistered at https://osf.io/vqrt9.

Metamemory research has focused on identifying which cues are used to predict memory and how those cues are used (i.e., analytically, non-analytically). Many cues have been found to underlie JOLs in laboratory studies such as concreteness (Begg et al., 1989; Witherby & Tauber, 2017), word frequency (Benjamin, 2003), or word pair relatedness (Undorf & Erdfelder, 2015). Further, there is ample evidence that cues are used directly via beliefs about memory and/or indirectly via experiences of 'ease' during learning (e.g., Frank & Kuhlmann, 2017; Mueller et al., 2014; Undorf et al., 2017; Undorf & Zimdahl, 2019). However, the question of cue learning has only been addressed in situations where there are systematic dissociations between metamemory and memory, by attempts to correct people's beliefs about memory, as previously discussed (e.g., Dunlosky & Hertzog, 2000; Koriat & Bjork, 2006a; Kornell & Bjork, 2009, Experiment 8; Rhodes & Castel, 2008, Experiment 4). Thus, much knowledge remains to be gained about how people learn information that they can rely on when making metamemory judgments.

In this manuscript, we investigated statistical learning as a mechanism for acquiring cues for JOLs. *Statistical learning* (SL) is defined as the extraction of statistical regularities from the environment (Saffran, Aslin, et al., 1996). It happens from repeated exposure to environmental input, without explicit instructions, feedback, social reinforcement, or intentional effort to learn (Batterink et al., 2015). SL was originally examined in the context of language acquisition as an experiential mechanism for segmenting fluent speech into words (Saffran, Aslin, et al., 1996; Saffran et al., 1997; Saffran, Newport, et al., 1996), but it has also been found to play a substantial role in other cognitive tasks such as visual search (Jones & Kaschak, 2012), sequence learning (Stadler, 1992), and causal learning (Sobel & Kirkham, 2007).

One type of regularity often used by SL studies is transitional probabilities. Transitional probabilities describe the predictive relationship between two elements such as syllables. For example, in the sequence “prettybaby”, assuming sufficient experience with English, “pre” is more predictive of “ty” than “ty” is of “ba”. This indicates that it is more likely that “pretty” is a word than “tyba” or “prettyba” are words. In fact, syllables with high transitional probabilities are likely to be part of the same word (Swingley, 2005).

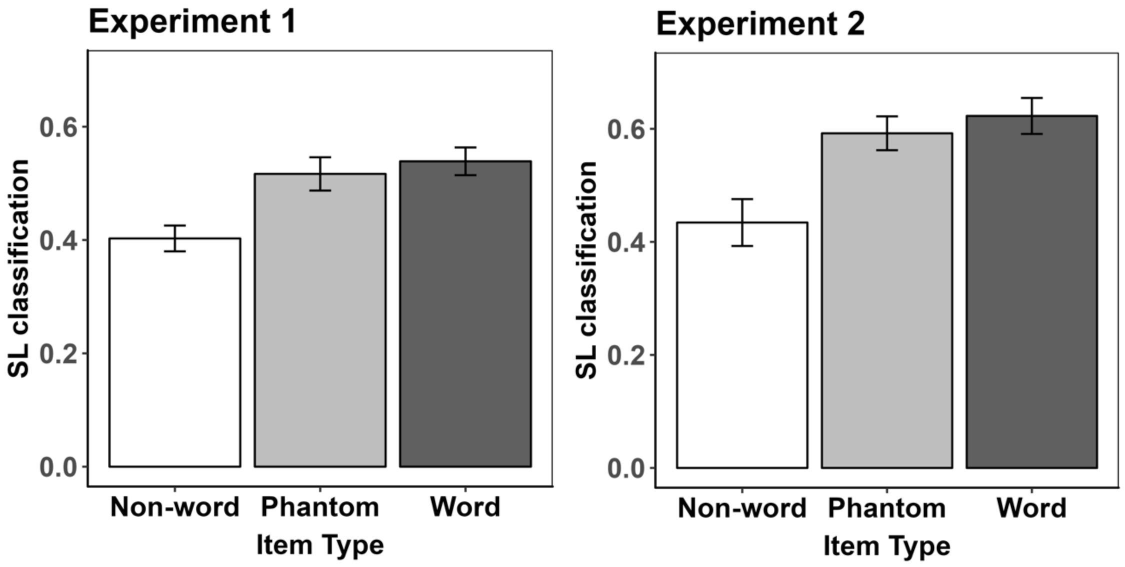
In a typical SL task, participants are exposed to input (e.g., auditory, pictorial) that is controlled by the researcher ensuring that it does not contain any other information than from the statistical information in transitional probabilities. In auditory SL studies, participants listen to a continuous stream of repeating three-syllable artificial words with high transitional probabilities within words and low transitional probabilities between words (i.e., certain syllables are more likely to appear together within a word than between words). Importantly, the auditory stream does not contain any other acoustic information such as pauses or stress differences. Above chance performance in a forced-choice-task in which each artificial word from the stream is presented together with a word foil that has the same syllables but does not follow the transitional probabilities from the input indicates that SL has taken place (e.g., Batterink et al., 2015; Endress & Mehler, 2009; Ordin & Polyanskaya, 2021; Perruchet & Poulin-Charronnat, 2012). There is evidence that learners do not show explicit verbal knowledge about transitional probabilities (Brady & Oliva, 2008; Conway & Christiansen, 2005; Turk-Browne et al., 2005). The literature strongly suggests that participants correctly choose words that follow transitional probabilities as part of the language learned based on an acquired wordlike quality even if those words have not been heard before (Endress & Mehler, 2009).

Across two experiments (N = 87, and N = 85), we combined the auditory SL paradigm (Saffran et al., 1997; Saffran, Newport, et al., 1996) and a metamemory task with JOLs (Undorf et al., 2018). In a familiarization phase, participants were exposed to a language that consisted of a continuous auditory stream of artificial words with fixed 0.5 transitional probabilities between adjacent syllables (e.g., *ro* → *se*, or *ro* → *ka*). Afterwards, they studied and made JOLs for items that were presented in the familiarization phase and follow the transitional probabilities (‘word’, e.g., *rosenu*), for items that were not presented in the familiarization phase but follow the transitional probabilities (‘phantom’, e.g., *roseti*), and for items that were neither presented in the familiarization phase nor followed the transitional probabilities (‘non-word’, e.g., *tasefa*). To verify whether participants learned the statistical structure of the language in the familiarization phase, we asked one group of participants to indicate whether each study item belonged to the language of the auditory stream before making their JOL (‘SL-assessment’ group). To rule out the possibility that SL effects on JOLs were only due to prompting participants to think about the language, another group made JOLs only (‘no SL- assessment’ group). Finally, all participants completed a recognition memory test (Experiment 1) or a 2-alternative-forced-choice memory test (Experiment 2). Based on prior SL findings (Endress & Mehler, 2009; Ordin & Polyanskaya, 2021), we expected that if SL takes place, words and phantoms will be perceived as wordlike because they follow transitional probabilities, and thus, they will be classified more often as belonging to the language than non-words. Further, we expected that JOLs would be influenced by wordlikeness. JOLs would be higher for words and phantoms than for non-words in both groups.

Experiment 1 and 2 results showed clear and marked SL effects. On average, words and phantoms were more likely to be classified as part of the language than non-words (see Figure 9). Further, SL classifications did not differ between words and phantoms indicating that SL classifications were based on the learned statistical structure of items rather than on the increased familiarity of words heard in the familiarization phase.

Figure 9

Mean proportion of SL classifications for each type of item in the SL-assessment group in Experiment 1 and Experiment 2

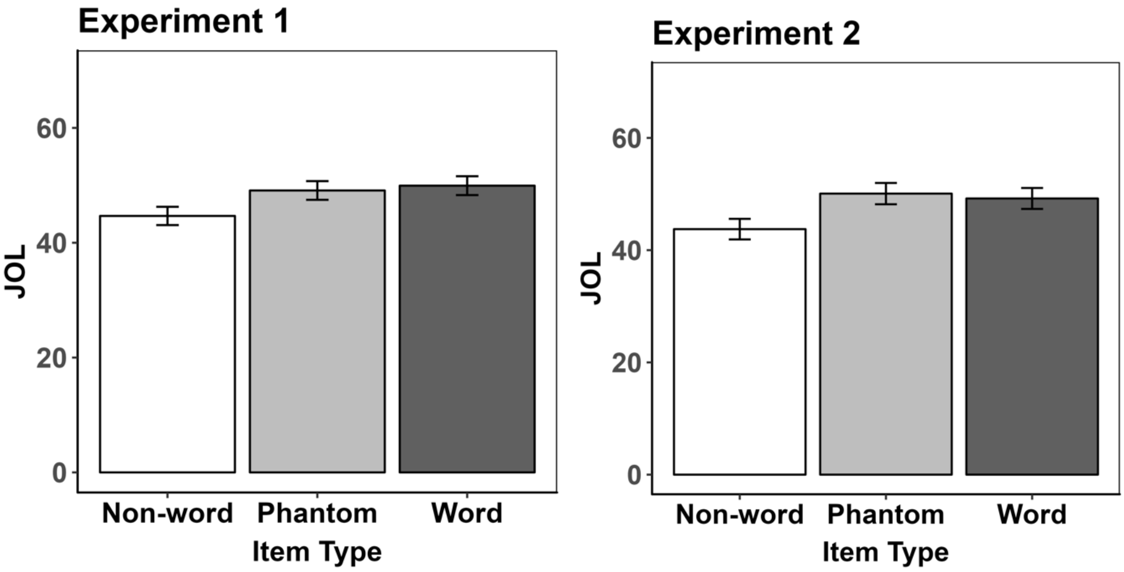


Note. Error bars represent one standard error of the mean.

JOLs were also clearly influenced by the wordlikeness of items arising from the transitional probabilities learned via SL: they were higher for words and phantoms than for non-words and did not differ between words and phantoms (see Figure 10). Additionally, mediation analyses showed that SL as assessed through SL classifications mediated the relationship between item type and JOLs. Notably, the non-significant interaction between item type and group on JOLs suggested that SL also occurred in the no-SL-assessment group. Overall, metamemory results indicated that participants learned the statistical relations between syllables which added a wordlike quality to the items and made JOLs on this basis.

Figure 10

Mean JOL for each type of item collapsed across the no-SL-assessment and SL-assessment groups in Experiment 1 and Experiment 2



Note. Error bars represent one standard error of the mean.

Regarding memory performance, in Experiment 1, recognition memory performance showed a different pattern of results than SL classifications and JOLs: Discrimination (*Pr*) was better for non-words than for words and phantoms (see Table 3). This was probably the case because items coherent with the statistical structure of the language (words and phantoms) appeared highly familiar in the test, regardless of whether they were studied or new. This probably impaired discrimination and promoted a lenient response bias. In contrast, high familiarity of non-words could stem only from their occurrence in the study list, which allows for accurate recognition memory responses and yielded better discrimination and a stricter response bias for non-word items.

Table 3

Mean (SDs) of Hits, FAs, Pr, and Br for each type of item in the No SL-assessment and SL-assessment group in Experiment 1

Group and measure	Item Type		
	Non-word	Phantom	Word
No-SL assessment			
Hits	.65 (.17)	.78 (.12)	.75 (.14)
FAs	.12 (.18)	.35 (.22)	.36 (.24)
Pr	.53 (.18)	.43 (.22)	.40 (.24)
Br	.19 (.28)	.58 (.27)	.53 (.28)
SL assessment			
Hits	.59 (.19)	.71 (.19)	.72 (.19)
FAs	.19 (.21)	.42 (.25)	.50 (.30)
Pr	.40 (.21)	.29 (.25)	.22 (.30)
Br	.25 (.25)	.57 (.27)	.60 (.25)

Experiment 2 used a 2-alternative-forced-choice (2-AFC) memory test to prevent any response tendencies or biases based on the type of item by having target and distractor items from the same type in each trial (i.e, word-word, phantom-phantom, or non-word-non-word). Results showed that performance in the 2-AFC test was similar for words, phantoms, and non-words (see Table 4). Discarding response bias in the 2-AFC memory test thus provided clear evidence that basing JOLs on wordlikeness resulted in discrepancies between JOLs and actual memory performance.

Table 4

Mean percentage of correct responses in the 2-AFC test for each type of item in the No SL-assessment and SL-assessment group in Experiment 2

Item Type	% correct	
	No-SL assessment	SL-assessment
Non-word	85.82 (34.95)	75 (43.40)
Phantom	79.08 (40.75)	81.14 (39.20)
Word	81.56 (38.85)	75.88 (42.88)

In this manuscript, we demonstrated that information on which metamemory judgments rely on can be acquired through SL. This is relevant for metamemory in real-world learning because regularities or patterns in the environment are abundant, and SL processes are present in many everyday situations. For instance, it has been demonstrated that SL supports the initial acquisition of word forms in second language learning (Alexander et al., 2023). It is thus possible that natural lexical regularities inform metacognitive judgments when learning a second language. In conclusion, this manuscript uncovers SL as an experiential mechanism for extracting regularities from the environment and using them as cues to predict memory performance even if those cues are invalid. It remains to be examined whether items congruent to the statistical structure learned have a memory advantage when measuring memory performance with another test such as free recall. This study lays the foundation for future research on statistical learning and metacognition.

4. General Discussion

In this thesis, various experimental paradigms were employed to investigate how cues informing metamemory judgments are learned. The three manuscripts contribute to the literature with important findings regarding the role of individual experiences for the acquisition of metacognitive knowledge about generic item memorability, regarding ways to mend metacognitive illusions, and regarding mechanisms through which environmental cues informing metamemory are acquired. The manuscripts of this thesis converge in underscoring the importance of cue learning in metacognition, referring not only to the acquisition of information in the form of cues but also to the use of such information for making monitoring judgments about learning and memory. The literature review and empirical data from this thesis aims to advance theoretical understanding of metamemory regarding how cue learning occurs.

In Manuscript I, we tested whether judgments of learning (JOLs) and memorability judgments (MJs) made for different memory criteria (one's own memory vs. generic item memorability) and made during different tasks (learning task vs. judgment-only task) would differ in their cue bases and relative accuracy. We ensured that the methods were similar in all other respects (i.e., judgment scale, pictorial materials, and memory criterion value). Using a within-subjects design with a JOL task and an MJ task in counterbalanced order, we found that people can predict not only their own future memory performance for scene pictures but also the general memorability of scene pictures with moderate accuracy and similar cue basis. Thus, demonstrating that discrepant findings on the accuracy of JOLs and MJs reported in prior work were largely due to methodological differences across prior studies. Crucially, MJs were more accurate and sensitive to valid cues after learning and testing in a JOL task. This shows that knowledge about general memorability that enhances MJs is acquired from experiences with one's own learning and testing.

In Manuscript II, we tested whether cognitive feedback known to correct illusions in judgments about the external world would alleviate illusions in JOLs. Participants studied single words, made JOLs, and completed a recall test across three study-test cycles with different study lists. Surprisingly, across two experiments, we showed that cognitive feedback about the recall status and JOL given to each studied item does not alleviate the font size illusion, the stability bias, and the font format illusion. In contrast, additional metacognitive feedback informing learners about possible metacognitions during the task, their biased nature, and ways to enhance JOLs, remedied the stability bias and thereby increased relative accuracy. In conclusion, this is the first demonstration that cognitive feedback on memory performance and JOLs at the item-level is insufficient for improving JOLs. An explanation is that more substantial declarative knowledge about metacognition than the one given in cognitive feedback is required to alleviate illusions. This is a promising approach to mending metacognitive illusions.

In Manuscript III, we tested statistical learning as a mechanism to learn cues informing JOLs. To this end, participants were exposed to a continuous stream with artificial three-syllable words with fixed transitional probabilities between adjacent syllables. Afterwards, they studied, made JOLs, and completed a memory test for words that follow the transitional probabilities either presented in the stream or not, and for words that do not follow the transitional probabilities. Results showed that JOLs were based on the wordlikeness quality arising from the transitional probabilities statistically learned: JOLs were higher for items following the transitional probabilities than not following them. Intriguingly, recognition memory performance was better for items not following the transitional probabilities than for items following them. Further, performance in a 2AFC test did not differ between item types. Thus, we found no correspondence of JOLs based on wordlikeness with actual recognition performance showing that the acquired wordlikeness of items was misleading. In conclusion, this is the first demonstration that cues influencing metacognition are acquired via statistical learning. This opens the field for future research on statistical learning and metacognition.

Taken together, the results of the three manuscripts presented in this thesis provide relevant knowledge about the acquisition of cues for metamemory judgments. As previously mentioned, while much is known about which and how cues are used in metamemory judgments, less is known about cue acquisition. The experimental evidence from each manuscript in this thesis contributes to close this gap.

4.1. Open Questions

Although each manuscript has implications on its own, it is important to integrate the common aspects in the three manuscripts to derive general conclusions about how cue learning in metamemory occurs. Further, this integration raises new questions and open new possibilities for future research. In the following, I will discuss some of the common aspects, open questions, and provide ideas for future research.

4.1.1. Learning From Experience?

Across the three manuscripts, cue learning from experience either occurred or did not occur. To illustrate this, in Manuscript I, cue learning occurred from experience with one's own learning and testing. In Manuscript II, cue learning did not occur from experience across multiple study-test cycles. Finally, in Manuscript III, cue learning occurred from experience with the environment. From a quick glimpse, this seems inconsistent, and one may ask if and under which conditions people really learn cues from experience. To answer this question, it is important to clarify the nature of the experience, and the cue content learned.

First, I will discuss Manuscripts I and II which are similar in the quality of the experience. This experience refers to personal experience with a standard prospective metamemory task with three components: study phase, making of JOLs, and test phase. Previous research has mostly been conducted using the same study lists across study-test cycles. This research indicates that test experience is especially relevant for improving JOLs because participants can rely on past performance, retrieval success or failure, for subsequent memory predictions (Finn & Metcalfe, 2007, 2008). In particular, two studies directly comparing the contributions of a study phase against a test phase indicate that testing experience is more effective than learning experience for enhancing JOL relative accuracy (Jang et al., 2012; Koriat & Bjork, 2006b). There is also evidence that self-paced study time is learned and used as a cue for metamemory judgments about others' memory only after learners have made JOLs for their own memory (Koriat & Ackerman, 2010; Undorf & Erdfelder, 2011). In Manuscript I, we disentangled the different components of the JOL task and found that the learning and testing phase enhanced the accuracy of generic item memorability predictions for a new set of items. Further, the linear mixed model analysis indicated that MJs aligned more with diagnostic cues after having a test than after making JOLs.

So, why then weren't illusions in Manuscript II remedied by experience with a previous JOL task with a learning phase and a recall test? One plausible explanation is the metamemory judgment target. In Manuscript I, the cue learning effects were observed in metamemory judgments made for the generic item of memorability (MJs) but not for memory predictions about one's own memory performance (JOLs). It might be that any knowledge acquired from task experience is easier to generalize to judgments about other's memory in contrast to one's own memory. This is because experiences during learning may interfere with the application of recently acquired knowledge in judgments about one's own cognition. We can find evidence for this in the *curse of knowledge* literature suggesting that people are very prone to use their own privileged knowledge and experience to judge what other people know even if this is erroneous (Birch & Bloom, 2003; Kelley & Jacoby, 1996). Additionally, generalizing knowledge to MJs might be easier than to JOLs because participants might rely more on past experiences when judging the general memorability of each item, rather than when predicting their chances of remembering the items in the future. This would explain why participants do not improve their JOLs after having had experiences with a previous study-test cycle in Manuscript II.

Another plausible explanation is that participants do not learn so much from experience in a free recall test (Manuscript II) in contrast to a recognition memory test (Manuscript I). As previously mentioned, previous research has mainly focused on experience across cycles with repeated study lists but not with new study lists. The two studies suggesting that test experience is more relevant than study experience used a cue-recall test (Jang et al., 2012; Koriat & Bjork, 2006b). In a cued-recall test and a recognition memory test, participants may more easily identify which are the items that they better/worse recall when cued or better/worse recognize when seen. This can be helpful for learning valid cues. For instance, a participant might realize during the test that she recognizes interesting pictures or pictures with people better than others. In contrast, learning valid cues from a free recall test might be more complex where one has the task to retrieve and search items in memory at the same time. This increase in cognitive load may hinder cue learning from a free recall test experience. Future studies could test this prediction.

Finally, alleviating illusions in Manuscript II involve that participants learn about the relations between the cues and recall as the criterion variable, this is, the cue validities. While participants learn from study and test experience in Manuscript I, we cannot be sure whether they explicitly learned cue-memory relations (i.e., cue validities). One possibility is that some cue information have reached the level of conscious awareness in the form of beliefs (e.g., Mueller et al., 2013, 2014). At the same time, it is also possible that cues have remained experiential at the level of subjective feelings that may not be fully articulated but serve as an inferential cue for metamemory judgments (e.g., Besken, 2016; Koriat & Levy-Sadot, 1999; Undorf et al., 2017). The same argument applies for Manuscript III, participants based their JOLs on statistical regularities recently learned. However, no participant except for one was able to articulate the syllable regularities. Thus, it is likely that the statistical regularities triggered feelings of fluency (Forest et al., 2022). In sum, experience may be helpful for learning cue information that probably remains at the experiential level, but not to develop explicit knowledge about cue-memory relations.

In conclusion, there are three possible answers to the question of whether learning from experience occurs; 1) learning from one's own previous experience may depend on the judgment target in which the cues acquired are transferred (i.e., likely for judgments about general item memorability, but not so likely for judgments about one's own memory), 2) learning from experience may depend on the content learned (i.e., likely for implicit knowledge at the level of subjective feelings, but not so likely for explicit knowledge about cue-memory relations), and 3) learning from experience may more likely occur from a cued-recall or a recognition test than from a free recall test. Further research would be needed to evaluate these predictions.

4.1.2. Learning From Feedback?

Another difference across the manuscripts is whether learning occurred from feedback or not. In Manuscript II, not only cognitive feedback on recall performance and JOLs at the item-level but also metacognitive feedback informing about illusory metacognition was required for alleviating the stability bias. In contrast, in Manuscript III, statistical learning took place and was reflected on JOLs without feedback or intention to learn. Also, in Manuscript I, explicit feedback was not required for participants to learn about the general memorability of items.

So, why then was a very extreme form of feedback (cognitive plus metacognitive feedback) required for cue learning to occur in Manuscript II? A plausible explanation to this is that very explicit cue-memory relations must be learned to remedy metacognitive illusions. To illustrate this, participants must learn that the number of future study opportunities is positively related with actual recall performance, and that the font size of words is not related to recall performance. This requires a deep or clear understanding of the cue validities that was only achieved with the additional use of metacognitive feedback. In contrast, in Manuscript I and Manuscript II, it is still possible that participants did not explicitly learn relations between cues and memory, but rather acquired a general sense of which are the pictures most likely to be recognized or which are the words that sound more memorable. The current thesis focused on learning mechanisms, *how people learn cues for their metamemory judgments*, but not on the inferential cue basis of the judgments (i.e., analytic and/or non-analytic). A second step, after having demonstrated that learning has taken place, is to examine the inferential cue basis. This would allow one to resolve whether knowledge at the level of explicit cue-memory relations can be acquired without feedback or not. This would be a worthwhile endeavor for future research.

4.1.3. What is the Cue Content Learned?

Finally, as previously mentioned, the cue knowledge acquired could be very different in nature. On the one hand, one can develop an intuition about which study items are more memorable than others without explicit knowledge about the cues that make the study item memorable (Koriat & Levy-Sadot, 1999). On the other hand, one can learn the specific cues that make study items more memorable. The latter implies that specific characteristics of the items are recognized and identified as memorable. This thesis focused on the acquisition of information of either kind of learning that influences item-by-item metamemory judgments. However, in future research, it is important to examine the nature of what was learned. This can be done by using one of the methods in the literature to assess whether a cue influences metamemory judgments via beliefs, for instance; global predictions and postdictions (Frank & Kuhlmann, 2017), post-experimental questionnaires (Undorf et al., 2017b), or vignette descriptions with global predictions (Mieth et al., 2021; Mueller et al., 2014). Since independent measures of fluency are important to obtain, a method to assess fluency such as feelings of "ease" (Alter & Oppenheimer, 2009) could be used, for instance; self-paced study times (Undorf & Erdfelder, 2015), or trials to acquisition (Undorf & Erdfelder, 2015).

For example, in Manuscript I and Manuscript II, this could be done by soliciting global predictions with a vignette description (Mieth et al., 2021; Mueller et al., 2014). In the vignette, participants are presented with a short description of the study. Then, they are asked to predict the number of items that they would remember by cue (e.g., *number of beautiful and not beautiful pictures that you would recognize, number of words in large and small font size that you would recall*). This approach could be used in combination with a fluency measure such as reaction times in a visual discrimination task with pictures in Manuscript I, or self-paced study times in Manuscript II. In Manuscript III, one could use a measure of fluency based on reaction times (Alexander et al., 2023; Batterink et al., 2015), where participants are presented with an auditory sequences of syllables and are asked to press a key whenever they hear a target item (i.e., following transitional probabilities or not). If reaction times are faster for items following transitional probabilities and this mediates the relation between transitional probabilities and JOLs, this would indicate that differences in processing fluency were acquired from statistical learning. To measure explicit cue knowledge acquired, the item stems with the first two syllables can be presented to participants for them to fill in the missing syllable. If participants can fill in the stem with above chance accuracy, this would indicate that they have acquired explicit knowledge about the statistical regularities. Further, once that it has been demonstrated that they have acquired explicit knowledge about the statistical regularities, a vignette explaining the item types and global predictions for these types can be used to measure whether they believe that items conforming to the statistical rules are more memorable than those that do not.

To sum up, clarifying what is the nature of the cue content learned (i.e., general sense of stimuli memorability, explicit cue-memory relations) would shed light on when and in which situations learning from experience and/or feedback can take place.

4.2. Further Future Directions

Next, I will outline additional future research directions inspired by the findings of this thesis.

4.2.1. From Accurate Cue Learning to Effective Regulation

According to the Nelson and Narens' (1990) framework of metamemory, students first monitor their learning and then control their learning by making decisions about how to study, what to study, and for how long to study. If monitoring is accurate, better decisions about study can be made. Several studies have demonstrated that participants can effectively control their learning when monitoring is accurate. For instance, better memory performance is observed when participants are given the opportunity to self-pace their study rather than in an experiment-paced condition or an others-paced condition (Koriat et al., 2006; Mazzoni & Cornoldi, 1993; Tullis & Benjamin, 2011). The same is true for item selection for restudy. Participants recall more items when their restudy choices are honored compared to dishonored (Dunlosky et al., 2021; Kornell & Metcalfe, 2006)

It is relevant to examine whether effects of cue learning from one's own learning and testing experience (Manuscript I), from metacognitive feedback (Manuscript II), and from statistical learning (Manuscript III) translate to effective control. For instance, this could be done by allowing participants to self-pace their study time. Effects of cue learning on control would be shown if, in addition to improved JOLs, memory performance is better in a self-paced condition than in an experimenter-paced condition after cue learning has occurred.

4.2.2. Effective Learning Paradigms

As previously mentioned, the research on cue learning in metamemory is still in its infancy. The work of this thesis shows that metacognitive feedback and learning mechanisms such as statistical learning are promising directions about how cues informing metamemory judgments are learned. However, future studies should urgently examine the generalizability of metacognitive feedback to other cues (e.g., study strategies). This would be helpful for knowing about the effectiveness of informing individuals about metacognition which could be tested in applied contexts such as educational settings. Further, it is important to examine whether items following learned regularities are better remembered than items not following such regularities in other memory tests such as a free recall test and a cued-recall test. This would inform about the benefits of statistical learning as a cue learning paradigm.

Finally, a closer investigation of the causal mechanisms underlying cue learning would be valuable, as learning cue-memory relations involves causal learning. Attributions of stimulus characteristics (i.e., cues) to memorability need to be made and integrated into beliefs that guide metacognitive judgments. These attributions can be facilitated, for instance, by simple reinforcement learning paradigms that combine feedback with monetary incentives. Although monetary incentives in metamemory often boost overconfidence in confidence judgments (Krawczyk, 2012; Lebreton et al., 2018, 2019) and JOLs (Bröder et al., 2024), they have never been used in conjunction with metacognitive feedback. This approach could shed light on the motivational aspects of cue learning.

4.3. Conclusion

Given the importance of accurate metacognitive monitoring for the effective regulation of behavior, this thesis investigated how people learn information to rely on for their metamemory monitoring judgments. The work in this thesis revealed that learning and testing experiences are relevant for acquiring knowledge about the general memorability of pictures. However, in metamemory judgments about one's own memory, neither previous task experience nor cognitive feedback on recall and JOL at the item level are helpful for acquiring and utilizing knowledge about the validity (or invalidity) of cues. Instead, declarative knowledge about metacognition as provided in an informative metacognitive feedback is required. Finally, pioneering research linking statistical learning and metacognition revealed that regularities extracted from experience with the environment are learned to judge one's own future memory. Altogether, the results in this thesis extend our understanding of how people acquire cues to evaluate their memory.

5. Bibliography

Alexander, E., Van Hedger, S. C., & Batterink, L. J. (2023). Learning words without trying: Daily second language podcasts support word-form learning in adults. *Psychonomic Bulletin & Review*, *30*(2), 751–762. https://doi.org/10.3758/s13423-022-02190-1

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*(1), 126–131. https://doi.org/10.1037/h0027455

Ariel, R., Hines, J. C., & Hertzog, C. (2014). Test framing generates a stability bias for predictions of learning by causing people to discount their learning beliefs. *Journal of Memory and Language*, *75*, 181–198. https://doi.org/10.1016/j.jml.2014.06.003

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of Cognitive Feedback on Performance. *Psychological Bulletin*, *106*(3), 410–433. https://doi.org/10.1037/0033-2909.106.3.410

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language*, *83*, 62–78. https://doi.org/10.1016/j.jml.2015.04.004

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610–632. https://doi.org/10.1016/0749-596X(89)90016-8

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*(2), 297–305. https://doi.org/10.3758/BF03194388

Besken, M. (2016). Picture-perfect is not perfect for metamemory: Testing the perceptual fluency hypothesis with degraded images. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1417–1433. https://doi.org/10.1037/xlm0000246

Birch, S. A. J., & Bloom, P. (2003). Children Are Cursed: An Asymmetric Bias in Mental-State Attribution. *Psychological Science*, *14*(3), 283–286. https://doi.org/10.1111/1467-9280.03436

Borkowski, J. G., Peck, V. A., Reid, M. K., & Kurtz, B. E. (1983). Impulsivity and Strategy Transfer: Metamemory as Mediator. *Child Development*, *54*(2), 459–473. https://www.jstor.org/stable/1129707

Brady, T. F., & Oliva, A. (2008). Statistical Learning Using Real-World Scenes: Extracting Categorical Regularities Without Conscious Intent. *Psychological Science*, *19*(7), 678–685. https://doi.org/10.1111/j.1467-9280.2008.02142.x

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica*, *45*(1–3), 223–241. https://doi.org/10.1016/0001-6918(80)90034-7

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

Bröder, A., Navarro-Báez, S., & Undorf, M. (2024). Reducing cheap talk? How monetary incentive affect the accuracy of metamemory judgments. Department of Psychology, University of Mannheim. Department of Psychology, Technical University of Darmstadt.

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

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*, 165–178. https://doi.org/10.1016/j.visres.2015.03.005

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*(1), 26–48. https://doi.org/10.3758/BF03210724

Caplan, J. B., Sommer, T., Madan, C. R., & Fujiwara, E. (2019). Reduced associative memory for negative information: Impact of confidence and interactive imagery during study. *Cognition and Emotion*, *33*(8), 1745–1753. https://doi.org/10.1080/02699931.2019.1602028

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, *36*(2), 429–437. https://doi.org/10.3758/MC.36.2.429

Chang, M., & Brainerd, C. J. (2022). Association and dissociation between judgments of learning and memory: A Meta-analysis of the font size effect. *Metacognition and Learning*, *17*(2), 443–476. https://doi.org/10.1007/s11409-021-09287-3

Cohen, R. L., Sandier, S. P., & Kcglcvich, L. (1991). The Failure of Memory Monitoring in a Free Recall Task. *Canadian Journal of Psychology*, *45*(4), 523–538. https://doi.org/10.1037/h0084303

Conway, C. M., & Christiansen, M. H. (2005). Modality-Constrained Statistical Learning of Tactile, Visual, and Auditory Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 24–39. https://doi.org/10.1037/0278-7393.31.1.24

Dougherty, M. R., Scheck, P., & Nelson, T. O. (2005). Using the past to predict the future. *Memory & Cognition*, *33*(6), 1096–1115. https://doi.org/10.3758/BF03193216

Dunlosky, J., & Hertzog, C. (2000). Updating Knowledge About Encoding Strategies: A Componential Analysis of Learning About Strategy Effectiveness From Task Experience. *Psychology and Aging*, *15*(3), 462–474. https://doi.org/10.1037/0882-7974.15.3.462

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. SAGE Publications. https://books.google.de/books?id=xHtJADBpp-IC

Dunlosky, J., Mueller, M. L., Morehead, K., Tauber, S. K., Thiede, K. W., & Metcalfe, J. (2021). Why Does Excellent Monitoring Accuracy Not Always Produce Gains in Memory Performance? *Zeitschrift Für Psychologie*, *229*(2), 104–119. https://doi.org/10.1027/2151-2604/a000441

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*(4), 374–380. https://doi.org/10.3758/BF03210921

Dunlosky, J., & Thiede, K. W. (2013). Metamemory. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 283–298). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0019

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351–367. https://doi.org/10.1016/j.jml.2008.10.003

Fiedler, K., Schott, M., Kareev, Y., Avrahami, J., Ackerman, R., Goldsmith, M., Mata, A., Ferreira, M. B., Newell, B. R., & Pantazi, M. (2020). Metacognitive myopia in change detection: A collective approach to overcome a persistent anomaly. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 649–668. https://doi.org/10.1037/xlm0000751

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 238–244. https://doi.org/10.1037/0278-7393.33.1.238

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*(1), 19–34. https://doi.org/10.1016/j.jml.2007.03.006

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive Psychology*, *1*(4), 324–340. https://doi.org/10.1016/0010-0285(70)90019-8

Forest, T. A., Siegelman, N., & Finn, A. S. (2022). Attention Shifts to More Complex Structures With Experience. *Psychological Science*, *33*(12), 2059–2072. https://doi.org/10.1177/09567976221114055

Frank, D. J., & Kuhlmann, B. G. (2017). More than just beliefs: Experience and beliefs jointly contribute to volume effects on metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 680–693. https://doi.org/10.1037/xlm0000332

Gilewski, M. J., Zelinski, E. M., & Schaie, K. W. (1990). The Memory Functioning Questionnaire for assessment of memory complaints in adulthood and old age. *Psychology and Aging*, *5*(4), 482–490. https://doi.org/10.1037/0882-7974.5.4.482

Groninger, L. D. (1976). Predicting recognition during storage: The capacity of the memory system to evaluate itself. *Bulletin of the Psychonomic Society*, *7*(5), 425–428. https://doi.org/10.3758/BF03337236

Groninger, L. D. (1979). Predicting Recall: The "Feeling-That-I-Will-Know" Phenomenon. *The American Journal of Psychology*, *92*(1), 45–58. https://doi.org/10.2307/1421478

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*(4), 208–216. https://doi.org/10.1037/h0022263

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning and Verbal Behavior*, *6*(5), 685–691. https://doi.org/10.1016/S0022-5371(67)80072-0

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 22–34. https://doi.org/10.1037/0278-7393.29.1.22

Hertzog, C., Price, J., Burpee, A., Frentzel, W. J., Feldstein, S., & Dunlosky, J. (2009). Why do people show minimal knowledge updating with task experience: Inferential deficit or experimental artifact? *Quarterly Journal of Experimental Psychology*, *62*(1), 155–173. https://doi.org/10.1080/17470210701855520

Hourihan, K. L. (2020). Misleading emotions: Judgments of learning overestimate recognition of negative and positive emotional images. *Cognition and Emotion*, *34*(4), 771–782. https://doi.org/10.1080/02699931.2019.1682972

Hourihan, K. L., & Bursey, E. (2017). A misleading feeling of happiness: Metamemory for positive emotional and neutral pictures. *Memory*, *25*(1), 35–43. https://doi.org/10.1080/09658211.2015.1122809

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011a). Understanding the intrinsic memorability of images. *Advances in Neural Information Processing Systems*, *24*, 2429–2437. https://doi.org/10.1167/12.9.1082

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011b). What makes an image memorable? *24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 145–152.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*(1), 186–200. https://doi.org/10.1037/a0025960

Jones, J. L., & Kaschak, M. P. (2012). Global statistical learning in a visual search task. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 152–160. https://doi.org/10.1037/a0026233

Kao, Y.-C., Davis, E. S., & Gabrieli, J. D. E. (2005). Neural correlates of actual and predicted memory formation. *Nature Neuroscience*, *8*(12), 1776–1783. https://doi.org/10.1038/nn1595

Karlsson, L., Juslin, P., & Olsson, H. (2004). Representational Shifts in a Multiple-Cue Judgment Task with Continuous Cues. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*(26), 648–653.

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*(1), 92–107. https://doi.org/10.3758/BF03211579

Kelley, C. M., & Jacoby, L. L. (1996). Adult Egocentrism: Subjective Experience versus Analytic Bases for Judgment. *Journal of Memory and Language*, *35*(2), 157–175. https://doi.org/10.1006/jmla.1996.0009

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of Knowing: The Influence of Retrieval Practice. *The American Journal of Psychology*, *93*(2), 329–343. https://doi.org/10.2307/1422236

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

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge Handbook of Consciousness* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511816789.012

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for Self and Other during self-paced study. *Consciousness and Cognition*, *19*(1), 251–264. https://doi.org/10.1016/j.concog.2009.12.010

Koriat, A., & Bjork, R. A. (2006a). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1133–1145. https://doi.org/10.1037/0278-7393.32.5.1133

Koriat, A., & Bjork, R. A. (2006b). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959–972. https://doi.org/10.3758/BF03193244

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting One's Own Forgetting: The Role of Experience-Based and Theory-Based Processes. *Journal of Experimental Psychology: General*, *133*(4), 643–656. https://doi.org/10.1037/0096-3445.133.4.643

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36–69. https://doi.org/10.1037/0096-3445.135.1.36

Koriat, A., & Levy-Sadot, R. (1999). Processes Underlying Metacognitive Judgments: Information based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 483-502). New York : Guilford Publications

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing Objective and Subjective Learning Curves: Judgments of Learning Exhibit Increased Underconfidence With Practice. *Journal of Experimental Psychology: General*, *131*(2), 147–162. https://doi.org/10.1037/0096-3445.131.2.147

Kornell, N., & Bjork, R. A. (2009). A Stability Bias in Human Memory: Overestimating Remembering and Underestimating Learning. *Journal of Experimental Psychology: General*, *138*(4), 449–468. https://doi.org/10.1037/a0017350

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 609–622. https://doi.org/10.1037/0278-7393.32.3.609

Krawczyk, M. (2012). Incentives and timing in relative performance judgments: A field experiment. *Journal of Economic Psychology*, *33*(6), 1240–1246. https://doi.org/10.1016/j.joep.2012.09.006

Kurtz, B. E., & Borkowski, J. G. (1987). Development of strategic skills in impulsive and reflective children: A longitudinal study of metacognition. *Journal of Experimental Child Psychology*, *43*(1), 129–148. https://doi.org/10.1016/0022-0965(87)90055-5

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLOS Computational Biology*, *15*(4), e1006973. https://doi.org/10.1371/journal.pcbi.1006973

Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., van Holst, R. J., & Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Sciences Advances, 4*(5). eaaq0668. https://doi.org/10.1126/sciadv.aaq0668

Leonesio, R. J., & Nelson, T. O. (1990). Do Different Metamemory Judgments Tap the Same Underlying Aspects of Memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 464–470. https://doi.org/10.1037/0278-7393.16.3.464

Luna, K., Martín-Luengo, B., & Albuquerque, P. B. (2018). Do delayed judgements of learning reduce metamemory illusions? A meta-analysis. *Quarterly Journal of Experimental Psychology*, *71*(7), 1626–1636. https://doi.org/10.1080/17470218.2017.1343362

Luna, K., Nogueira, M., & Albuquerque, P. B. (2019). Words in larger font are perceived as more important: Explaining the belief that font size affects memory. *Memory*, *27*(4), 555–560. https://doi.org/10.1080/09658211.2018.1529797

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 509–527. https://doi.org/10.1037/a0014876

Matvey, G., Dunlosky, J., Shaw, R. J., Parks, C., & Hertzog, C. (2002). Age-related equivalence and deficit in knowledge updating of cue effectiveness. *Psychology and Aging*, *17*(4), 589–597. https://doi.org/10.1037/0882-7974.17.4.589

Mazzoni, G., & Cornoldi, C. (1993). Strategies in Study Time Allocation: Why Is Study Time Sometimes Not Effective? *Journal of Experimental Psychology: General*, *122*(1), 47–60. https://doi.org/10.1037/0096-3445.122.1.47

McDonough, I. M., Enam, T., Kraemer, K. R., Eakin, D. K., & Kim, M. (2021). Is there more to metamemory? An argument for two specialized monitoring abilities. *Psychonomic Bulletin & Review*, *28*(5), 1657–1667. https://doi.org/10.3758/s13423-021-01930-z

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Mieth, L., Schaper, M. L., Kuhlmann, B. G., & Bell, R. (2021). Memory and metamemory for social interactions: Evidence for a metamemory expectancy illusion. *Memory & Cognition*, *49*(1), 14–31. https://doi.org/10.3758/s13421-020-01071-z

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2015). Why is knowledge updating after task experience incomplete? Contributions of encoding experience, scaling artifact, and inferential deficit. *Memory & Cognition*, *43*(2), 180–192. https://doi.org/10.3758/s13421-014-0474-2

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, *20*(2), 378–384. https://doi.org/10.3758/s13423-012-0343-6

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1287–1306. https://doi.org/10.1037/a0036914

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., & Dunlosky, J. (1991). When People's Judgments of Learning (JOLs) are Extremely Accurate at Predicting Subsequent Recall: The "Delayed-JOL Effect." *Psychological Science*, *2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Narens. (1990). Metamemory: A Theoretical Framework and New Findings. In *Psychology of Learning and Motivation* (Vol. 26, pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5

Newell, B. R., Weston, N. J., Tunney, R. J., & Shanks, D. R. (2009). The Effectiveness of Feedback in Multiple-Cue Probability Learning. *Quarterly Journal of Experimental Psychology*, *62*(5), 890–908. https://doi.org/10.1080/17470210802351411

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *19*(2), 155–160. https://doi.org/10.1037/h0082899

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Ordin, M., & Polyanskaya, L. (2021). The role of metacognition in recognition of the content of statistical learning. *Psychonomic Bulletin & Review*, *28*(1), 333–340. https://doi.org/10.3758/s13423-020-01800-0

Pan, S. C., & Rivers, M. L. (2023). Metacognitive awareness of the pretesting effect improves with self-regulation support. *Memory & Cognition*, *51*(6), 1461–1480. https://doi.org/10.3758/s13421-022-01392-1

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. https://doi.org/10.1016/j.jml.2012.02.010

Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency – Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction*, *44*, 31–40. https://doi.org/10.1016/j.learninstruc.2016.01.012

Price, J., & Harrison, A. (2017). Examining what prestudy and immediate judgments of learning reveal about the bases of metamemory judgments. *Journal of Memory and Language*, *94*, 177–194. https://doi.org/10.1016/j.jml.2016.12.003

Price, J., Hertzog, C., & Dunlosky, J. (2008). Age-Related Differences in Strategy Knowledge Updating: Blocked Testing Produces Greater Improvements in Metacognitive Accuracy for Younger than Older Adults. *Aging, Neuropsychology, and Cognition*, *15*(5), 601–626. https://doi.org/10.1080/13825580801956225

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615–625. https://doi.org/10.1037/a0013684

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review*, *16*(3), 550–554. https://doi.org/10.3758/PBR.16.3.550

Rhodes, M. G. (2016). Judgments of Learning: Methods, Data, and Theory. In J. Dunlosky & S. (Uma) K. Tauber (Eds.), *The Oxford Handbook of Metamemory* (1st ed., pp. 65-80). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.4

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615–625. https://doi.org/10.1037/a0013684

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review*, *16*(3), 550–554. https://doi.org/10.3758/PBR.16.3.550

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, *35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental Language Learning: Listening (and Learning) Out of the Corner of Your Ear. *Psychological Science*, *8*(2), 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x

Schmoeger, M., Deckert, M., Loos, E., & Willinger, U. (2020). How influenceable is our metamemory for pictorial material? The impact of framing and emotionality on metamemory judgments. *Cognition*, *195*(104112), 1–10. https://doi.org/10.1016/j.cognition.2019.104112

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review*, *1*(3), 357–375. https://doi.org/10.3758/BF03213977

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, *38*(7–8), 608–625. https://doi.org/10.1016/j.ergon.2008.01.007

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 156–163. https://doi.org/10.1016/S0022-5371(67)80067-7

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of Measuring Recognition Memory: Applications to Dementia and Amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science*, *10*(3), 298–306. https://doi.org/10.1111/j.1467-7687.2007.00589.x

Spellman, B. A., & Bjork, R. A. (1992). When Predictions Create Reality: Judgments of Learning May Alter What They Are Intended to Assess. *Psychological Science*, *3*(5), 315–317. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

Stadler, M. A. (1992). Statistical Structure and Implicit Serial Learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*(2), 318–327. https://doi.org/10.1037/0278-7393.18.2.318

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207–222. https://doi.org/10.1080/14640747308400340

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, *19*(2), 73–74. https://doi.org/10.1080/14640747308400340

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin & Review*, *18*(5), 973–978. https://doi.org/10.3758/s13423-011-0114-9

Susser, J. A., & Mulligan, N. W. (2015). The effect of motoric fluency on metamemory. *Psychonomic Bulletin & Review*, *22*(4), 1014–1019. https://doi.org/10.3758/s13423-014-0768-1

Susser, J. A., Panitz, J., Buchin, Z., & Mulligan, N. W. (2017). The motoric fluency effect on metamemory. *Journal of Memory and Language*, *95*, 116–123. https://doi.org/10.1016/j.jml.2017.03.002

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86–132. https://doi.org/10.1016/j.cogpsych.2004.06.001

Tauber, S. K., Dunlosky, J., Urry, H. L., & Opitz, P. C. (2017). The effects of emotion on younger and older adults' monitoring of learning. *Aging, Neuropsychology, and Cognition*, *24*(5), 555–574. https://doi.org/10.1080/13825585.2016.1227423

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language*, *64*(2), 109–118. https://doi.org/10.1016/j.jml.2010.11.002

Tullis, J. G., & Benjamin, A. S. (2012). The effectiveness of updating metacognitive knowledge in the elderly: Evidence from metamnemonic judgments of word frequency. *Psychology and Aging*, *27*(3), 683–690. https://doi.org/10.1037/a0025838

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429–442. https://doi.org/10.3758/s13421-012-0274-5

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language*, *95*, 124–137. https://doi.org/10.1016/j.jml.2017.03.003

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The Automaticity of Visual Statistical Learning. *Journal of Experimental Psychology: General*, *134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

Undorf, M. (2020). Fluency Illusions in Metamemory. In A. M. Cleary & B. L. Schwartz (Eds.), *Memory Quirks* (1st ed., pp. 150–174). Routledge. https://doi.org/10.4324/9780429264498-12

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629–642. https://doi.org/10.1177/1747021819882308

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264–1269. https://doi.org/10.1037/a0023719

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition*, *43*(4), 647–658. https://doi.org/10.3758/s13421-014-0479-x

Undorf, M., Navarro-Báez, S., & Bröder, A. (2022). "You don't know what this means to me" – Uncovering idiosyncratic influences on metamemory judgments. *Cognition*, *222*(105011), 1–9. https://doi.org/10.1016/j.cognition.2021.105011

Undorf, M., Navarro-Báez, S., & Zimdahl, M. F. (2022). Metacognitive illusions. In R. F. Pohl, *Cognitive Illusions* (3rd ed., pp. 307–323). Routledge. https://doi.org/10.4324/9781003154730-22

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., & Zimdahl, M. F. (2019). Metamemory and memory for a wide range of font sizes: What is the contribution of perceptual fluency? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 97–109. https://doi.org/10.1037/xlm0000571

Undorf, M., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language*, *92*, 293–304. https://doi.org/10.1016/j.jml.2016.07.003

Witherby, A. E., & Tauber, S. K. (2017). The concreteness effect on judgments of learning: Evaluating the contributions of fluency and beliefs. *Memory & Cognition*, *45*(4), 639–650. https://doi.org/10.3758/s13421-016-0681-0

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485–3492. https://doi.org/10.1109/CVPR.2010.5539970

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*(7), 918–933. https://doi.org/10.1037/xge0000177

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. https://doi.org/10.3758/BF03329756

Ziglar. Z. (1977). *See you at the top*. Pelican Publishing Company.

Zimmerman, C. A., & Kelley, C. M. (2010). "I'll remember this!" Effects of emotionality on memory predictions versus memory performance. *Journal of Memory and Language*, *62*(3), 240–253. https://doi.org/10.1016/j.jml.2009.11.004

A. Acknowledgments

> "What you get by achieving your goals is not as important as what you become by achieving your goals."
>
> **— Zig Ziglar (1977)**

This journey through my PhD has taught me much more than I expected, not only scientific skills, but also helped me grow as a person. For this, I am extremely grateful to the people who helped me making it possible and who have accompanied me along the way.

First, I would like to thank my supervisors, Arndt Bröder and Monika Undorf, for training my scientific thinking, teaching me how to conduct good experiments, and always being there to provide feedback and advice in each step of the research process. Specifically, Arndt Bröder for teaching me to keep a bird view and don't get lost in detail. Monika Undorf, for teaching me to rigorously evaluate every part of an experiment and every word in a sentence, thank you for taking the time to listen to me even when I was just organizing my thoughts. I also would like to thank the cognitive psychology group in Mannheim for creating a stimulating scientific environment. Specially, I would like to thank Edgar Erdfelder and Beatrice Kuhlmann, for their willingness to evaluate my thesis.

Second, I would like to thank my colleagues in Mannheim, Sophie, Tong, Franzi, David, Pascal, and Barbara. Sophie, for support with settling down in a new country, for your cheerful attitude, and advice. Tong, for your care and friendship. Franzi, for helping me organizing data collection since the first day and still being supportive in organizing our group in Darmstadt. Thank you all for your support and for the nice time during lunch breaks, hikes, and other social events. I also would like to thank my colleagues outside of the Mannheim experimental psychology group, Fabiola, Theresa, Pedro, and Carolin for the support and the long talks that enormously motivated me. Thanks to my friends, Nidia, Gaby, and Daniela, for visiting me during this last year when I needed fun and good times.

Finally, I would like to thank my family in Mexico who have always believed in me. Your love and care during each trip to Mexico were very energizing for me. I am grateful to my grandma, although she is not here to see me now, I can still feel one of her big hugs. Thank you, Mama, without your love I would not have come this far in pursuing my dreams. I am grateful to my father who was a great inspiration for me and encouraged my curiosity to know more about the mind since I was very young. Ultimately, thank you, Wilken, you experienced most of the ups and downs of this PhD with me and supported me in every way possible, thank you for being my safe place.

B. Statement of Originality

1. I hereby declare that the presented doctoral dissertation with the title Cue Learning in Metamemory: Understanding How People Learn to Judge Memory is my own work.
2. I did not seek unauthorized assistance of a third party and I have employed no other sources or means except the ones listed. I clearly marked any quotations derived from the works of others.
3. I did not present this doctoral dissertation or parts of it at any other higher education institution in Germany or abroad.
4. I hereby conform the accuracy of the declaration above.
5. I am aware of the significance of this declaration and the legal consequences in case of untrue or incomplete statements.

I affirm in lieu of oath that the statements above are to the best of my knowledge true and complete.

Signature:

Date: 30.07.2024

C. Co-Authors’ Statements

D. Copies of Articles

Original Article

QJEP

Quarterly Journal of Experimental Psychology
1–20
© Experimental Psychology Society 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17470218241239829
qjep.sagepub.com
Sage

Predicting the memorability of scene pictures: Improved accuracy through one's own experience

Sofia Navarro-Báez[1,2], **Monika Undorf**[2] **and Arndt Bröder**[1]

Abstract

There are conflicting findings regarding the accuracy of metamemory for scene pictures. Judgements of stimulus memorability in general (*memorability judgements* [MJs]) have been reported to be unpredictive of actual image memorability. However, other studies have found that *judgements of learning* (JOLs)—predictions of one's own later memory performance for recently studied items—are moderately predictive of people's own actual recognition memory for pictures. The current study directly compared the relative accuracy and cue basis of JOLs and MJs for scene pictures. In Experiments 1 and 2, participants completed an MJ task and a JOL task in counterbalanced order. In the MJ task, they judged the general memorability of each picture. In the JOL task, they studied pictures and made JOLs during a learning phase, followed by a recognition memory test. Results showed that MJs were predictive of general scene memorability and relied on the same cues as JOLs, but MJ accuracy considerably improved after the JOL task. Experiment 3 demonstrated that prior learning experiences drove this increase in MJ accuracy. This work demonstrates that people can predict not only their own future memory performance for scene pictures with moderate accuracy but also the general memorability of scene pictures. In addition, experiences with one's own learning and memory support the ability to assess scene memorability in general. This research contributes to our understanding of the basis and accuracy of different metamemory judgements.

Keywords

Metamemory accuracy; judgements of learning; memorability judgements; scene memorability

Received: 28 April 2023; revised: 30 October 2023; accepted: 26 February 2024

It is helpful to know which pictures are memorable. For instance, an illustrator may benefit from such knowledge when choosing pictures for advertisements. The ability to assess and to know about memory is termed as *metamemory* (Dunlosky & Thiede, 2013). An advantage of accurate metamemory is the effective regulation of future memory performance (Bjork et al., 2013). However, studies investigating metamemory accuracy of naturalistic scene pictures have yielded conflicting evidence: It is not clear how accurate people are at predicting which pictures will be remembered and which will not. Thus, although it is well known that the human visual memory storage for pictures is astonishing (Nickerson, 1965; Shepard, 1967; Standing, 1973), how good metamemory for pictures is remains to be examined.

The conflicting findings on metamemory accuracy for scene pictures stem from studies using either *memorability judgements* (MJs)—judgements of stimulus memorability in general—or *judgements of learning* (JOLs)—predictions of one's own later memory performance for recently studied items. Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) found that MJs are unpredictive of actual picture memorability. This is surprising since different people tend to remember and forget the same pictures (Isola, Parikh, et al., 2011; Isola,

[1] Department of Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany
[2] Department of Psychology, Technical University of Darmstadt, Darmstadt, Germany

Corresponding author:
Sofia Navarro-Báez, Alexanderstr. 10 (S1|15), 64283 Darmstadt, Germany.
Email: sofia.navarro@tu-darmstadt.de

Xiao, et al., 2011, 2014). In contrast, JOLs have been found to be moderately predictive of actual individual recognition memory for pictures (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). The current study aims to test whether these differences are due to MJs referring to memorability as a generic item attribute, whereas JOLs refer to one's own chances of remembering a recently studied item. This endeavour will enhance our understanding of the accuracy and basis of MJs and JOLs and extend our knowledge about different metamemory judgements.

Memorability of scene pictures

Although items may naturally vary in their actual memorability between individuals due to idiosyncratic encoding (Hintzman, 1980; Undorf et al., 2022), recent work has indicated that actual memorability of scene pictures is quite consistent across participants (Bainbridge et al., 2013; Bylinskii et al., 2015; Isola, Parikh, et al., 2011; Isola, Xiao, et al., 2011, 2014). In a large-scale series of studies, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) measured the memorability of more than 2,000 images of real-world scenes from the SUN database (Xiao et al., 2010). They used a repeat detection task in which participants saw sequences of 120 images and were asked to detect whenever there was a repetition of an image. *Image memorability* was measured as the percentage of correct detections by participants. To investigate how consistent image memorability is across participants, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) randomly split the sample into two independent halves and correlated the image memorability values from the two halves. Repeating this procedure over 25 times, the average correlation was strong ($\rho = .75$) and indicated that people tend to recognise and miss the same pictures.

Given consistency of image memorability across participants, a further step is to explain what makes an image memorable. Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) identified attributes contributing to memorability. Highly memorable images had semantic attributes such as enclosed spaces, telling a story, and people present. In contrast, less memorable images displayed open spaces, aesthetic settings, and were peaceful. Interestingly, perceptual image features such as colour (e.g., hue, saturation) and object statistics (e.g., number of objects, coverage of pixels over objects) were unrelated to memorability. Overall, image memorability was mainly predicted by the high-level semantic information conveyed in the picture (but see Lin et al., 2021). Nevertheless, a large proportion of image memorability variance remained unexplained.

MJs

If image memorability tends to be the same across participants, it is reasonable to ask whether people can assess the memorability of pictures. To address this question, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) obtained MJs in two tasks; in the first task, 30 participants were asked, "Is this a memorable image? Yes/No," and in the second task, 30 other participants were asked, "If you were to come across this image in the morning, and then happen to see it again at the end of the day, do you think you would realize that you have seen this image earlier in the day? Yes/No." Results showed that MJs did not predict image memorability: correlations between MJs and memorability were $\rho = -0.19$ in the first task and $\rho = -0.02$ in the second task. Instead, MJs were highly correlated with average ratings of semantic image attributes from a norming sample (aesthetics, $\rho = .83$; interestingness, $\rho = .86$) that were inversely related with image memorability (aesthetics, $\rho = -.36$; interestingness, $\rho = -.23$). These results suggest that people have the misconception that beautiful and interesting images are highly memorable and, more generally, indicate that people lack insight into item memorability.

JOLs

JOLs are commonly studied metamemory judgements. When people make JOLs, they predict their own future memory performance for recently studied items. Crucially, JOLs are elicited after learning each item and are compared with participant's own later memory performance. Higher-order monitoring processes of learning and memory are involved when making JOLs (Nelson & Narens, 1990). Inferential accounts of metamemory assume that JOLs are inferences based on available cues and heuristics because there is no direct access to the strength of the memory trace (Koriat, 1997). Cues for JOLs are classified into three different types (Koriat, 1997). *Intrinsic cues* are characteristics inherent to the studied items, such as word concreteness or the aesthetics of a picture. *Extrinsic cues* are related to the study conditions in which items are learned, such as the number of study repetitions and encoding strategies used. *Mnemonic cues* are sensitive to the effects of extrinsic and intrinsic cues and derive from the quality of processing items during learning, such as ease of encoding or retrieval fluency.

Evidence for inferential accounts of metamemory comes from situations in which metamemory judgements are dissociated from actual memory, leading to metamemory illusions (see Undorf, 2020; Undorf et al., 2022, for a review). For pictorial materials, there have been very few illusions found. One of them is for picture emotionality. Recognition memory performance is reduced for emotional pictures, but JOLs tend to be higher for emotional pictures compared with neutral ones (Caplan et al., 2019; Hourihan, 2020; Hourihan & Bursey, 2017). However, on a free recall test in which participants verbally described studied pictures, JOLs accurately predict better memory for emotional pictures (Schmoeger et al., 2020; Tauber et al., 2017). Thus, people recognise the positive validity of picture emotionality on free recall but fail to consider the differential negative effects of emotionality in a recognition memory test.

Accurate JOLs, in contrast, imply that these are based on cues that are predictive of people's actual memory performance (Chandler, 1994; Dunlosky & Metcalfe, 2009; Koriat, 1997). An important aspect of metamemory accuracy is relative accuracy (or resolution)—the extent to which metamemory judgements discriminate between items that will be remembered and those that will not be remembered. Importantly, most of the few JOL studies using pictures of scenes have found that JOLs are relatively accurate in terms of relative accuracy and track cue effects on actual memory performance (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). This is illustrated in the study by Undorf and Bröder (2021), in which a total of six intrinsic and extrinsic cues in pictures from the SUN database (Xiao et al., 2010) were manipulated across three experiments. Results showed that recognition memory performance was better for scenes that were contextually distinctive, coloured (vs. grayscale), telling a story, twice (vs. once) presented, and containing persons, whereas recognition memory performance was worse for peaceful scenes. At the same time, people's JOLs were higher for all cues that helped memory and only failed to reflect that peacefulness hindered memory. Moreover, JOLs showed moderate relative accuracy, suggesting that reliance on valid probabilistic cues is an important factor for relative accuracy. Similarly, studies with verbal materials found that JOLs are based on multiple cues most of which have predictive validity and are moderate in their relative accuracy (e.g., Bröder & Undorf, 2019; Koriat, 1997; Undorf et al., 2018).

Differences between JOLs and MJs

As mentioned above, prior research by Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) suggest that MJs are unpredictive of actual image memorability across participants, while JOLs are relatively accurate at predicting participants' own memory performance (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). This is interesting because both judgements refer to picture memorability and, consequently the same judgement target.

A potential reason for differences in accuracy between JOLs and MJs could be that they refer to different aspects of memorability. JOLs are predictions of one's own memory performance, whereas MJs are estimations of memorability as a general attribute (i.e., picture memorability) and do not focus on one's own experiences during learning and remembering. It is possible that people use different cues to inform judgements about memorability as a general attribute as opposed to their own learning and memory (Tullis & Fraundorf, 2017). In addition, people may use different cues for metamemory judgements made during a learning task versus a judgement-only task. For instance, pre-study JOLs made prior to learning items based on information about cue levels only (e.g., "You are about to study an emotional item") show lower relative accuracy than standard immediate JOLs. This is because pre-study JOLs can only be based on beliefs about how cue values affect memorability (e.g., "It is an emotional item, so it is easy to remember"; Price & Harrison, 2017; Undorf & Bröder, 2020), but not on learning. Similarly, ease-of-learning judgements made prior to learning (e.g., "How easy or difficult will it be to learn this item?") show lower relative accuracy than immediate JOLs (Kelemen et al., 2000; Leonesio & Nelson, 1990; Pieger et al., 2016). Furthermore, JOLs that are elicited immediately after studying each item are less accurate than JOLs elicited with a delay (Dunlosky & Nelson, 1992, 1994; Nelson & Dunlosky, 1991). This is because the cues available after learning might be more diagnostic than those during learning where item information is still present in working memory. Taken together, JOLs might rely more on diagnostic cues than MJs because they are made for one's own memory during a learning task.

In addition, the accuracy between JOLs and MJs might differ because of the memory tasks used to measure picture memorability and the memory criterion value used for accuracy. Picture memorability measures may, for example, vary between memory tasks. JOL studies with pictures of scenes often use a learning phase followed by an old/new recognition memory test (Caplan et al., 2019; Hourihan, 2020; Hourihan & Bursey, 2017; Kao et al., 2005; Undorf & Bröder, 2021). In contrast, in their study on MJs, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)) employed a repeat detection task in which participants simultaneously encoded images and detected image repetitions. Moreover, regarding the memory criterion value, JOLs are related to participant's own individual memory performance (i.e., correlation of JOLs with item recognition memory by participant), whereas MJs are related to image memorability at the item level (i.e., correlation of MJs with item recognition memory aggregated across participants from other samples). Although it has been demonstrated that image memorability is highly consistent across participants (Bainbridge et al., 2013; Bylinskii et al., 2015; Isola, Parikh, et al., 2011; Isola, Xiao, et al., 2011, 2014), there might be individual differences contributing to judgement predictive accuracy. By aggregating recognition memory performance across participants, idiosyncratic information influencing memory and metamemory is not considered (Tullis & Fraundorf, 2017; Undorf et al., 2022). This might be another reason contributing to the lower accuracy of MJs.

It is important to mention that a recent study showed that judgements of perceived memorability and JOLs for pictures of real-world objects and faces were both predictive of actual stimulus memorability (Saito et al., 2023). Given differences in the stimuli materials, parallels between that research and the current study are difficult to draw. Scene pictures are complex and high-dimensional stimuli that cannot be easily recoded with a simple verbal label. In contrast, real-world objects are easier to encode and retrieve because they benefit from an imaginal/verbal dual-coding processing (Paivio, 1991). Also, face processing is highly specialised and may not be comparable to the processing of other visual stimuli (Bruce & Young, 1986; Schwaninger et al., 2004). Thus, MJs might have shown low accuracy in prior studies due to the complexity and high dimensionality of scene pictures.

The current study

The aim of this study was to directly compare the relative accuracy and cue basis of JOLs and MJs for pictures of scenes. To achieve this aim, participants made two types of metamemory judgements, JOLs and MJs, for different aspects of picture memorability (one's own future memory vs. memorability as a generic item attribute) and during a learning versus judgement-only task, respectively. At the same time, we ensured that the MJ and JOL procedures were as similar as possible in all other respects. Specifically, we used the same judgement scale for both judgements, manipulated identical cues, and investigated the relative accuracy of JOLs and MJs with respect to the same memory criterion, namely, actual population memorability of scenes. *Population scene memorability* was defined as the proportion of recognition hits minus the proportion of false alarms per scene in each experiment's recognition memory task. This measure corresponds to the proportion of corrected hit rates (also known as *Pr*, Snodgrass & Corwin, 1988) and prevents that false memories contribute to the actual memorability of scenes.[1] We also investigated the relative accuracy of JOLs with respect to the participants' own memory performance criterion.[2] If discrepant findings regarding the accuracy of JOLs and MJs reported in prior work are mainly due to differences in the judgements' cue basis, we expect to see clear differences in cue use and judgement accuracy across JOLs and MJs. In contrast, if discrepant findings are largely due to methodological differences across studies obtaining JOLs and MJs, we expect to see similar accuracy and cue basis.

In Experiments 1 and 2, we used a within-subjects design by presenting a JOL task and an MJ task to the same participants, with the order of tasks counterbalanced between participants. In the JOL task, participants studied and made JOLs for a set of pictures, completed a distraction task, and finally completed a recognition memory test. In the MJ task, participants judged the memorability of another set of pictures. In Experiment 1, we orthogonally manipulated aesthetics and interestingness in two clearly distinguishable levels to compare the cue basis of MJs and JOLs. In Experiments 2 and 3, we used pictures that represented the whole range of normed image memorability in a continuous way. To foreshadow the results, we found that MJs and JOLs had similar cue bases and were both predictive of scene memorability, but that the relative accuracy of MJs improved after a JOL task (Experiment 1). This effect was completely unexpected, so we replicated it in Experiment 2. To gain further insight in this unexpected and theoretically relevant result, we designed Experiment 3 to disentangle which component of the JOL task drives the improvement in MJ accuracy. For this, we used a four-group design in which participants completed either (1) the learning phase with JOLs and a memory test (i.e., the full JOL task as in the previous experiments), (2) the learning phase without JOLs plus a memory test, (3) the learning phase with JOLs but no memory test, or (4) no learning phase with JOLs and no memory test (i.e., no component of the JOL task) before completing the MJ task. We found that the learning phase by itself was sufficient for the improvement in MJ accuracy. In addition, we found that MJs were more sensitive to cue effects after a memory test than after making JOLs. This was in line with Pearson correlations showing higher MJ accuracy when participants previously took a test than when they made JOLs.

Experiment 1

In Experiment 1, we examined the relative accuracy and cue basis of JOLs and MJs for pictures of naturalistic scenes that varied in aesthetics and interestingness. Aesthetics and interestingness were the image attributes identified as negative predictors of image memorability, but positively affecting MJs in Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014)). The JOL task was similar to the one used by Undorf and Bröder (2021). Participants provided a JOL after studying each picture from a set of 120 pictures, and, following the learning phase, completed a recognition memory test with 240 pictures. In the MJ task, participants gave an MJ for each picture from another set of 120 pictures. They were explicitly instructed not to study the pictures, but only to judge their general memorability. This procedure was very similar to that used by Isola, Parikh, et al. (2011), Isola, Xiao, et al. (2011, 2014), with the exception that Isola et al. obtained binary ratings, whereas we used the same 11-point scale for MJs and JOLs. This was critical to prevent that potential accuracy differences between the two types of judgements could stem from using different judgement scales. To manipulate aesthetics and interestingness, we presented scenes from all possible combinations of high and low interestingness and aesthetics to participants. As Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) found that aesthetics and interestingness negatively affected memory performance, we expected that memory performance for pictures would be worse for scenes high

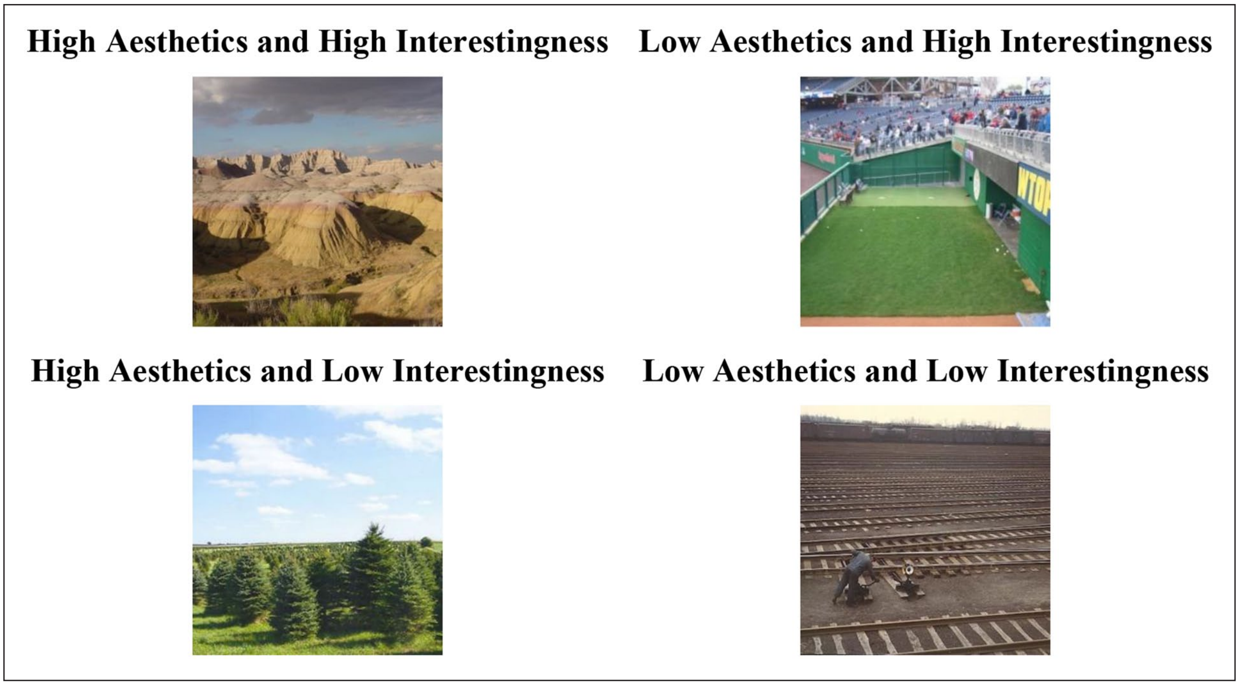

High Aesthetics and High Interestingness
Low Aesthetics and High Interestingness
High Aesthetics and Low Interestingness
Low Aesthetics and Low Interestingness

Figure 1. Example pictures for each combination of aesthetics and interestingness used in Experiment 1.

in aesthetics and interestingness. Furthermore, given that JOLs for pictures were usually moderately accurate and relied on valid cues (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021), we predicted accurate JOLs that would decrease with aesthetics and interestingness. It was an open question whether MJs would be accurate and relying on valid cues.

Method

Design and materials. The design was a 2 (aesthetics: low vs. high) × 2 (interestingness: low vs. high) × 2 (task order condition: JOLs first vs. MJs first) mixed design, with aesthetics and interestingness as within-participants factors and task order as a between-participants factor. Half of the participants were randomly allocated to the JOLs-first condition ($n=26$). The other half of participants were allocated to the MJs-first condition ($n=26$). Aesthetics and interestingness were manipulated by selecting different sets of normed scene pictures. Stimuli were 360 pictures from the SUN database (Xiao et al., 2010). Normed values for aesthetics and interestingness were taken from Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) who asked 30 participants "Is this an aesthetic image?" and "Is this an interesting image?" Yes/No. Ninety scenes each were low in aesthetics and interestingness, low in aesthetics and high in interestingness, high in aesthetics and low in interestingness, and high in aesthetics and interestingness (see Figure 1).[3] We divided the scenes into three parallel sets each with 120 scenes in total, 30 of them from each combination of aesthetics and interestingness. For each participant, scenes from one randomly selected set served as study items in the JOL task, scenes from another randomly selected set served as distractors in the test phase of the JOL task, and scenes from the third set were used in the MJ task.

Participants. We aimed at recruiting at least 50 participants from the Prolific online subject pool. This sample size provides a statistical power of $(1-\beta)=.94$ to detect medium-sized main and interaction effects ($f=.25$, equivalent to $\eta_p^2=.06$) with $\alpha=.05$ in a mixed ANOVA when assuming a correlation of .50 between repeated measures (G*Power 3; Faul et al., 2007). We recruited participants who were 18–61 years old, reported English as their first language, and had at least a high school diploma as highest degree. The experiment took approximately 40 min and participants were paid £5.

To ensure high data quality, our criteria for not accepting submissions of participants in Prolific were: (1) study timed out, based on a time limit set by Prolific based on the estimated completion time ($n=4$), (2) completing the study with a different device than a desktop computer ($n=0$), or (3) low effort throughout the experiment operationally defined as writing gibberish in the filler task ($n=0$) or corrected hit rates of or very close to zero ($n=1$).[4] We accepted submissions from 57 participants. Our criteria for excluding accepted submissions from analysis were: participants reported technical problems ($n=5$), admitted having used helping tools during the study ($n=0$), or admitted having completed the study with the help of someone else ($n=0$). The final sample included 52 participants (37 females, 14 males, and 1 other). Their mean age was 32.42 years ($SD = 11.1$), 3 participants were between 18 and 20 years in age, 29 participants were between 21 and 30 years in age, 7 participants were between 31 and 40 years in age, 7 participants were between 41 and 50 years in age, and 6 participants were between 51 and 61 years in age.

Procedure. The experiment consisted of a JOL task and an MJ task. Participants in the JOLs-first condition completed the JOL task first and then completed the MJ task. Task order was reversed for participants in the MJs-first condition. At the beginning of the experiment, we asked participants to comply with the following requirements: maximising the size of the web browser so that it covers the entire screen, completing the study in a single session, not leaving the study to engage in other tasks, completing the study in an environment that is free of noise or distraction, and not using any helping tools to complete the tasks.

In the JOL task, participants were instructed that their task was to remember 120 scene pictures for a later memory test in which studied photos would be intermixed with new ones and they would be asked to indicate whether each photo presented was studied or new. They were also instructed to predict the chances that they would personally recognise the photo on the test immediately after learning each photo. At learning, each scene picture was centred in the top half of the screen and displayed for 1 s, preceded by a 500-ms fixation cross that appeared in the same location. Immediately afterwards, participants indicated their chances of recognising the picture at test. To make their self-paced JOL, participants clicked on one of 11 buttons labelled 0, 10, . . ., 90, and 100. Following the learning phase, participants performed a semantic filler task for 3 min. On each filler trial, participants had 20 s to type in one word from each of three categories (i.e., animal, meal, and city) that started with a given letter. Finally, participants completed a self-paced recognition test with 240 scenes that included the 120 studied and 120 new scenes. At test, each scene picture was centred in the top half of the screen and participants indicated whether they had studied the picture before by clicking on buttons labelled "yes" and "no."

In the MJ task, participants were told that they would be presented with 120 scene pictures and their task is to judge how memorable each scene is. Participants were informed that they need not study the pictures themselves. At judging, each scene picture was centred in the top half of the screen and participants indicated how memorable the picture was for people who are asked to memorise the photo and later recognise it among new pictures. To make their self-paced MJ, participants clicked on one of 11 buttons labelled 0 (not memorable at all), 10, . . ., 90, and 100 (very memorable). For each participant, scene pictures were presented in a new random order in the learning phase and recognition memory test of the JOL task, and in the MJ task.

Data analysis

We report three different measures of judgement resolution. In all measures of judgement resolution, we used population scene memorability at the item level as memory criterion for MJ and JOL accuracy. Population scene memorability corresponds to the corrected hit rate for each scene, and it was calculated by subtracting the false alarm rate from the hit rate per scene. As the memory criterion for JOL accuracy at the individual level (i.e., participants' own memory performance), we used uncorrected hit rates because it is impossible to correct hit rates for both individual participants and individual items. Our main measure of judgement resolution is the within-subject Goodman–Kruskal gamma correlation between metamemory judgements and memory performance (Nelson, 1984). This is one of the most used measures of relative metamemory accuracy. Because population scene memorability was a continuous variable, we also report Pearson correlation coefficients. Furthermore, because gamma has been criticised due to inflated Type 1 errors (Murayama et al., 2014), discarded ties (Masson & Rotello, 2009; Spellman et al., 2014), and variation with liberal or conservative response criteria in recognition memory (Masson & Rotello, 2009), we additionally conducted a mixed-effects model analysis predicting population scene memorability from MJs and JOLs (Murayama et al., 2014).

Results

Resolution of JOLs and MJs. Table 1 and Figure 2 show mean gamma correlations between metamemory judgements and population scene memorability for each task in each task order condition. All correlations were significantly positive, $t >= 5.02$, $p < .001$, indicating that not only JOLs but also MJs captured differences in population scene memorability. A 2 (task: JOL vs. MJ; within-participants) × 2 (task order condition: JOLs-first vs. MJs-first; between-participants) mixed ANOVA revealed no main effects, task: $F(1, 50) = 2.24$, $p = .14$, $\eta_p^2 = .04$, task order condition: $F(1, 50) = 1.34$, $p = .25$, $\eta_p^2 = .03$, but a significant interaction, $F(1, 50) = 25.30$, $p < .001$, $\eta_p^2 = .34$. Follow-up t-tests indicated that gamma correlations for JOLs did not differ between conditions, $t(50) = 1.36$, $p = .18$, $d = 0.38$, whereas gamma correlations for MJs were higher in the JOLs-first condition than in the MJs-first condition, $t(50) = 3.09$, $p < .01$, $d = 0.88$, which indicates higher relative accuracy of MJs when made after the JOL task. Equivalent results were found with the mixed-effects model analysis (see the Supplementary Material 2). Similar results were found with Pearson correlations, except for a main effect of task indicating higher Pearson correlations for MJs than for JOLs (see the Supplementary Material 1).

Table 1 shows mean gamma correlations between JOLs and participant's own memory performance (individual

Table 1. Means (*SDs*) of the gamma correlation between population scene memorability (hit rate corrected per scene) or participant's own memory performance and JOLs or MJs in each task order condition of Experiments 1, 2, and 3.

Experiment and condition	Accuracy criterion		
	Population scene memorability		Own memory performance
	JOLs	MJs	JOLs
Experiment 1			
JOLs first	.21 (.14)	.35 (.17)	.36 (.15)
MJs first	.27 (.17)	.19 (.19)	.40 (.16)
Experiment 2			
JOLs first	.23 (.14)	.31 (.15)	.33 (.25)
MJs first	.25 (.18)	.19 (.17)	.45 (.18)
Experiment 3			
MJ-task-only	-	.20 (.22)	-
Full-JOL-task	.26 (.14)	.32 (.15)	.38 (.22)
Study-and-JOL-task	.28 (.18)	.26 (.20)	-
Study-and-test-task	-	.31 (.17)	-

Note. JOLs = judgements of learning, MJs = memorability judgements, Population scene memorability = hit rate minus false alarm rate per scene across participants in each experiment.

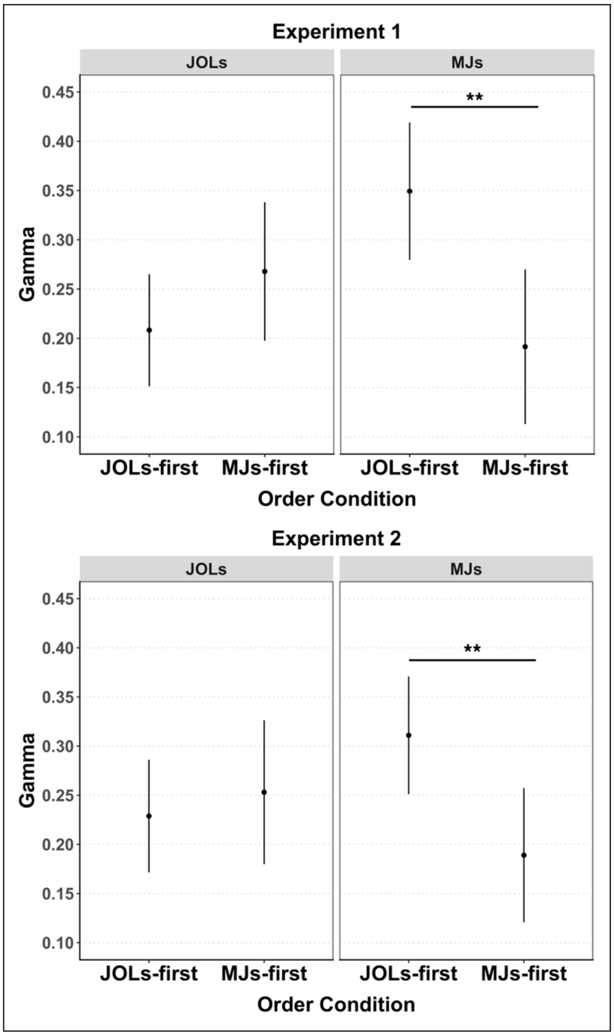


Figure 2. Gamma correlations between population scene memorability (hit rate corrected per scene) and judgements of learning (JOLs) or memorability judgements (MJs) in each task order condition of Experiments 1 and 2.
Note. Error bars represent one standard error of the mean.
Experiment 1: p = .003
Experiment 2: p = .008

memory performance) in each task order condition. Both gamma correlations were significantly positive, $t \geq 12.04$, $p < .001$, and they did not differ between order conditions, $t < 1$.

Cue effects on JOLs and individual memory performance. Figure 3 presents JOLs and corrected hit rates from the JOL task by aesthetics and interestingness in the JOLs-first and MJs-first condition. A 2 (aesthetics: low vs. high) × 2 (interestingness: low vs. high) × 2 (task order condition: JOLs first vs. MJs first) mixed ANOVA on JOLs revealed no main effect of aesthetics, $F(1, 50) = 0.77$, $p = .\ 39$, $\eta_p^2 = .02$, a main effect of interestingness, $F(1, 50) = 90.02$, $p < .001$, $\eta_p^2 = .64$, indicating higher JOLs for scenes high in interestingness than low in interestingness, no main effect of task order condition, $F(1, 50) = 1.64$, $p = .21$, $\eta_p^2 = .03$, and a significant interaction between order condition and interestingness, $F(1, 50) = 7.75$, $p < .01$, $\eta_p^2 = .13$. Follow-up *t*-tests indicated that interestingness affected JOLs in both conditions, but more so in the MJs-first condition, JOLs-first condition: $t(25) = 5.34$, $p < .001$, $d = 1.07$; MJs-first condition: $t(25) = 7.87$, $p < .001$, $d = 1.58$. No other interactions were significant, $F < 1$, $p \geq .37$.

A similar ANOVA on corrected hit rates (*Pr*) revealed better recognition memory performance for scenes low in aesthetics than high in aesthetics, $F(1, 50) = 54.73$, $p < .001$, $\eta_p^2 = .52$, and for scenes high in interestingness than low in interestingness, $F(1, 50) = 4.15$, $p = .047$, $\eta_p^2 = .08$, no other effects were significant, $F \leq 1.64$, $p \geq .21$.[5] We thus replicated Isola, Parikh, et al.'s (2011) and Isola, Xiao, et al.'s (2011, 2014) findings of better memory performance for scenes low in aesthetics, but did not replicate better memory performance for scenes low in interestingness. Instead, we found that memory performance was better for scenes high in interestingness. We will return to this point in the "Discussion" section.

Cue effects on MJs. Figure 3 presents MJs from the MJ task by aesthetics and interestingness in the JOLs-first and MJs-first condition. A 2 (aesthetics: low vs. high) × 2 (interestingness: low vs. high) × 2 (task order condition: JOLs first vs. MJs first) mixed ANOVA on MJs revealed no main effect of aesthetics, $F(1, 50) = 3.65$, $p = .06$, $\eta_p^2 = .07$, a main effect of interestingness, $F(1, 50) = 224.02$, $p < .001$, $\eta_p^2 = .82$, indicating higher MJs for scenes high

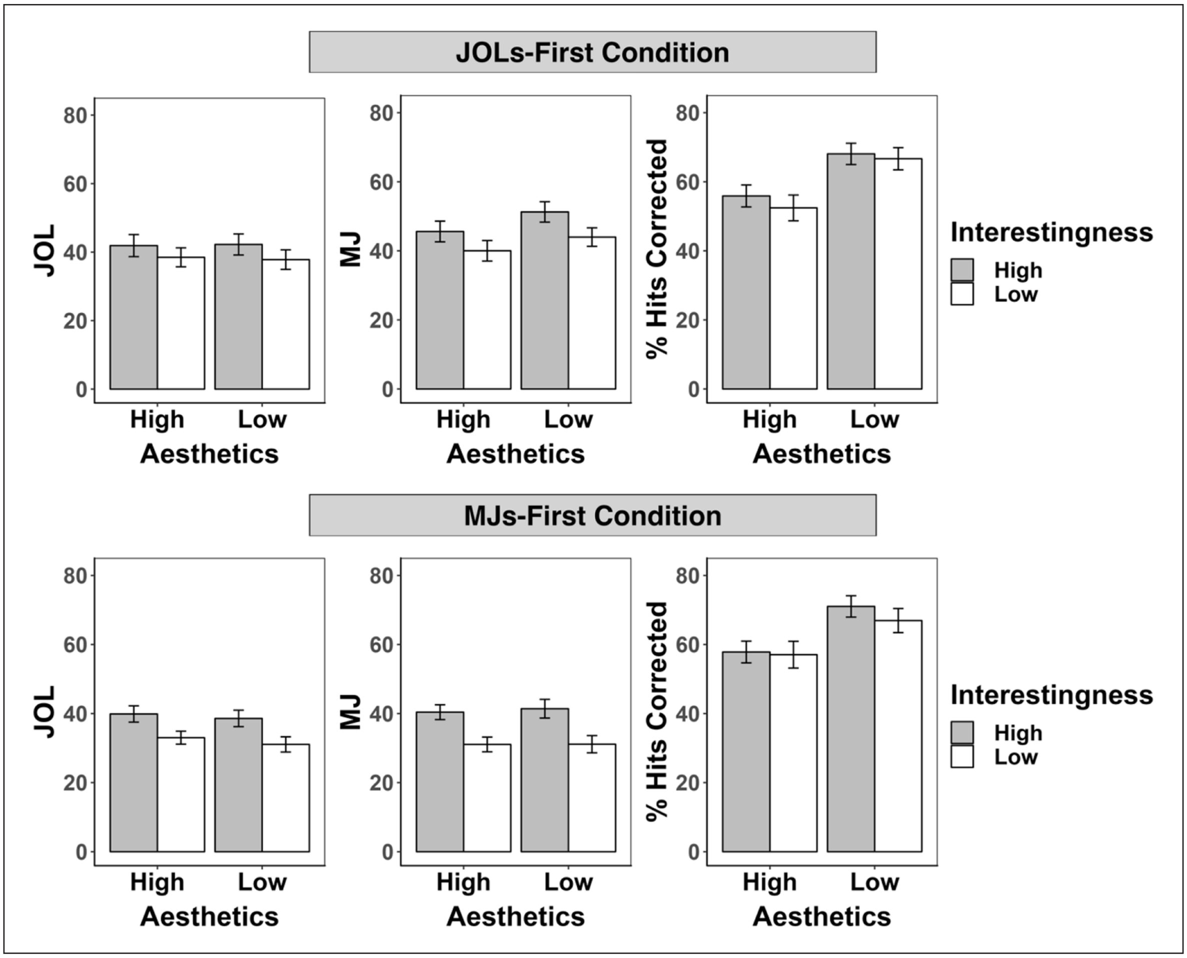


Figure 3. Mean judgements of learning (JOL), memorability judgements (MJ) and corrected hit rates (% hits corrected) by aesthetics and interestingness in the JOLs-first (top panel) and MJs-first (bottom panel) conditions of Experiment 1.
Note. Error bars represent one standard error of the mean.

in interestingness than low in interestingness, a main effect of task order condition, $F(1, 50) = 7.42$, $p < .01$, $\eta_p^2 = .13$, indicating higher MJs in JOLs-first condition than in the MJs-first condition, and a significant interaction between interestingness and task order condition, $F(1, 50) = 9.71$, $p < .01$, $\eta_p^2 = .16$. Follow-up *t*-tests indicated that interestingness affected MJs in both conditions, but more so in the MJs-first condition, JOLs-first condition: $t(25) = 9.33$, $p < .001$, $d = 1.87$; MJs-first condition: $t(25) = 11.71$, $p < .001$, $d = 2.34$. All other interactions, $F \leq 2.33$, $p \geq .13$.

Discussion

Recognition memory was affected by the two image characteristics aesthetics and interestingness. As in Isola et al., we found better memory performance for scenes low rather than high in aesthetics. Contrary to Isola et al., we found better memory performance for scenes high rather than low in interestingness. Potential explanations for this difference in results include that we used a different recognition memory paradigm (i.e., learning phase followed by an old/new memory test), and that aesthetics and interestingness were strongly correlated in Isola et al.'s study ($\rho = .85$) but manipulated orthogonally here. Furthermore, the finding that JOLs and MJs were both unaffected by aesthetics, but higher for pictures high rather than low in interestingness suggests a similar cue basis of JOLs and MJs. The finding that aesthetics did not affect either metamemory judgement fits with prior findings indicating that people sometimes fail to factor valid cues in their JOLs for scene pictures (e.g., peacefulness, Undorf & Bröder, 2021).

Despite people's failure to recognise the predictive validity of aesthetics in their JOLs and MJs, reliable resolution showed that both metamemory judgements captured differences in the relative population memorability of scenes. Thus, by directly comparing JOLs and MJs in a within-subjects design using the same memory criterion, our results showed that both types of judgement had moderate resolution.

A new and unexpected finding was that the accuracy of MJs improved substantially after completing a JOL task, whereas completing an MJ task did not affect JOL accuracy.[6] The order of the tasks also did not affect JOL accuracy when using participant's own memory performance as memory criterion. This finding shows that the accuracy of JOLs and MJs differs in whether it is affected by the order of the tasks. Previous metamemory studies using multiple study-test cycles for the same materials have reported changes in the resolution and absolute accuracy of JOLs. From the second study-test cycle onward, reliance on past memory performance increases JOL resolution and, at the same time, under-confidence lowers absolute accuracy of JOLs (Finn & Metcalfe, 2008; King et al., 1980; Koriat et al., 2006). Importantly, the current finding that MJ resolution improves after a JOL task is novel in that it demonstrates increased accuracy of judging the general memorability of *new* pictures after actively engaging in a learning phase with JOLs and a memory test.

In conclusion, the results of Experiment 1 indicate that (1) JOLs and MJs for scenes were predictive of population scene memorability and that (2) both types of metamemory judgements had a similar cue basis (i.e., based on interestingness, but not on aesthetics). A surprising finding was that (3) having completed a learning task with JOLs and a recognition memory test improved the accuracy of MJs, whereas making MJs did not improve the accuracy of JOLs. A potential mechanism for this improvement in relative accuracy is that participants gained experience by intentionally learning pictures and reflecting about their own memory performance. If MJs are made without this experience, participants probably lack knowledge about how to assess the abstract image feature memorability. This interpretation suggests that experiences with one's own memory precede the understanding of memory in general, and it would help to explain why MJ and JOL accuracy is sometimes comparable and sometimes not. To rule out that improved MJ accuracy after the JOL task was an accidental result in Experiment 1, Experiment 2 aimed to conceptually replicate this finding.

Experiment 2

Experiment 2 aimed to replicate findings of Experiment 1 and, specifically, the unexpected finding that having completed a JOL task with learning pictures, making JOLs, and taking a recognition memory test improved the relative accuracy of MJs. Because JOLs and MJs were based on similar cues in Experiment 1, we did not manipulate individual cues in Experiment 2, but instead used scenes that varied widely in scene memorability. Based on the findings obtained in Experiment 1, we expected that both JOLs and MJs would be similarly impacted by scene memorability. As in Experiment 1, all participants completed a JOL task and an MJ task with the task order manipulated between participants. We expected that, as in Experiment 1, JOLs and MJs would be predictive of population scene memorability. At the same time, given the experience and knowledge people gained in the JOL task, we hypothesised that the relative accuracy of MJs would be higher in the JOL-first condition than in the MJ-first condition.

Method

Design and materials. The design was a 10 (scene memorability: 10 levels from low to high) × 2 (task order condition: JOLs-first vs. MJs-first) mixed design, with scene memorability as a within-participants factor and task order as a between-participants factor. Scene memorability was manipulated by selecting different sets of scene pictures that varied in corrected hit rates (i.e., hit rate minus false alarm rate per scene) reported in Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014). We used the deciles of the frequency distribution of scene memorability as cutoff values and selected 36 scenes from each of the 10 levels, resulting in a total of 360 pictures (see Figure 4).[7] We divided the scenes from each level of memorability into three parallel sets with 12 scenes. Recombining these, we thus created 3 parallel sets of 120 pictures each (12 of each level). Sets were also similar in aesthetics and interestingness. For each participant, scenes from one randomly selected set served as study items in the JOL task, scenes from another randomly selected set served as distractors in the test phase of the JOL task, and scenes from the third set were used in the MJ task.

Participants. We aimed at recruiting 50 participants from the Prolific online subject pool who were 18 to 61 years old, reported English as their first language, and had at least a high-school diploma as highest degree. Power analysis was identical to that of Experiment 1. The experiment took approximately 40 min and participants were paid £5. Based on the same criteria as in Experiment 1, we did not accept submissions in Prolific when the study timed out ($n=2$), was completed on a different device than a desktop computer ($n=1$), or there was low effort throughout the experiment ($n=0$).

We accepted submissions from 55 participants. Based also on the same criteria of Experiment 1, we excluded accepted submissions from analysis when participants reported technical problems ($n=5$), admitted having used helping tools during the study ($n=0$), or admitted having completed the study with the help of someone else ($n=0$). The final sample included 50 participants (30 females, 20 males). The mean age of participants was 34.64 ($SD=9.94$), 2 participants were between 18 and 20 years in age, 21 participants were between 21 and 30 years in age, 13 participants were between 31 and 40 years in age, 8 participants were between 41 and 50 years in age, and 6 participants were between 51 and 61 years in age.

Procedure. The procedure was identical to Experiment 1.

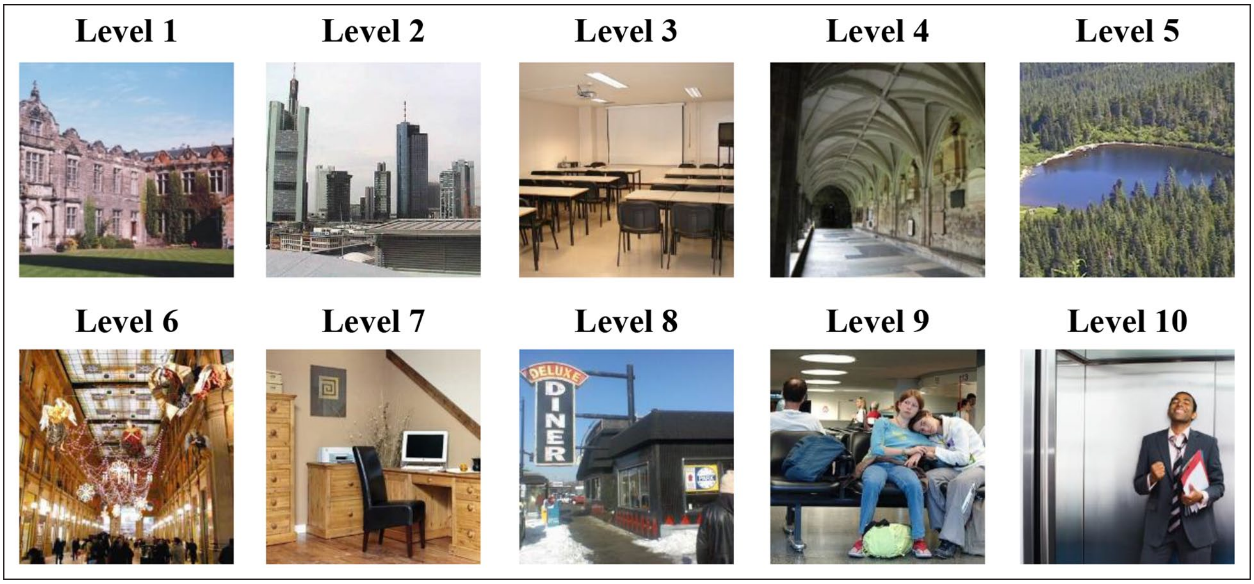


Figure 4. Example pictures of each of the 10 levels of scene memorability (from 1 = lowest to 10 = highest) used in Experiments 2 and 3.

Results

Resolution of JOLs and MJs. Table 1 and Figure 2 show mean gamma correlations between metamemory judgements and population scene memorability for each task in each task order condition. All correlations were significantly positive, $t > 5.72$, $p < .001$. A 2 (task: JOL vs. MJ; within-participants) × 2 (task order condition: JOLs-first vs. MJs-first; between participants) mixed ANOVA revealed no main effects: task, $F < 1$, task order condition, F (1, 48) = 1.52, p = .22, η_p^2 = .02, but a significant interaction, F(1, 48) = 12.99, $p < .001$, η_p^2 = .21. Planned comparisons indicated that gamma correlations for JOLs did not differ between conditions, $t < 1$, p = .59, whereas gamma correlations for MJs were higher in the JOLs-first condition than in the MJs-first condition, t(48) = 2.77, $p < .01$, d = 0.80. As in Experiment 1, this again shows higher relative accuracy of MJs after learning items, making JOLs, and completing a recognition memory test. Equivalent results were found with Pearson correlations and a mixed-effects model analysis (see the Supplementary Materials 1 and 2).

Table 1 shows mean gamma correlations between JOLs and participant's own memory performance in each task order condition. Both correlations were significantly positive, $t \geq 6.65$, $p < .001$, and did not differ between conditions, t(48) = 1.88, p = .07, d = 0.57.

Cue effects. Figure 5 presents JOLs, corrected hit rates, and MJs. We used a mixed-effects model (Bates et al., 2015) to evaluate whether JOLs and MJs increased monotonically with scene memorability. This approach allowed us to directly test for a linear increase in metamemory judgements with scene memorability as a fixed-effects predictor with 10 levels. To evaluate whether the scene memorability slope differed between order conditions, we included order condition and its interaction with scene memorability as additional fixed-effects predictors in the model. We specified random intercepts for participants and uncorrelated random slopes for scene memorability. Scene memorability was mean-centred, and task order condition was effect coded (−1 = MJs-first, 1 = JOLs-first). We used a logistic regression model to evaluate a linear increase in hit rates with scene memorability.

Cue effects on JOLs and individual memory performance. Regressing JOLs on scene memorability, task order condition, and their interaction revealed a significantly positive unstandardized coefficient for scene memorability, b = 2.01, (SE = 0.22), t = 9.18, $p < .001$, indicating that JOLs increased with scene memorability. No other effects were significant, order condition: b = 4.02, (SE = 2.01), t = 2.00, p = .05, interaction: $t < 1$. A logistic regression model revealed that hit rates increased with scene memorability, b = 0.20, (SE = 0.01), z = 17.72, $p < .001$. No other effects were significant, $z \leq 1.41$, $p \geq .16$.

Cue effects on MJs. Regressing MJs on scene memorability, task order condition, and their interaction revealed significantly positive unstandardized coefficients for scene memorability, b = 1.91, (SE = 0.21), t = 8.90, $p < .001$, indicating that MJs increased with scene memorability. The model also revealed significantly positive unstandardized coefficients for order condition, b = 7.38, (SE = 1.79), t = 4.12, $p < .001$, indicating that MJs were higher in the JOLs-first condition, and for the interaction between scene

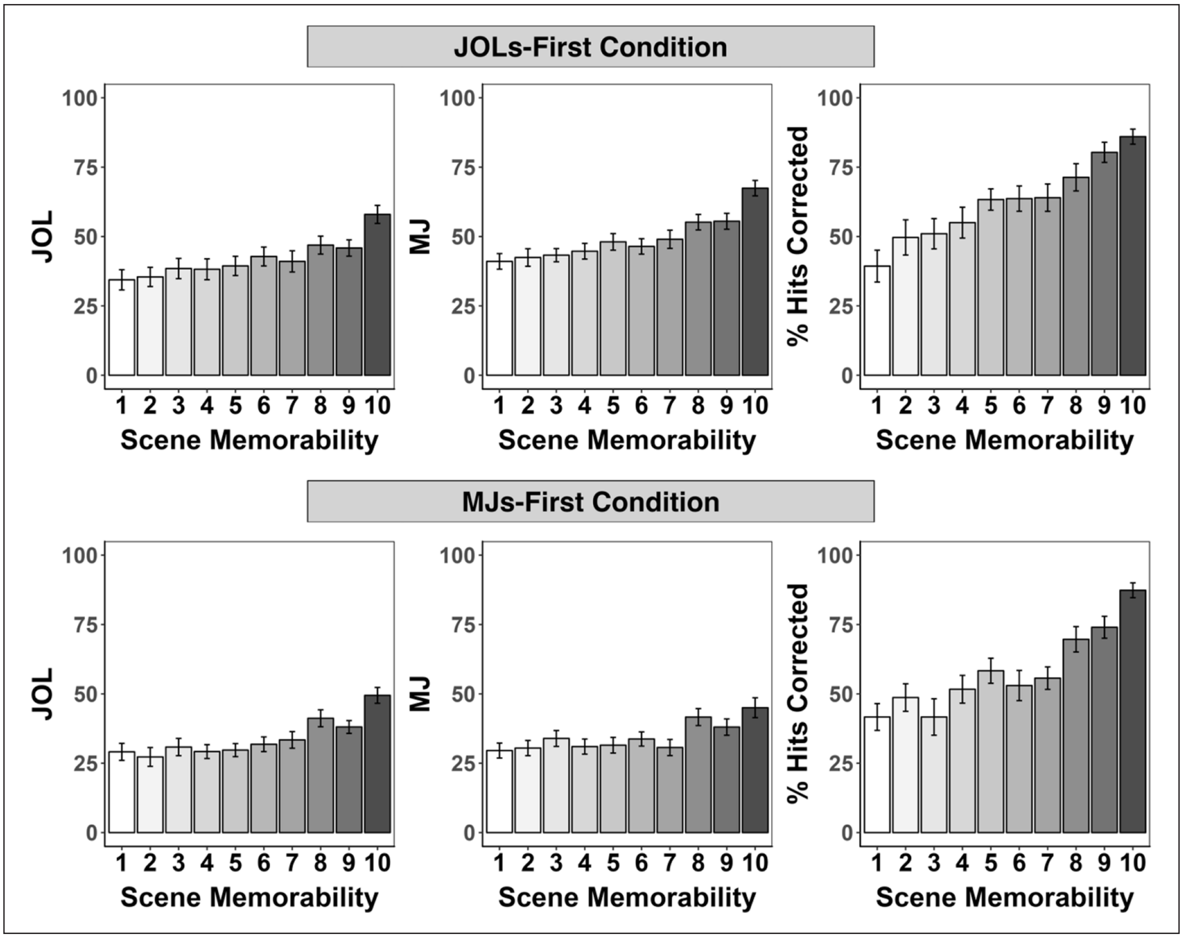


Figure 5. Mean judgements of learning (JOL), memorability judgements (MJ), and corrected hit rates (% hits corrected) by scene memorability in the JOLs-first (top panel) and MJs-first (bottom panel) conditions of Experiment 2.
Note. Error bars represent one standard error of the mean.

memorability and order condition, $b=0.51$, $(SE=0.21)$, $t=2.38$, $p=.022$, indicating differences in the effects of scene memorability on MJs across order conditions. Separate follow-up regression models for each order condition revealed that MJs increased with scene memorability in both conditions, but more so in the JOLs-first condition, JOLs-first condition: $b=2.42$, $(SE=0.28)$, $t=8.52$, $p<.001$; MJs-first condition: $b=1.40$, $(SE=0.32)$, $t=4.36$, $p<.001$.

Discussion

As in Experiment 1, both JOLs and MJs were predictive of differences in population scene memorability. We again found that the relative accuracy of MJs improved after a JOL task, whereas prior experiences with making MJs did not improve JOL accuracy. Thus, when using the same memory criterion for accuracy at the item level, differences in accuracy between JOLs and MJs tied to the order of tasks still emerged. Importantly, this shows that the beneficial impact of a preceding JOL task on MJ accuracy is a robust effect that merits further scrutiny. The finding that task order did not affect JOL resolution was independent of the memory criterion used for accuracy (i.e., population scene memorability, or participant's own memory performance).

As expected, both JOLs and MJs monotonically increased with increasing scene memorability. This again indicated that the cue basis of the two metamemory judgements is similar and suggested that several cues diagnostic of memorability underlie each type of metamemory judgement (for evidence that multiple cues are integrated in JOLs for scene pictures and for verbal materials, see, for example, Undorf & Bröder, 2021; Undorf et al., 2018). Importantly, MJs increased more strongly with scene memorability in the JOLs-first than in the MJs-first condition, indicating that MJs become more sensitive to scene memorability effects after a JOL task. In contrast, scene memorability effects on JOLs were unaffected by the task order condition. This finding supports our hypothesis that participants learn about the general memorability of scenes by completing a JOL task and make MJs for new set of pictures on an updated basis.

In summary, results from Experiment 2 again show that experience with a JOL task provides a viable basis for assessing the general memorability of scenes. Therefore, this novel result from Experiment 1 was proven to be replicable and one may ask for an explanation. One step in this direction is to investigate which component of the JOL task (i.e., learning phase, making JOLs, recognition memory test) drives the improvement in MJ accuracy. Regarding the potential contribution of a learning phase to the relative accuracy of metamemory, literature is scarce. Two studies investigating the effects of prior learning versus prior testing on metamemory accuracy found that test experience was more effective than learning experience (Jang et al., 2012; Koriat & Bjork, 2006a). However, in the current study, it might be possible that a learning phase provides participants with the experience required to make accurate MJs. Regarding JOL experience, making JOLs for oneself might increase MJ accuracy because monitoring one's own learning can increase sensitivity towards diagnostic cues. For instance, previous studies have found that processing fluency as indicated by short self-paced study times is used as a cue for other's memory predictions only after learners had made JOLs for their own memory (Koriat & Ackerman, 2010; Undorf & Erdfelder, 2011). Therefore, from a cue-weighting perspective (Undorf et al., 2018), it may be that completing a learning phase with JOLs fosters the use of valid cues for MJs. These valid cues might include mnemonic cues such as the ease of encoding (Begg et al., 1989; Chandler, 1994; Hertzog et al., 2003) or perceiving pictures (Besken, 2016; Fei-Fei et al., 2007; Undorf et al., 2017) and intrinsic cues such as emotionality and concreteness (Undorf & Bröder, 2020). Alternatively, or additionally, test experience might improve MJ accuracy for a new set of scenes by providing participants with feedback regarding the memorability of scene pictures. Specifically, monitoring one's recognition memory performance for scene pictures during the test may provide hints of the features or feature combinations that make a picture memorable (Mitton & Fiacconi, 2020). For example, a participant might realise during the test that she recognises interesting pictures or pictures with people better than others. Thus, based on prior work, it is plausible that learning scene pictures, making JOLs, and taking a recognition memory test underlie the improvement in MJ accuracy observed in the previous experiments separately or in combination.

Experiment 3

Experiment 3 was preregistered (https://osf.io/3fujm) and aimed at disentangling which component of the JOL task drives the improvement in MJ accuracy. For this, in the first part of the experiment, different groups of participants completed either the full JOL task (*full-JOL-task* condition), a learning phase with JOLs, but without a test (*study-and-JOL-task* condition), a learning phase without JOLs, but with a recognition memory test (*study-and-test-task* condition), or no component of the JOL task (*MJ-task-only* condition). In the second part of the experiment, all participants completed an MJ task. In this four-group design, making JOLs (yes, no) and taking a test (yes, no) are fully crossed with the MJ-task-only condition being the control. However, the MJ-task-only condition (i.e., no JOLs, no test) additionally differs from the other three conditions by not including a learning phase. Because one cannot make JOLs or take a test without having learned the pictures, completing the learning phase (yes, no) cannot be fully crossed with the other variables (i.e., making JOLs, taking a test). Nevertheless, the imbalanced design allows for all crucial tests: If all experimental conditions show a similar improvement in MJ accuracy compared with the control condition, then completing a learning phase is the critical factor driving MJ accuracy. If MJ accuracy is higher in the full JOL-task condition than in the conditions in which participants make JOLs but do not take a test or take a test but do not make JOLs, then making JOLs and taking a test have additive effects on MJ accuracy. Finally, differences in MJ accuracy across the conditions in which participants make JOLs but do not take a test or take a test but do not make JOLs will reveal the relative importance of making JOLs or taking a test for improved MJ accuracy.

Method

Design and materials. The design was a 10 (scene memorability: 10 levels from low to high) × 4 (condition: full-JOL-task, study-and-test-task, study-and-JOL-task, MJ-task-only) mixed design, with scene memorability as a within-participants factor and condition as a between-participants factor. We used the same sets of pictures as in Experiment 2.

Participants. We aimed at recruiting $N=212$ participants from the Prolific online subject pool ($n=53$ in each condition) who were 18 to 61 years old, reported English as their first language, and had at least a high-school diploma as highest degree. This sample size provides a statistical power of $(1-\beta)=.95$ to detect medium-sized effects ($f=.25$, equivalent to $\eta_p^2=.06$) with $\alpha=.05$ in a fixed-effects ANOVA employed to test power for contrasts with $df=1$ and $df=4$ in the numerator and the denominator, respectively (G*Power 3; Faul et al., 2007; Perugini et al., 2018). The experiment took approximately 40 min and participants were paid £5. Participants were randomly allocated to one of the four conditions. Based on the same criteria as in Experiments 1 and 2, we did not accept submissions in Prolific when the study timed out ($n=2$), was completed on a different device than a desktop computer ($n=1$), or there was low effort throughout the experiment ($n=0$). We accepted submissions from 212 participants. Based also on the same criteria as Experiments 1 and 2, we excluded data from analysis when participants reported technical problems ($n=0$), admitted having used helping tools during the study ($n=3$), admitted completing the study with the help of someone else ($n=4$), or admitted having just clicked through the study without taking part seriously ($n=0$). The final sample included 205 participants ($n=51$ in the full-JOL-task, study-and-JOL-task, and MJ-task-only condition, $n=52$ in the study-and-test-task condition). They were 113 females, 91 males, and 1 other. The mean age of participants was 37.29 years ($SD=10.54$), 6 participants were between 18 and 20 years in age, 59 participants were between 21 and 30 years in age, 60 participants were between 31 and 40 years in age, 52 participants were between 41 and 50 years in age, and 28 participants were between 51 and 61 years in age.

Procedure. The experiment consisted of two parts. In the first part of the experiment, participants in the full-JOL-task condition completed the same JOL task as in Experiments 1 and 2 (i.e., learning phase with JOLs, semantic filler task, and recognition memory test). Participants in the study-and-JOL-task condition completed a learning phase with JOLs, and a semantic filler task, but no recognition memory test. They received the same initial instructions as participants in the full-JOL-task condition but learned at the end of the experiment that we wanted to examine the accuracy of memorability estimates in participants who had not taken a memory test, which is why they had skipped the memory test. Participants in the study-and-test-task condition completed a learning phase without JOLs, a semantic filler task, and a recognition memory test. Participants in the MJ-task-only condition completed the semantic filler task only. In the second part of the experiment, participants from all conditions completed the same MJ task as in Experiments 1 and 2.

Results

Resolution of JOLs and MJs. As in Experiments 1 and 2 and as preregistered, we examined the extent to which MJs predicted differences in the actual population memorability of scenes in this experiment using within-subject Goodman–Kruskal gamma correlations, Pearson correlations (see the Supplementary Material 1), and a mixed-effects model analysis (see the Supplementary Material 2). To assess differences in MJ accuracy between conditions, we transformed our hypotheses into a set of orthogonal contrasts using Helmert coding. This approach has two main advantages: (1) testing specific group differences that are independent and more informative than an omnibus F-test, and (2) greater statistical power than follow-up *t*-tests (Rosenthal et al., 2000). The first contrast tested the difference between the control condition (MJs-only; −3/4) and all three experimental conditions (full-JOL-task, study-and-JOL-task, study-and-test-task; coded all as +1/4). The second contrast tested the difference between the full-JOL-task condition (+2/3) and the other two experimental conditions (study-and-JOL-task group, the study-and-test-task group; coded both as −1/3). The third contrast tested the difference between the study-and-JOL-task condition (−1/2) and the study-and-test-task condition (+1/2).

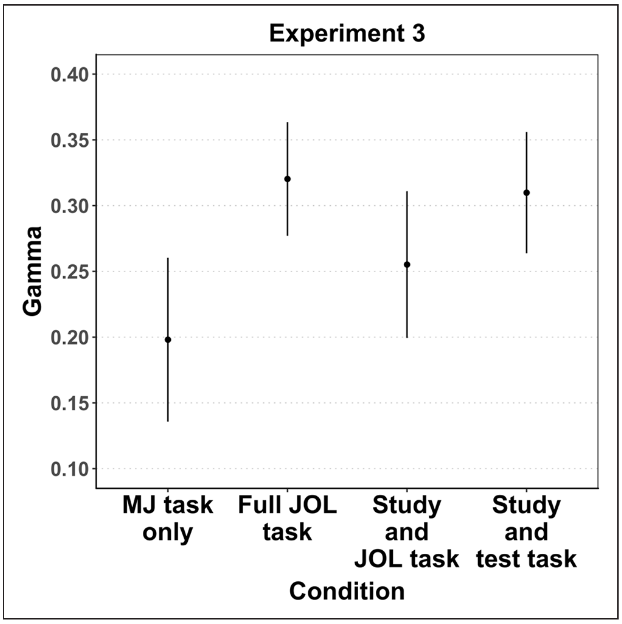


Figure 6. Gamma correlations between population scene memorability (hit rate corrected per scene) and memorability judgements (MJs) in each condition of Experiment 3.
Note. Error bars represent one standard error of the mean.

Table 1 and Figure 6 show mean gamma correlations between MJs and population scene memorability in each condition of Experiment 3. All correlations were significantly positive, $t \geq 6.39$, $p < .001$, indicating that MJs in all conditions captured differences in population scene memorability. Planned contrasts revealed that the learning phase present in all experimental conditions improved MJ accuracy compared with only making MJs, $t(201)=3.22$, $p < .01$. They also revealed that MJ accuracy did not differ between the full-JOL-task condition and the conditions with one component of the JOL task only (i.e., study-and-JOL-task, study-and-test-task), $t(201)=1.18$, $p=.24$, showing that there were no additive effects of making JOLs and taking a test. Finally, MJ accuracy did not differ between the study-and-JOL-task condition and the study-and-test-task condition, $t(201)=1.49$, $p=.14$, suggesting that making JOLs is not more beneficial than taking a test,

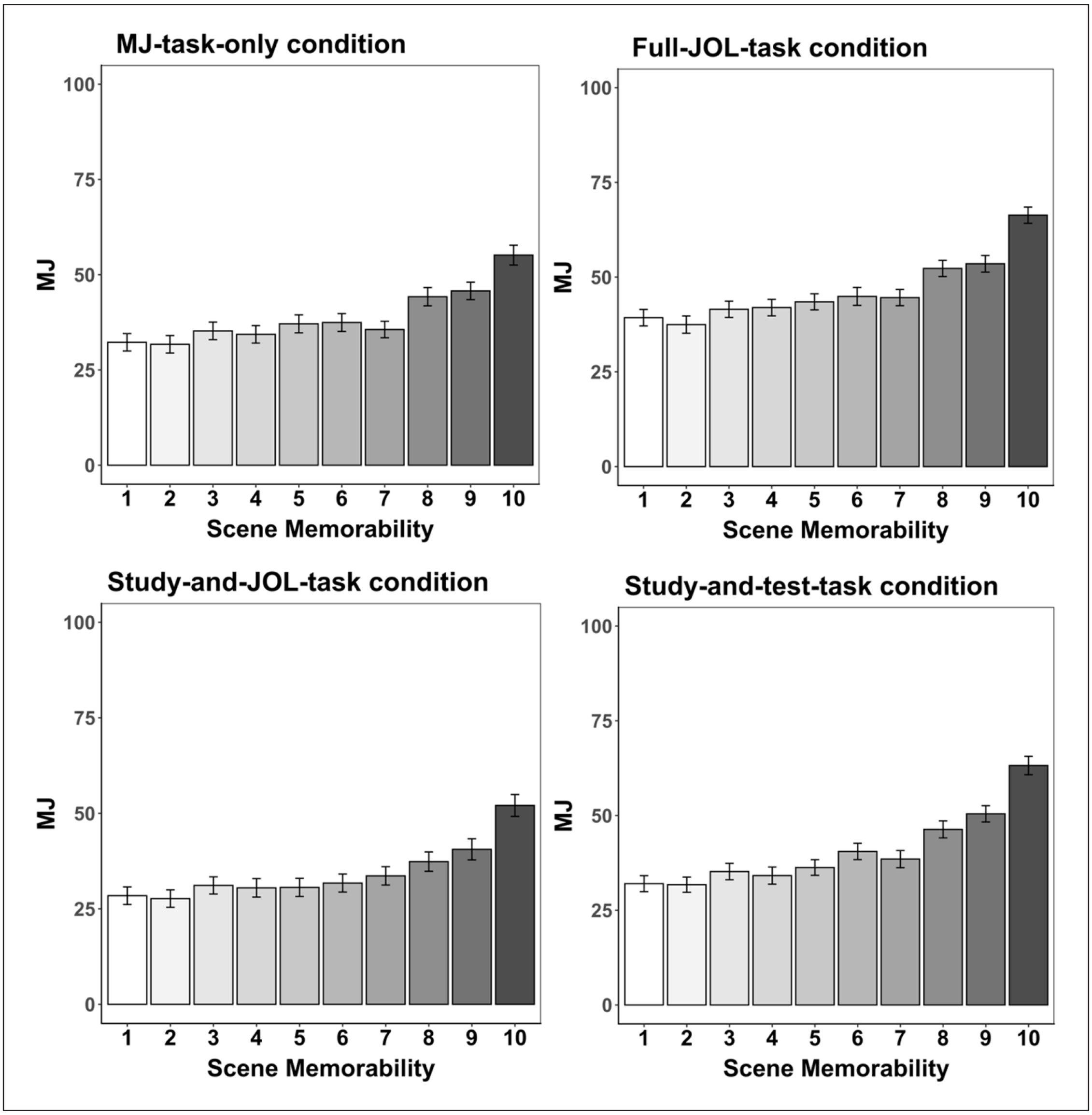


Figure 7. Mean memorability judgements (MJ) by scene memorability in each condition of Experiment 3.
Note. Error bars represent one standard error of the mean.

or vice versa. The latter result, however, was not supported by Pearson correlations which instead suggested that taking a test improved MJ accuracy more than making JOLs (see the Supplementary Material 1).

Cue effects on MJs. Figure 7 presents MJs in each condition of Experiment 3. As in Experiment 2 and as preregistered, we used a mixed-effects model to examine whether MJs increased with scene memorability. We included condition and its interaction with scene memorability as fixed-effects predictors in the model to evaluate whether the scene memorability slope differs between conditions. We specified random intercepts for participants and uncorrelated random slopes for scene memorability. Scene memorability was grand mean-centred, and condition was coded with the same Helmert contrasts as in the resolution analysis.

A significantly positive unstandardized coefficient for scene memorability, $b=2.43$, $(SE=0.19)$, $t=22.43$, $p<.001$, indicated that MJs again increased with scene memorability. Significantly positive unstandardized coefficients for the second and third contrasts coding condition; $b=8.93$, $(SE=2.56)$, $t=3.48$, $p<.001$, $b=6.47$, $(SE=2.95)$, $t=2.19$, $p<.05$, indicated higher MJs in the full-JOL-task condition than in the study-and-JOL-task and the study-and-test-task conditions, and higher MJs in the study-and-test-task condition than in the study-and-JOL-task condition. More importantly, a significant interaction between scene memorability and the third contrast coding condition revealed differences in scene memorability effects on MJs between the study-and-JOL-task and the study-and-test-task conditions, $b=0.85$, $(SE=0.31)$, $t=2.78$, $p<.01$. Separate follow-up regression models for each condition revealed that MJs increased with scene memorability in both conditions, but more so in the study-and-test-task condition, study-and-JOL-task condition: $b=2.09$, $(SE=0.19)$, $t=11.05$, $p<.001$; study-and-test-task condition: $b=2.94$, $(SE=0.26)$, $t=11.51$, $p<.001$. All other effects were nonsignificant, $t\leq 1.52$.

Cue effects on JOLs and individual memory performance. We used a similar mixed-effects model to evaluate whether JOLs in the full-JOL-task and the study-and-JOL-task conditions increase with scene memorability. Condition was effect coded (−1 = study-and-JOL-task condition, 1 = full-JOL-task condition). This model revealed a significantly positive unstandardized coefficient for scene memorability, $b=2.17$, $(SE=0.13)$, $t=16.79$, $p<.001$, indicating that JOLs again increased with scene memorability. All other effects were nonsignificant, $t\leq 0.75$.

A logistic regression model was used to evaluate whether individual recognition memory performance in the full-JOL-task and the study-and-test-task conditions increases with scene memorability. Condition was effect coded (−1 = study-and-test-task condition, 1 = full-JOL-task condition). This model revealed that hit rates increased with scene memorability, $b=0.16$, $(SE=0.01)$, $z=20.40$, $p<.001$, that hit rates were higher in the full-JOL-task than in the study-and-test-task condition, $b=0.53$, $(SE=0.09)$, $z=5.88$, $p<.001$, and that scene memorability effects on hit rates differed between the full-JOL-task and the study-and-test-task conditions, $b=0.05$, $(SE=0.01)$, $z=5.85$, $p<.001$. Separate follow-up regression models for the latter two conditions revealed that hit rates increased with scene memorability in both conditions, but more so in the full-JOL-task condition, study-and-test-task condition: $b=0.11$, $(SE=0.01)$, $z=11.53$, $p<.001$; full-JOL-task condition: $b=0.20$, $(SE=0.01)$, $z=16.92$, $p<.001$.

Discussion

The aim of Experiment 3 was to disentangle which component of the JOL task drives the improvement in MJ accuracy observed in the previous experiments. All measures of resolution showed that MJ accuracy improved in all experimental conditions (full-JOL-task, study-and-JOL-task, study-and-test-task) in comparison to the control condition (MJs-only). As the learning phase is the common factor in all experimental conditions, our result suggests that a learning phase by itself provides experiences that are beneficial for subsequently assessing the memorability of pictures. Moreover, given that MJ accuracy was not better in the full-JOL-task condition than in the other two experimental conditions (study-and-JOLs-task condition, study-and-test-task condition), we did not find evidence for additive effects of making JOLs and taking a test on MJ accuracy. Regarding the individual effects of making JOLs and taking a test on MJ accuracy, gamma correlations suggested that neither making JOLs nor taking a test improves MJ accuracy, as did the mixed-effects model analysis. In contrast, the analysis of Pearson correlations reported in the Supplementary Material 1 suggested that a recognition memory test improves MJ accuracy more than making JOLs.

As in Experiment 2, we found that scene memorability influenced MJs, JOLs, and recognition memory performance. Importantly, MJs were influenced more strongly by scene memorability in the study-and-test-task condition than in the study-and-JOL-task condition. This result suggests that completing a recognition memory test is more beneficial for MJ accuracy than making JOLs, which is consistent with the Pearson correlation analysis showing that MJ accuracy is higher when having previously taken a test than having made JOLs, but not with the gamma correlation analysis or the mixed-effects model analysis.

General discussion

Previous research revealed inconsistent results on the accuracy of metamemory for pictures of naturalistic scenes. Specifically, Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) found that MJs are unpredictive of scene memorability, whereas other studies have found that JOLs are moderately predictive of individual memory performance for scene pictures (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). One potential explanation for these discrepant results are differences in the cue basis underlying the two types of metamemory judgements. JOLs might rely more on diagnostic cues than MJs because such cues might be more available when making a judgement for one's own memory during a learning task. Alternatively, methodological differences across studies such as the memory task used for measuring stimulus memorability (i.e., classical old/new recognition memory task versus repeat detection task) or the memory criterion value used for accuracy (i.e., recognition memory performance aggregated across participants versus each participant's own memory performance) might be responsible for the discrepant results. The current study differentiated between these possibilities by systematically investigating the relative accuracy and cue basis of MJs and JOLs for pictures of scenes.

Our three experiments revealed that both MJs and JOLs are moderately accurate at predicting differences in the population memorability of scenes. This finding held across three different measures of judgement resolution (within-subjects gamma correlations, within-subjects Pearson correlations, and a mixed-effects model analysis). Our experiments also revealed that MJs and JOLs have a similar cue basis when pictures differed in aesthetics and interestingness (Experiment 1) or represented a broad range of scene memorability (Experiments 2 and 3). Thus, JOLs and MJs had similar accuracy and cue basis when obtained by similar procedures and differed only in that JOLs referred to people's own memory and were made during a learning task, whereas MJs referred to memorability is a generic item attribute and were made during a judgement-only task. This implies that discrepant findings on the accuracy of JOLs and MJs reported in prior work were largely due to methodological differences across studies.

We also found one crucial difference between MJs and JOLs. That is, MJ accuracy improved considerably when MJs were made after rather than before completing the JOL task. In contrast, this was not true for JOLs. Their accuracy was similar in both task order conditions. Experiment 3 was designed to disentangle which component of the JOL task provides the experiences participants subsequently rely on to make more accurate MJs. Results showed that a learning phase is sufficient for improving MJ accuracy as indicated by all measures of judgement resolution. In addition, Pearson correlations (but not Gamma correlations or a mixed-effects model analysis) indicated that the recognition memory test improved MJ accuracy more than making JOLs. This result is consistent with the finding that MJs were more closely related to normed values of scene memorability after having taken a memory test than after having made JOLs.

Accuracy of MJs and JOLs

Our finding that MJs are predictive of differences in the memorability of scenes at the item level contrasts with Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) results. MJ accuracy was instead consistent with the moderate accuracy of JOLs in our study and other metamemory studies (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). Our MJ results were also in line with Saito et al. (2023), who found that judgements of perceived memorability were predictive of real-world objects and faces memorability. So, evidence is accumulating that people can predict the general memorability of different types of images. This makes it even more interesting to ask for the reasons for the discrepancy in MJ results between our study and Isola et al.'s study.

One potential explanation is that we used a fine-grained judgement scale, while Isola et al. used a binary scale (yes, no). Our participants could therefore make more nuanced scene memorability predictions. However, future research will be needed to test whether the opportunity to make fine-grained distinctions between the memorability of scenes really contributes to MJ resolution. So far, one relevant prior study found that the range of confidence scales does not affect confidence accuracy in a recognition memory task (Tekin & Roediger, 2017).

Another potential explanation for why MJs were accurate in our study but not in Isola et al.'s studies might be that we measured scene memorability in an old/new recognition memory test that followed upon a learning phase, while Isola et al. used a repeat detection task. However, this explanation is inconsistent with two aspects of our results. First, Experiments 2 and 3 showed that MJs and JOLs increased with the normed values of scene memorability obtained in Isola et al.'s detection task. Second, relating MJs with normed values of scene memorability revealed very similar results as did relating MJs with the population scene memorability measure obtained in this study. These observations suggest that MJ accuracy is similar for classical old/new recognition memory tasks and repeat detection task.

Regarding the criterion for accuracy, MJs had similar moderate accuracy as JOLs at the item and individual level in our study and other metamemory studies (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). This suggests that the lack of MJ accuracy reported in Isola et al. was not due to analysing accuracy at the item level. This is not to say, however, that differences in accuracy between the item and individual level cannot exist. Quite to the contrary, idiosyncratic influences on memory and metamemory that can only contribute to judgement predictive accuracy at the individual level have been obtained in several studies (see, for example, Tullis & Fraundorf, 2017; Undorf et al., 2022).

Order effects on MJ accuracy

Experiment 3 finding that a learning phase improves MJ accuracy suggests that having seen and intentionally learned pictures for oneself provides a good basis for assessing the general memorability of pictures. Interestingly, we did not find evidence that making JOLs *per* se improved MJ accuracy. This is in line with West et al. (2023), who showed that the well-documented increase in JOL accuracy through repeated trials does not rely on making JOLs. Thus, evidence so far indicates that experience with making metamemory judgements *per* se is not essential for subsequent metamemory accuracy.

Regarding the individual contribution of the memory test on MJ accuracy, Pearson correlations showed that a recognition memory test enhances MJ accuracy relative to merely having made JOLs. This finding is consistent with positive effects of test experience on metamemory accuracy reported in studies using the same verbal materials across multiple study-test cycles (Finn & Metcalfe, 2008; Hertzog et al., 2013; King et al., 1980; Koriat & Bjork, 2006a; Touron et al., 2010; but see Mitton & Fiacconi, 2020). However, it should be considered with caution, because it did not replicate in analyses based on gamma correlations or linear mixed models. Nevertheless, in the current study with different material across trials, prior testing experience reliably increased the cue sensitivity of metamemory judgements for a new set of structurally similar scenes. This implies that participants extracted information diagnostic of memorability from testing and used this information to subsequently judge the general memorability of scenes.

To sum up, participants learned about scene memorability by experience with their own learning and testing. This illustrates what Flavell (1979) suggested in his seminal work about metacognition by saying that experiences can "affect the metacognitive knowledge base by adding to it, deleting from it, or revising it" (p. 908).

Future research directions

Given the finding that both JOLs and MJs are predictive of scene memorability, it is important to ask if participants are aware of image features diagnostic of memorability. Based on metamemory research, it is likely that some cue information reaches the level of conscious awareness (e.g., Mueller et al., 2013, 2014). However, it is also plausible that some cues remain experiential at the level of subjective feelings that may not be fully articulated, but nevertheless serve as an inferential basis for the metamemory judgements (e.g., Besken, 2016; Koriat & Levy-Sadot, 1999; Undorf et al., 2017). Prominent inferential accounts of metamemory (Koriat, 1997), distinguish between two types of processes through which cues affect metamemory judgements: theory-based and experience-based processes. Theory-based processes imply the deliberate application of explicit beliefs and knowledge about memory in general and one's own memory. In contrast, experience-based processes imply a non-analytic inferential process that operates below full awareness through which by-products of the cognitive processing of items such as the feeling of "ease" influence metamemory judgements.

Shedding light on participants awareness of stimulus memorability by examining the contributions of theory-based and experience-based processes on metamemory judgements for scene pictures would be an interesting avenue for future research. For instance, this could be done by soliciting pre-study metamemory judgements or using survey designs for assessing the contributions of beliefs to metamemory judgements about item memorability in general.

Limitations

A limitation of Experiment 3 is that we did not include a control group that completed the MJ task twice. We thus cannot fully exclude the possibility that experience with materials during an MJ task might be sufficient to increase MJ accuracy on a second trial. We do, however, regard it unlikely because completing the MJ task did not improve JOL accuracy. Nevertheless, more research will be needed to test whether completing the MJ task repeatedly improves accuracy and if so, whether the improvement is comparable to the one observed after learning pictures for oneself.

Another limitation is that our MJ task was not fully identical to Isola et al.'s task. Our aim was to examine whether differences in accuracy between JOLs and MJs were due to differences in their cue basis arising from the different aspects of memorability judged (one's own vs. generic item attribute) in different tasks (during learning vs. judgement-only). For this, it was necessary to make their procedures similar in all other respects. A potential drawback of this approach is that we cannot know which procedural change or combination of procedural changes are responsible for the differences in MJ accuracy obtained in Isola et al.'s study and the current study.

Conclusion

In conclusion, we found that the predictive accuracy of MJs is not necessarily different from that of JOLs. This stands in stark contrast to Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) findings but is consistent with evidence that metamemory for scene pictures is moderately accurate (Kao et al., 2005; Schmoeger et al., 2020; Tauber et al., 2017; Undorf & Bröder, 2021). Our work shows that people can predict not only their own future memory performance for scene pictures but also the general memorability of scene pictures with moderate accuracy. At the same time, we did find a notable difference between JOLs and MJs: MJ accuracy improves with prior learning and testing experience, whereas JOL accuracy is independent of prior assessments of general memorability. This shows that reflections about and experiences with one's own learning and memory contribute to our understanding and knowledge about metamemory and memory processes in general.

Acknowledgements

The authors thank Franziska Schäfer and Anna Fee Wefelmeier for help with material preparation.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Grants UN 345/2-1 (M.U.) and BR 2130/14-1 (A.B.) from the Deutsche

Forschungsgemeinschaft, by the University of Mannheim's Graduate School of Economic and Social Sciences (S.N.-B.) and by a Margarete von Wrangell fellowship from the state of Baden-Württemberg (M.U.).

ORCID iDs

Sofia Navarro-Báez https://orcid.org/0000-0002-6467-654X
Monika Undorf https://orcid.org/0000-0002-0118-824X

Data accessibility statement

The data and materials from the present experiment are publicly available at the Open Science Framework website: https://osf.io/hpy6q/. Experiments 1 and 2 were not preregistered. Experiment 3 was preregistered at https://osf.io/3fujm.

Supplementary material

The supplementary material is available at qjep.sagepub.com.

Notes

1. Isola, Parikh, et al. (2011) and Isola, Xiao, et al. (2011, 2014) used uncorrected hit rates as a measure of scene memorability. We think that this is not legitimate in our experiments because FA rates ranged between 9% (Experiment 1) and 16% (Experiment 3). Please note that the use of corrected hit rates is a deviation from the preregistration of Experiment 3. Importantly, all results were identical when using uncorrected hit rates except for the main effect of task in the Pearson correlation analysis in Experiment 1 and the interactive effect in the mixed-effects model analysis in Experiment 3.
2. It was impossible to investigate the relative accuracy of MJs with respect to the participant's own memory performance because participants did not complete a recognition memory test on MJ items.
3. Means and (SDs) of aesthetics and interestingness, respectively, were: 0.13 (0.07) versus 0.49 (0.11) for scenes low in aesthetics and interestingness, 0.14 (0.07) versus 0.83 (0.06) for scenes low in aesthetics and high in interestingness, 0.52 (0.11) versus 0.50 (0.10) for scenes high in aesthetics and low in interestingness, and 0.52 (.10) vs. 0.83 (0.05) for scenes high in aesthetics and interestingness.
4. Please note that low memory was a valid reason for rejection on Prolific when we collected data for Experiment 1 in 2020 (Prolific guidelines have changed in this respect in the meantime).
5. Hit rates revealed the same pattern, aesthetics: $F(1, 50) = 22.71$, $p < .001$, $\eta_p^2 = .31$, interestingness: $F(1, 50) = 11.19$, $p < .01$, $\eta_p^2 = .18$, no other effects were significant, $F <= 2.96$, $p >= .09$.
6. The MJ task numerically improved JOL resolution, but the effect was not reliable and only half the size of that for MJs.
7. Means and SDs of each of level of scene memorability were 27.63 and 4.96 (Level 1), 38.72 and 2.72 (Level 2), 44.71 and 1.38 (Level 3), 48.98 and 1.42 (Level 4), 53.80 and 1.42 (Level 5), 58.22 and 1.38 (Level 6), 62.95 and 1.52 (Level 7), 67.75 and 1.52 (Level 8), 72.52 and 1.60 (Level 9), 83.04 and 4.55 (Level 10).

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610–632. https://doi.org/10.1016/0749-596X(89)90016-8

Besken, M. (2016). Picture-perfect is not perfect for metamemory: Testing the perceptual fluency hypothesis with degraded images. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1417–1433. https://doi.org/10.1037/xlm0000246

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

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*, 165–178. https://doi.org/10.1016/j.visres.2015.03.005

Caplan, J. B., Sommer, T., Madan, C. R., & Fujiwara, E. (2019). Reduced associative memory for negative information: Impact of confidence and interactive imagery during study. *Cognition and Emotion*, *33*(8), 1745–1753. https://doi.org/10.1080/02699931.2019.1602028

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, *22*(3), 273–280. https://doi.org/10.3758/BF03200854

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

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*(4), 374–380. https://doi.org/10.3758/BF03210921

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language*, *33*(4), 545–565. https://doi.org/10.1006/jmla.1994.1026

Dunlosky, J., & Thiede, K. W. (2013). *Metamemory*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0019

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

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 1–29. https://doi.org/10.1167/7.1.10

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*(1), 19–34. https://doi.org/10.1016/j.jml.2007.03.006

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 22–34. https://doi.org/10.1037/0278-7393.29.1.22

Hertzog, C., Hines, J. C., & Touron, D. R. (2013). Judgments of learning are influenced by multiple cues in addition to memory for past test accuracy. *Archives of Scientific Psychology*, *1*(1), 23–32. https://doi.org/10.1037/arc0000003

Hintzman, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, *87*(4), 398–410. https://doi.org/10.1037/0033-295X.87.4.398

Hourihan, K. L. (2020). Misleading emotions: Judgments of learning overestimate recognition of negative and positive emotional images. *Cognition and Emotion*, *34*(4), 771–782. https://doi.org/10.1080/02699931.2019.1682972

Hourihan, K. L., & Bursey, E. (2017). A misleading feeling of happiness: Metamemory for positive emotional and neutral pictures. *Memory*, *25*(1), 35–43. https://doi.org/10.1080/09658211.2015.1122809

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). Understanding the intrinsic memorability of images. *Advances in Neural Information Processing Systems*, *24*, 2429–2437. https://doi.org/10.1167/12.9.1082

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In 24th IEEE conference on computer vision and pattern recognition (CVPR) (pp. 145–152).

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*(1), 186–200. https://doi.org/10.1037/a0025960

Kao, Y.-C., Davis, E. S., & Gabrieli, J. D. E. (2005). Neural correlates of actual and predicted memory formation. *Nature Neuroscience*, *8*(12), 1776–1783. https://doi.org/10.1038/nn1595

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*(1), 92–107. https://doi.org/10.3758/BF03211579

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology*, *93*(2), 329–343. https://doi.org/10.2307/1422236

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

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and Cognition*, *19*(1), 251–264. https://doi.org/10.1016/j.concog.2009.12.010

Koriat, A., & Bjork, R. A. (2006a). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959–972. https://doi.org/10.3758/BF03193244

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments. *Process Theories*, *20*, 483–502.

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 595–608. https://doi.org/10.1037/0278-7393.32.3.595

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 464–470. https://doi.org/10.1037/0278-7393.16.3.464

Lin, Q., Yousif, S. R., Chun, M. M., & Scholl, B. J. (2021). Visual memorability in the absence of semantic content. *Cognition*, *212*, 104714. https://doi.org/10.1016/j.cognition.2021.104714

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 509–527. https://doi.org/10.1037/a0014876

Mitton, E. E., & Fiacconi, C. M. (2020). Learning from (test) experience: Testing without feedback promotes metacognitive sensitivity to near-perfect recognition memory. *Zeitschrift für Psychologie*, *228*(4), 264–277. https://doi.org/10.1027/2151-2604/a000424

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, *20*(2), 378–384. https://doi.org/10.3758/s13423-012-0343-6

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1287–1306. https://doi.org/10.1037/a0036914

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., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, *2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation* (*Vol. 26*, pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *19*(2), 155–160. https://doi.org/10.1037/h0082899

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(3), 255–287. https://doi.org/10.1037/h0084295

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), 20. https://doi.org/10.5334/irsp.181

Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency—Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction*, *44*, 31–40. https://doi.org/10.1016/j.learninstruc.2016.01.012

Price, J., & Harrison, A. (2017). Examining what prestudy and immediate judgments of learning reveal about the bases of metamemory judgments. *Journal of Memory and Language*, *94*, 177–194. https://doi.org/10.1016/j.jml.2016.12.003

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Saito, J. M., Kolisnyk, M., & Fukuda, K. (2023). Judgments of learning reveal conscious access to stimulus memorability. *Psychonomic Bulletin & Review*, *30*, 317–330. https://doi.org/10.3758/s13423-022-02166-1

Schmoeger, M., Deckert, M., Loos, E., & Willinger, U. (2020). How influenceable is our metamemory for pictorial material? The impact of framing and emotionality on metamemory judgments. *Cognition*, *195*(104112), 1–10. https://doi.org/10.1016/j.cognition.2019.104112

Schwaninger, A., Wallraven, C., & Bülthoff, H. H. (2004). Computational modeling of face recognition based on psychophysical experiments. *Swiss Journal of Psychology*, *63*(3), 207–215. https://doi.org/10.1024/1421-0185.63.3.207

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 156–163. https://doi.org/10.1016/S0022-5371(67)80067-7

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2014). Measuring memory and metamemory. In *Handbook of metamemory and memory*. Routledge. https://doi.org/10.4324/9780203805503.ch6

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207–222. https://doi.org/10.1080/14640747308400340

Tauber, S. K., Dunlosky, J., Urry, H. L., & Opitz, P. C. (2017). The effects of emotion on younger and older adults' monitoring of learning. *Aging, Neuropsychology, and Cognition*, *24*(5), 555–574. https://doi.org/10.1080/13825585.2016.1227423

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications*, *2*(1), 49. https://doi.org/10.1186/s41235-017-0086-z

Touron, D. R., Hertzog, C., & Speagle, J. Z. (2010). Subjective learning discounts test type: Evidence from an associative learning and transfer task. *Experimental Psychology*, *57*(5), 327–337. https://doi.org/10.1027/1618-3169/a000039

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language*, *95*, 124–137. https://doi.org/10.1016/j.jml.2017.03.003

Undorf, M. (2020). Fluency illusions in metamemory. In A. M. Cleary & B. L. Schwartz (Eds.), *Memory quirks* (1st ed., pp. 150–174). Routledge. https://doi.org/10.4324/9780429264498-12

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629–642. https://doi.org/10.1177/1747021819882308

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

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264–1269. https://doi.org/10.1037/a0023719

Undorf, M., Navarro-Báez, S., & Bröder, A. (2022). "You don't know what this means to me"—Uncovering idiosyncratic influences on metamemory judgments. *Cognition*, *222*(105011), 1–9. https://doi.org/10.1016/j.cognition.2021.105011

Undorf, M., Navarro-Báez, S., & Zimdahl, M. F. (2022). Metacognitive illusions. In R. F. Pohl (Ed.), *Cognitive illusions* (3rd ed., pp. 307–323). Routledge. https://doi.org/10.4324/9781003154730-22

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., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language*, *92*, 293–304. https://doi.org/10.1016/j.jml.2016.07.003

West, J. T., Kuhns, J. M., Touron, D. R., & Mulligan, N. W. (under review). *Increased metamemory accuracy with practice does not require practice with metamemory*.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3485–3492). https://doi.org/10.1109/CVPR.2010.5539970

Supplementary Material 1

Pearson correlations between population scene memorability and JOLs or MJs in Experiment 1 and 2

Table 1 shows mean Pearson correlations between metamemory judgments and population scene memorability for each task in each task order condition of Experiments 1, 2, and 3. All correlations were significantly positive, t >= 5.45, $p < .001$ (Experiment 1), t >= 5.85, $p < .001$ (Experiment 2), corroborating results from gamma correlations indicating that JOLs and MJs captured differences in population scene memorability. A 2 (task: JOL vs. MJ) × 2 (task order condition: JOLs-first vs. MJs-first) repeated-measures ANOVA on Fisher-Z-transformed values revealed a main effect of task in Experiment 1, $F(1, 50) = 5.15$, $p = .028$, $\eta_p^2 = .09$, no main effect of task in Experiment 2, $F < 1$, no main effects of task order condition, $F < 1$ (Experiment 1), $F(1, 48) = 2.27$, $p = .14$, $\eta_p^2 = .04$ (Experiment 2), and significant interactions, $F(1, 50) = 26.31$, $p < .001$, $\eta_p^2 = .35$ (Experiment 1), $F(1, 48) = 12.99$, $p < .001$, $\eta_p^2 = .21$ (Experiment 2). As with the gamma correlations, planned comparisons indicated that Pearson correlations (Fisher-Z-transformed) for JOLs did not differ between conditions, $t(50) = 1.67$, $p = .10$, $d = 0.47$ (Experiment 1), $t < 1$, $p = 0.80$, $d = 0.07$ (Experiment 2), whereas Pearson correlations for MJs were higher in the JOLs-first condition than in the MJs-first condition, $t(50) = 2.85$, $p < .01$, $d = 0.81$ (Experiment 1), $t(48) = 2.95$, $p < .01$, $d = 0.85$ (Experiment 2), indicating higher relative accuracy of MJs when made after a JOL task with a recognition memory test.

Pearson correlations between population scene memorability and MJs in Experiment 3

All correlations were significantly positive, t >= 9.06, $p < .001$, corroborating results from gamma correlations indicating that MJs in all conditions captured differences in population scene memorability. As with gamma correlations, planned contrasts on Fisher-Z-transformed values revealed that the learning phase present in all experimental conditions improved MJ accuracy compared to only making MJs, $t(201) = 3.46$, $p < .001$. They also revealed that MJ accuracy did not differ between the full-JOL-task condition and the conditions with one component of the JOL task only (i.e., study-and-JOL-task, study-and-test-task), $t(201) = 0.87$, $p = .38$. In contrast to gamma correlations, MJ accuracy was higher in the study-and-test-task condition than in the study-and-JOL-task condition, $t(201) = 2.10$, $p = .036$, indicating that taking a test improved MJ accuracy more than making JOLs.

Table 1

Means (SDs) of the Pearson Correlation Between Population Scene Memorability and JOLs or MJs in Each Task Order Condition of Experiments 1, 2, and 3

Experiment and condition	Metamemory judgments	
Experiment 1	JOLs	MJs
JOLs first	.23 (.15)	.40 (.19)
MJs first	.31 (.20)	.24 (.23)
Experiment 2		
JOLs first	.29 (.16)	.39 (.17)
MJs first	.30 (.22)	.23 (.20)
Experiment 3		
MJ-task-only	-	.26 (.21)
Full-JOL-task	.32 (.16)	.39 (.17)
Study-and-JOL-task	.34 (.15)	.32 (.18)
Study-and-test-task	-	.40 (.20)

Note. JOLs = judgments of learning, MJs = memorability judgments, Population scene memorability = hit rate minus false alarm rate per scene across participants in each experiment

Supplementary Material 2

Mixed-effects model analysis predicting population scene memorability from metamemory judgments in Experiment 1, 2, and 3

Experiment 1

In the mixed-effects model analysis, we predicted population scene memorability from task (effect coded: -1 = MJ, 1 = JOL), task order condition (effect coded: -1 = MJs first, 1 = JOLs first), standardized metamemory judgments, and their interactions. We treated task, task order condition, and metamemory judgments as fixed effects. We specified random participant slopes for metamemory judgments to account for differences in accuracy across participants. We did not specify random intercepts for participants because population scene memorability was calculated by aggregating memory performance across participants. This model provided a significantly better fit to the data than an otherwise identical model without metamemory judgments as a fixed-effects predictor, $\chi^2(4) = 117.97$, $p < .001$, indicating that the metamemory judgments capture differences in the population memorability of scenes. This conclusion was also corroborated by the finding of a significantly positive regression weight for metamemory judgments, $b = 0.06$ ($SE = 0.01$), $t = 10.89$, $p < .001$. At the same time, main effects of task, $b = 0.01$ ($SE = 0.002$), $t = 5.88$, $p < .001$, and task order condition, $b = -0.01$ ($SE = 0.002$), $t = -3.15$, $p < .01$, as well as significant interactions between task and metamemory judgments, $b = -0.01$ ($SE = 0.002$), $t = -4.61$, $p < .001$, and among task, task order condition, and metamemory judgments, $b = -0.01$ ($SE = 0.002$), $t = -6.25$, $p < .001$, suggested differences in the relative accuracy of metamemory judgments across tasks and order conditions.

Separate models for each task revealed that JOLs and MJs significantly predicted population scene memorability, $b = 0.05$ ($SE = 0.01$), $t = 9.08$, $p < .001$, $b = 0.07$ ($SE = 0.01$), $t = 9.73$, $p < .001$; respectively. In the JOL task, a main effect of order condition, $b =$ -

0.01 (*SE* = 0.003), *t* = -4.37, *p* < .01, and no interaction between JOLs and order condition, *b* = -0.01 (*SE* = 0.01), *t* = -1.66, *p* = .10, showed that the predictive value of JOLs for population scene memorability did not differ between order conditions. In contrast, in the MJ task, the interaction between MJs and order condition, *b* = 0.02 (*SE* = 0.01), *t* = 2.68, *p* < .01, indicated that MJs were more accurate when made after the JOL task. The main effect of order condition was not significant, *b* = -0.001 (*SE* = 0.003), *t* = -0.39, *p* = .70.

Experiment 2

A mixed-effects model with standardized metamemory judgments as fixed-effects predictor provided a significantly better fit to the data than an otherwise identical model with task, task order condition and without metamemory judgments as a fixed-effects predictor, $\chi^2(4) = 97.60$, *p* < .001. Results showed that metamemory judgments significantly predicted population scene memorability, *b* = 0.06 (*SE* = 0.01), *t* = 10.73, *p* < .001. There was also a main effect of task, *b* = 0.01 (*SE* = 0.002), *t* = 4.67, *p* < .001, a main effect of task order condition, *b* = -0.02 (*SE* = 0.002), *t* = -9.32, *p* < .001, and significant interactions between task and order condition, *b* = 0.01 (*SE* = 0.002), *t* = 6.95, *p* < .001, between task and metamemory judgments, *b* = -0.004 (*SE* = 0.002), *t* = -2.04, *p* = .041, and among task, task order condition, and metamemory judgments, *b* = -0.01 (*SE* = 0.002), *t* = -5.83, *p* < .001, suggesting differences in the relative accuracy of metamemory judgments across tasks and order conditions.

In separate models for each task, JOLs and MJs significantly predicted population scene memorability, *b* = 0.06 (*SE* = 0.01), *t* = 8.16, *p* < .001, *b* = 0.06 (*SE* = 0.01), *t* = 9.69, *p* < .001; respectively. In the JOL task, a main effect of order condition, *b* = -0.01 (*SE* = 0.003), *t* = -2.26, *p* = .024, and no interaction, *b* = -0.01 (*SE* = 0.01), *t* = -0.74, *p* = .46, indicated that the predictive value of JOLs for population scene memorability did not differ between order conditions. In the MJ task, there was not only a main effect of order condition, *b* = -0.03 (*SE* = 0.003), *t* = -9.81, *p* < .001, but also the interaction was significant, *b* = 0.02 (*SE* = 0.01), *t* = 3.11, *p* < .01, indicating that the predictive value of MJs for population scene memorability was higher after a JOL task.

Experiment 3

In the mixed-effects model analysis, we predicted population scene memorability from condition using Helmert coding[1] (see also pre-registration, https://osf.io/3fujm), standardized MJs, and their interactions. We treated condition, and MJs as fixed effects predictors. We specified random participant slopes for MJs to account for differences in accuracy across participants. We did not specify random intercepts for participants because population scene memorability was calculated by aggregating memory performance across participants.

This model provided a significantly better fit to the data than an otherwise identical model without MJs as a fixed-effects predictor, $\chi2(4) = 242.02$, p < .001, indicating that MJs capture differences in the population memorability of scenes as in Experiment 1 and 2. At the same time, there was a significantly positive regression weight for MJs, *b* = 0.06 (*SE* = 0.003), *t* = 20.84, *p* < .001, which corroborates the above conclusion that MJs are predictive of scene memorability. There were also main effects of condition: the first contrast indicated that scene memorability was higher in the MJ-task-only condition vs. all other conditions, *b* = -0.01 (*SE* = 0.003), *t* = -4.25, *p* < .001, the second contrast indicated that scene memorability was higher in the study-and-JOL-task and study-and-test-task conditions vs. the full-JOL-task condition, *b* = -0.02 (*SE* = 0.003), *t* = -7.78, *p* < .001, and the third contrast indicated that

[1] The first contrast tested the difference between the control condition (MJs-only; -3/4) and all three experimental conditions (full-JOL-task, study-and-JOL-task, study-and-test-task; coded all as +1/4). The second contrast tested the difference between the full-JOL-task condition (+2/3) and the other two experimental conditions (study-and-JOL-task group, the study-and-test-task group; coded both as -1/3). The third contrast tested the difference between the study-and-JOL-task condition (-1/2) and the study-and-test-task condition (+1/2).

scene memorability was higher in the study-and-JOL-task condition vs. study-and-test-task condition, b = -0.01 (SE = 0.003), t = -2.28, p = .022. As in Gamma and Pearson correlations, the significant interaction between MJs and the first contrast, b = 0.01 (SE = 0.01), t = 2.08, p = .038, indicated that MJs were more predictive of population scene memorability in the conditions with any component of the JOL task (i.e., full-JOL-task condition, study-and-JOL-task condition, study-and-test-task condition) than in the MJ-task-only condition. The interactions of scene memorability and the second and third contrast were non-significant t < 1.13.

Supplementary Material

Experiment 1

Table S1

Mean (SDs) of Hit Rates (Hits), False Alarm Rates (FAs), and Hit Rates Corrected (Pr) of Scenes used in Experiment 1

Interestingness	Aesthetics	Hits %	FAs %	*P*r
Low	Low	74.12 (14.19)	6.94 (9.74)	67.17 (17.19)
Low	High	64.14 (19.44)	9.61 (11.30)	54.53 (22.45)
High	Low	76.16 (16.63)	6.24 (7.87)	69.93 (19.08)
High	High	69.08 (19.68)	12.11 (13.53)	56.97 (24.17)

Experiment 2

Table S2

Mean (SDs) of Hit Rates (Hits), False Alarm Rates (FAs), and Hit Rates Corrected (Pr) of Scenes used in Experiment 2

Scene Memorability Level	Hits %	FAs %	*P*r
1	54.19 (15.07)	14.29 (11.33)	39.90 (15.60)
2	61.85 (12.37)	14.04 (12.19)	47.81 (17.18)
3	63.10 (13.55)	16.54 (13.10)	46.56 (16.65)
4	64.25 (14.61)	11.89 (8.00)	52.36 (15.47)
5	71.12 (13.11)	10.77 (10.04)	60.36 (14.51)
6	71.56 (13.44)	13.46 (12.55)	58.10 (19.78)

7	73.80 (12.99)	14.31 (12.74)	59.49 (16.72)
8	78.25 (10.79)	8.06 (9.88)	70.19 (14.38)
9	82.61 (11.71)	6.16 (8.19)	76.45 (15.67)
10	90.93 (9.16)	4.49 (5.27)	86.44 (10.33)

Experiment 3

Table S3

Mean (SDs) of Hit Rates (Hits), False Alarm Rates (FAs), and Hit Rates Corrected (Pr) of Scenes used in Experiment 3

Scene Memorability Level	Hits %	FAs %	*P*r
1	52.18 (14.82)	21.03 (9.85)	31.16 (15.00)
2	58.07 (9.09)	18.59 (8.93)	39.47 (11.87)
3	62.58 (11.17)	21.82 (8.74)	40.76 (13.02)
4	62.55 (12.75)	17.70 (9.88)	44.85 (12.45)
5	65.41 (11.16)	15.99 (8.57)	49.41 (12.48)
6	68.43 (8.72)	16.60 (11.07)	51.83 (12.51)
7	67.84 (8.53)	15.15 (8.97)	52.69 (10.38)
8	71.56 (10.97)	11.28 (8.09)	60.28 (11.56)
9	76.08 (8.89)	11.64 (8.91)	64.44 (11.66)
10	80.96 (9.95)	6.96 (4.52)	74.00 (11.28)

Mending metacognitive illusions requires metacognitive feedback

Sofia Navarro-Báez[12], Arndt Bröder[1], & Monika Undorf [2]

Word Count: 9,659

Authors Note

[1]Department of Psychology, Technical University of Darmstadt.

[2]Department of Psychology, School of Social Sciences, University of Mannheim.

All data, materials, and analysis code are available at

https://osf.io/vgy7d/?view_only=5d02381f12754484ad1c422cd1d4f44b

Abstract

The metacognitive monitoring of cognitive processes, such as learning and memory, is not always accurate. Metacognitive illusions occur when metacognitive judgments rely on invalid information or fail to rely on valid information. Given the importance of accurate metacognitive monitoring for the effective regulation of behavior, this study tested the effectiveness of feedback in mending metacognitive illusions. Across three experiments, participants completed three study-test cycles in which they studied three different study lists. Participants made judgments of learning (JOLs) and received feedback or no feedback after each cycle. In Experiments 1 and 2, cognitive feedback about one's own recall performance and JOL for each studied item was provided. In Experiment 3, additional metacognitive feedback informed learners about possible metacognitions during the task, their biased nature, and ways to enhance JOLs. Results showed that cognitive feedback was not effective for mending the font size illusion (Experiments 1 and 2), the stability bias (Experiment 1), or the font format illusion (Experiment 2). In contrast, metacognitive feedback remedied the stability bias and in turn improved relative JOL accuracy (Experiment 3). Moreover, the font size illusion decreased across cycles when manipulated orthogonally to a valid cue (Experiments 1 and 3), but not to an invalid cue (Experiment 2). In conclusion, this study shows that cognitive feedback alone is not enough for improving the cue basis and resolution of JOLs but rather metacognitive feedback with an in-depth explanation of biased metacognition is needed.

Keywords: metacognitive illusions; metamemory; judgments of learning; cognitive feedback, metacognitive feedback

Introduction

Accurate *metacognitive monitoring* — the real-time assessment of cognitive processes — is important because it guides behavior (Nelson & Narens, 1990). For example, a student with accurate metacognitive monitoring may correctly identify which topics she has mastered sufficiently for the exam and continue to study those that she has not yet mastered. In fact, experimental studies show that participants with higher monitoring accuracy can regulate their learning better by selecting material to restudy more appropriately (Dunlosky et al., 2021; Thiede et al., 2003; Tullis & Benjamin, 2012). This ultimately leads to better grades as shown by a meta-analysis demonstrating the positive relationship between metacognition and academic performance even when controlling for intelligence (Ohtani & Hisasaka, 2018). However, unfortunately, metacognition is not always helpful because monitoring is sometimes incorrect. Metacognitive monitoring judgments are inferential in nature and rely on cues, which are broadly categorized into characteristics of the study condition (e.g., study strategies used), inherent characteristics of the study material (e.g., concreteness of study words), and subjective experiences of ease (Koriat, 1997). The accuracy of metacognitive judgments suffers when these rely on invalid cues or fail to rely on the valid cues (Undorf et al., 2022).

Improving self-regulated learning thus requires metacognitive awareness of cue validity. This is that people's metacognitive judgments rely on valid cues and ignore invalid cues. *How to improve metacognitive awareness of cue validity* is therefore a practically relevant question. So far, however, research has mainly proven that correcting metacognitive illusions (i.e., systematic dissociations between people's metacognitions and cognitions) is very difficult, often does not work at all, and when it does, the improvement is very small (e.g., Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Kornell & Bjork, 2009; Mueller et al., 2015; Pan & Rivers, 2023; Yan et al., 2016). In the current study, we tested the effectiveness of two forms of feedback, *cognitive feedback* (Balzer et al., 1989) alone and with additional *metacognitive feedback* (Fiedler et al., 2020). We tested whether these can improve metacognitive awareness of cue validity in *judgments of learning* (JOLs) – predictions of future memory performance (Rhodes, 2016). In the following, we will review the methods that have and have not been successful in promoting metacognitive awareness of the validity (or invalidity) of cues. We will then discuss the cognitive feedback method tested in Experiment 1 and 2 which has been demonstrated to improve the cue basis and accuracy of judgments about the external world (e.g., Karlsson et al., 2004; Little & Lewandowsky, 2009; Newell et al., 2009; Smithson et al., 2023).

Methods used to foster metacognitive awareness

Several studies have used experience across multiple learning-test cycles as a method to improve metacognitive awareness of cues (Castel, 2008; Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Mueller et al., 2015; Pan & Rivers, 2023; Sungkhasettee et al., 2011; Tauber & Rhodes, 2010). The idea is that encoding and retrieval experiences from learning and test phases aid learners to become aware of cues that are helpful for their learning and memory and this fosters adaptive cue utilization in metacognitive judgments. This is especially the case in studies using novel item lists across cycles. When making metacognitive judgments for the same materials, memory of past performance can be used as a heuristic (Ariel & Dunlosky, 2011; Finn & Metcalfe, 2008; Koriat & Bjork, 2006a; Tauber & Rhodes, 2012). This strong heuristic cue might prevent cue discovery by overshadowing subtler cues. The former studies with novel item lists across cycles illustrate that participants acquire correct knowledge about the effectiveness of study strategies (e.g., imagery is better than repetition) as demonstrated by global predictions and strategy effectiveness ratings, but fail to use this knowledge in their item-by-item metacognitive judgments (Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Mueller et al., 2015). Other studies have shown that task experience is not helpful for metacognitive judgments to become sensitive to word orientation (i.e., greater memory performance for words presented inverted rather than upright, as shown by Sungkhasettee et al., 2011), serial position effects in memory (Castel, 2008), and the benefit of pre-testing as a study strategy (Pan & Rivers, 2023). An exception is a study in which metacognitive judgments became more accurate at predicting greater memory performance for occupations than for surnames (Tauber & Rhodes, 2010). Overall, despite the easy accessibility of direct task experience, it does not generally increase the sensitivity of metacognitive judgments to valid cues.

Warnings are another method researchers have used in attempts to foster that metacognitive judgments rely on valid cues and ignore invalid cues in situations where metacognitive illusions occur. For instance, Kornell and Bjork (2009) warned participants to keep in mind that their future memory will improve with the number of study opportunities, however, participants continued underestimating their future learning. Rhodes and Castel (2008) warned participants that the font size of words will not affect their future memory, but participants still relied on font size for their JOLs even though actual memory performance was unaffected by font size. Yan et al. (2016) informed participants about the superiority of interleaving exemplars of to-be-learned categories over blocking exemplars by category, but participants still predicted benefits of blocking for themselves. Although most studies have proven that warnings are ineffective, warnings can be successful when detailed information that is tailored and delivered individually to participants is provided (Koriat & Bjork, 2006b; Miller & Geraci, 2011). However, this comes with the limitation that information tailored and delivered individually is difficult or even impossible to implement. To sum up, these studies demonstrate the persistence of metacognitive illusions and lack of reliance on valid cues even when the experiment instructions provide very explicit warnings.

Increasing the salience of relevant aspects of the task can be a successful intervention for fostering cue awareness. Participants accurately predicted the primacy and recency effect in memory when the serial position of items was presented before studying each item (Castel, 2008). Studies asking participants to study items with differentially effective strategies on a within-subjects basis have made the strategy cue salient by having strategy blocks (e.g., first all imagery items, then all repetition items, or vice versa). In general, having blocks instead of a random intermix of items is partially effective for making the strategy cue salient and making JOLs rely on the strategy cue. In the study by Yan et al. (2016), this method was successful when the highly effective strategy was presented first but not when it was presented second. In the study by Price et al., (2008), there was a moderate improvement of JOLs. In sum, making cues salient can sometimes be an effective intervention for improving judgment cue sensitivity.

Finally, performance feedback has been suggested as a method for participants to identify effective study strategies from task experience according to the *inferential deficit hypothesis* (Dunlosky & Hertzog, 2000; Matvey et al., 2002). This hypothesis states that there are limited cognitive resources to monitor test performance and make inferences about valid cues. Thus, studies have provided performance feedback related to the cues manipulated with the aim to support participants to distinguish the validity of cues from their test performance (Pan & Rivers, 2023; Tullis et al., 2013). The study by Tullis et al. (2013) provided participants with the number of correctly recalled restudied and pre-tested items. This resulted in a correct identification that pre-testing is a more effective strategy than re-studying. However, the study by Pan and Rivers (2023) had to prompt participants to recall their predictions in addition to providing them with feedback on their test performance to enhance their awareness of the superiority of the pre-testing strategy. Importantly, both studies used global predictions, which largely reflect metamemory beliefs, rather than item-by-item predictions, where learning experiences contribute more. One study with item-by-item JOLs providing aggregated performance feedback (i.e., number of words recalled per item type) showed that feedback was effective for improving global predictions but not immediate JOLs (Mueller et al., 2015). Overall, there is evidence that performance feedback is effective for improving global judgments, but its effectiveness for item-by-item JOLs remains to be further examined.

Cognitive feedback

Research on judgment and decision making assumes that people rely on cues when making judgments about the external world. A famous paper by Brehmer (1980) reviewed studies on the ability to learn from experience in probabilistic situations such as clinical inference. He came to the pessimistic conclusion that people do not learn from experience with mere outcome feedback (i.e., knowing the outcome) in complex and uncertain tasks because of the number of biases that prevent them from learning. He argues that learning is not a passive process of just observing the world, but rather an active process of hypothesis testing. However, people usually confirm hypothesis rather than refute incorrect hypothesis, focus on positive information, neglect negative information, and have the implicit assumption that the world is deterministic rather than probabilistic. Because of all these reasons, any knowledge acquired by experience is not necessarily valid.

In contrast, a review by Balzer et al. (1989) revealed that so-called *cognitive feedback* is effective for improving the cue basis of judgments compared to mere outcome feedback. Cognitive feedback consists of task information, cognitive information, and functional validity information. Task information refers to the relations between cues and criterion (i.e., task system) – “Which cues are valid?*”*. Cognitive information refers to the subject’s cognitive system – “How do people use the cues for their judgments?”. Functional validity information refers to the relation of the cognitive system to the task system – “What is the difference between the former two?". Several studies have demonstrated that cognitive feedback improves the cue basis and accuracy of judgments about the external world (Karlsson et al., 2004; Little & Lewandowsky, 2009; Newell et al., 2009; Seong & Bisantz, 2008; Smithson et al., 2023). For example, Newell et al. (2009) showed that cognitive feedback helped participants to distinguish the predictive validity of different cues for the share price of a fictional stock company. Similarly, in the study by Seong & Bisantz (2008), participants learned to identify an aircraft moving on a simulated radar screen as either hostile or friendly with four key aircraft parameters: speed, altitude, range, and time. This identification process was aided by an automated decision-aid and cognitive feedback about the functioning of the decision-aid was given.

Since judgments about external criteria are like judgments about one's own cognition in that both judgments rely on probabilistic cues, cognitive feedback may be beneficial for learning to distinguish the different predictive validities of cues in metacognition as well. Specially, the opportunity to relate JOLs to actual memory performance (i.e., functional validity information) should support an understanding of which cues are valid for memory performance and in which direction (i.e., this helps versus impairs memory), and which ones are not valid at all (i.e., this does not affect memory). The critical aspect is to transfer the acquired knowledge from cognitive feedback to the online monitoring situation during learning. This might be challenging when other varying cues that are not under the control of experimenter such as idiosyncratic cues are present (Bröder & Undorf, 2019; Undorf, et al., 2022). Given the importance of finding effective ways to train metacognition, it is worth testing cognitive feedback as an intervention to improve the cue use and the accuracy of item-by-item JOLs. Thus, in Experiment 1 and 2, we tested the effectiveness of cognitive feedback. Since the cognitive feedback intervention was surprisingly ineffective, in

Experiment 3, we additionally included a different form of feedback, *metacognitive feedback,* which we will discuss later.

This study

In this study, we tested the effectiveness of different forms of feedback in improving the cue basis and accuracy of judgments of learning (JOLs). Across three experiments, participants completed three study-test cycles with JOLs and received either feedback or no feedback after each study-test cycle. In all three experiments, we orthogonally manipulated two cues in total. In all experiments, we manipulated font size (18 pt vs. 48 pt), a cue that is overweighted in JOLs — large-font words elicit higher JOLs than small-font words but font size has a very small or no effect on recall performance (Chang & Brainerd, 2022; Luna et al., 2018; Rhodes & Castel, 2008). In Experiment 1 and 3, additional to font size, we manipulated the number of study opportunities (1 vs. 2), a cue that is underweighted in JOLs — JOLs do not differ between once-learned and twice-learned words while twice-learned words are better recalled (Kornell & Bjork, 2009). In Experiment 2, the additional cue manipulated was font format (standard vs. aLtErnAtiNg), a cue that is overweighted in JOLs — standard-format words elicit higher JOLs than alternating-format words but font format has no effect on recall performance (Rhodes & Castel, 2008)[1].

We hypothesized that feedback would lead to JOLs relying increasingly on valid cues (i.e., number of study opportunities) and ignore invalid cues (i.e., font size, font format). This means that JOLs should be higher for twice-learned than once-learned words after feedback. Further, JOLs should not differ between large-font and small-font words, and between standard-font and alternating-font words after feedback. If, in contrast, these improvements occur in both groups, this would indicate that study-test experience is beneficial for learning cue validities and implementing them in JOLs.

[1] Mueller et al. (2013) found worse recall performance for alternating-format than standard-format word pairs.

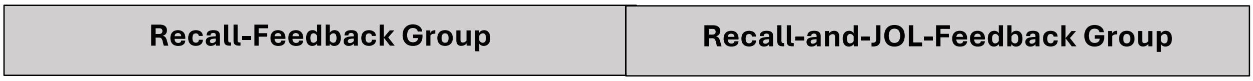
Experiment 1

Experiment 1 aimed to test the effectiveness of cognitive feedback for increasing JOL reliance on number of study opportunities (i.e., valid cue) and for decreasing JOL reliance on font size (i.e., invalid cue). Experiment 1 entailed four between-subjects groups. In the control group, participants received no feedback, so they only had their own memory of test performance as feedback on cue validity. In the outcome feedback group (recall-feedback group), participants saw the words they had recalled and not recalled, organized by the two cues (see Figure 1). In the cognitive feedback group (recall-and-JOL-feedback group), the list was accompanied by the JOL participants had given to each word during study. This enables to compare the actual recall (i.e., task information) with their prediction (i.e., cognitive information). Finally, the social-reference-feedback group was informed about the cues manipulated in the experiment and their cue validity, accompanied by a table showing the average performance of other participants doing this task.

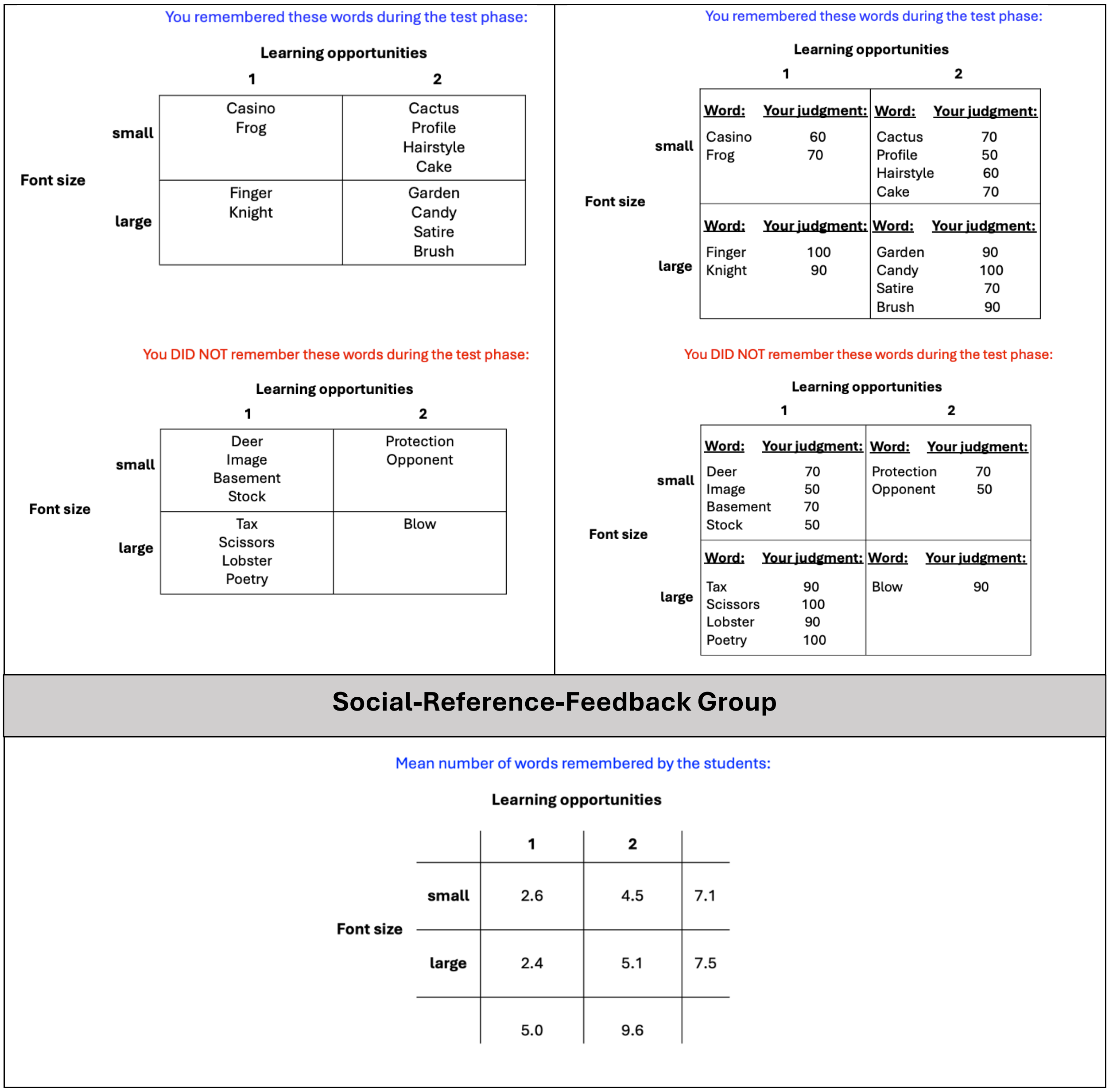
We hypothesized that the cognitive information in the recall-and-JOL-feedback group would improve cue weighting in JOLs (i.e., effect of number of study opportunities and no effect of font size) and, in turn, increase relative accuracy. At the same time, it was an open question whether the outcome feedback in the recall-feedback group would lead to better cue weighting and accuracy as suggested by the inferential deficit hypothesis (Hertzog et al., 2009; Matvey et al., 2002) or whether social reference information would be sufficient for improvements or even go beyond the cognitive feedback.

Figure 1

Example Feedback Presented to Participants in the Recall-Feedback, Recall-and-JOL-Feedback, and Social-Reference-Feedback Group in Experiment 1



Recall-Feedback Group	Recall-and-JOL-Feedback Group



You remembered these words during the test phase:

Learning opportunities

Font size		1	2
	small	Casino Frog	Cactus Profile Hairstyle Cake
	large	Finger Knight	Garden Candy Satire Brush

You DID NOT remember these words during the test phase:

Learning opportunities

Font size		1	2
	small	Deer Image Basement Stock	Protection Opponent
	large	Tax Scissors Lobster Poetry	Blow

You remembered these words during the test phase:

Learning opportunities

Font size		1		2	
		Word:	Your judgment:	Word:	Your judgment:
	small	Casino Frog	60 70	Cactus Profile Hairstyle Cake	70 50 60 70
		Word:	Your judgment:	Word:	Your judgment:
	large	Finger Knight	100 90	Garden Candy Satire Brush	90 100 70 90

You DID NOT remember these words during the test phase:

Learning opportunities

Font size		1		2	
		Word:	Your judgment:	Word:	Your judgment:
	small	Deer Image Basement Stock	70 50 70 50	Protection Opponent	70 50
		Word:	Your judgment:	Word:	Your judgment:
	large	Tax Scissors Lobster Poetry	90 100 90 100	Blow	90

Social-Reference-Feedback Group

Mean number of words remembered by the students:

Learning opportunities

Font size		1	2	
	small	2.6	4.5	7.1
	large	2.4	5.1	7.5
		5.0	9.6	

Method

Transparency and Openness

In this and all subsequent experiments, we report how we determined our sample size, all data exclusions, all manipulations, and all measures. All data, materials, and analysis code are available at https://osf.io/vgy7d/?view_only=5d02381f12754484ad1c422cd1d4f44b. Designs and analyses of all experiments were not preregistered.

Design

The design was a mixed design with font size (18 point, 48 point), number of study presentations (1, 2), and study-test cycle (1, 2, 3) as within-subjects factors and feedback group (no-feedback, recall-feedback, recall-and-JOL feedback, social-reference-feedback) as a between-subjects factor.

Materials

Stimuli were 120 German six-letter nouns. All words were of neutral valence (M = 0.36, SD = 0.93; rated on a 7-point scale, -3 = *very negative* to 3 = *very positive*), moderate arousal (M = 2.59, SD = 0.80; rated on a 5-point scale, 1 = *low arousal* to 5 = *high arousal*), and moderate concreteness (M = 4.87, SD = 1.67; rated on a 7-point scale, 1 = *low imageability* to 7 = *high imageability*). All normed values were taken from Võ et al. (2009). We constructed three study lists of 40 items that were parallel in all word characteristics. For each participant, study lists were randomly assigned to study-test cycles. For each participant, 20 randomly chosen words were presented once for study and the remaining 20 words were presented twice for study. One randomly selected half of the once- and twice presented words were displayed in small 18-point Arial font or in a large 48-point Arial font. The first four items represented each combination of number of study opportunities and font size and were used as buffer items that were not included in the analysis. Items were presented in a new random order for each participant.

Participants

Participants were 160 University of Mannheim undergraduates. When assuming a correlation of .50 between repeated measures, this sample size provides a statistical power of ($1 - \beta$) > .99 for detecting medium-sized (f = .25, equivalent to η_p^2 = .06) main effects of the within-subjects factors and interaction effects with α = .05 in a mixed ANOVA and a statistical power of ($1 - \beta$) = .87 for detecting medium-sized main effects of the between-subjects factor (G*Power 3; Faul et al., 2007). Participants were randomly allocated to the four feedback groups (*n* = 40 in each group). We excluded participants who assigned the same JOL to all items in one or more study phases (*n* = 2) or who had zero recall performance in one or more test phases (*n* = 2). The final sample included 156 participants with a mean age of 23 years (*SD* = 4.10), *n* = 37 in the no-feedback group, *n* = 41 in the recall-feedback group, *n* = 40 in the recall-and-JOL-feedback group, *n* = 38 in the social-reference-feedback group.

Procedure

The experiment consisted of three study-test cycles, each of which included a study phase with JOLs, a distractor task, and a free recall test. Instructions informed participants that in each cycle they would study 40 words and would be asked to recall as many words as they could remember in a memory test. Participants were also told that they would be asked to predict the chance of recalling each word immediately after studying it and that they would have an extra study opportunity for some words before the test. At study, each word appeared on the screen for 4 s. Immediately afterwards, the number of study presentations (1 vs 2) and the JOL prompt *Chance of recall (0-100)?* appeared on the screen, and participants pressed on one of 11 keys labeled 0, 10, …, 90, and 100 to make their JOL. Consequently, we obtained one JOL for once-presented words and two JOLs for twice-presented words. A 200-ms blank screen preceded the presentation of each word. Following a 1-min numerical filler task, participants had 4 min to write down as many studied words as they could remember. At the end of each study-test cycle, participants in the no-feedback group typed examples of one randomly chosen category (i.e., mountains, capitals, or rivers) for 3 minutes. Participants in the recall-feedback group saw an overview of the words they had remembered and had not remembered organized by font size and number of study presentations (see Figure 1). Recall feedback remained on the screen for as long as they wished and for a minimum duration of 45 s. Participants in the recall-and-JOL-feedback group saw the same organized overview of remembered and not-remembered words as the recall-feedback group, however, complemented by their own JOLs from the study phase (see Figure 1). Participants in the social-reference-feedback group received the following information: *This experiment deals with misjudgments of one's own memory performance. One reason for misjudgments is that learners do not know exactly how certain features of the study situation affect memory. In this experiment, you will have the opportunity to learn how 1) the font size during study and 2) the number of study opportunities influence memory.* They were also told that they would see the memory performance of students who participated in the same experiment and learned and remembered the same words as they themselves. They then saw an overview of the average number of remembered words by font size and number of study presentations from other participants from a previous experiment. This overview remained on the screen for as long as participants wished and for a minimum duration of 45 s (see Figure 1). Afterwards, participants were told the following: *Many student participants have overestimated the influence of font size on their memory in this part of the experiment. In addition, many students have underestimated the influence of an additional study opportunity on their memory. Now please consider: Did you maybe make these mistakes in Part 1 (2, 3)?* Immediately prior to the next block, participants from all feedback groups then responded to the question: *What did you learn about your learning and memory from this feedback?*

Results

Effects are considered significant based on an alpha level of .05 and a Greenhouse-Geisser correction was applied when the sphericity assumption was violated.

JOLs

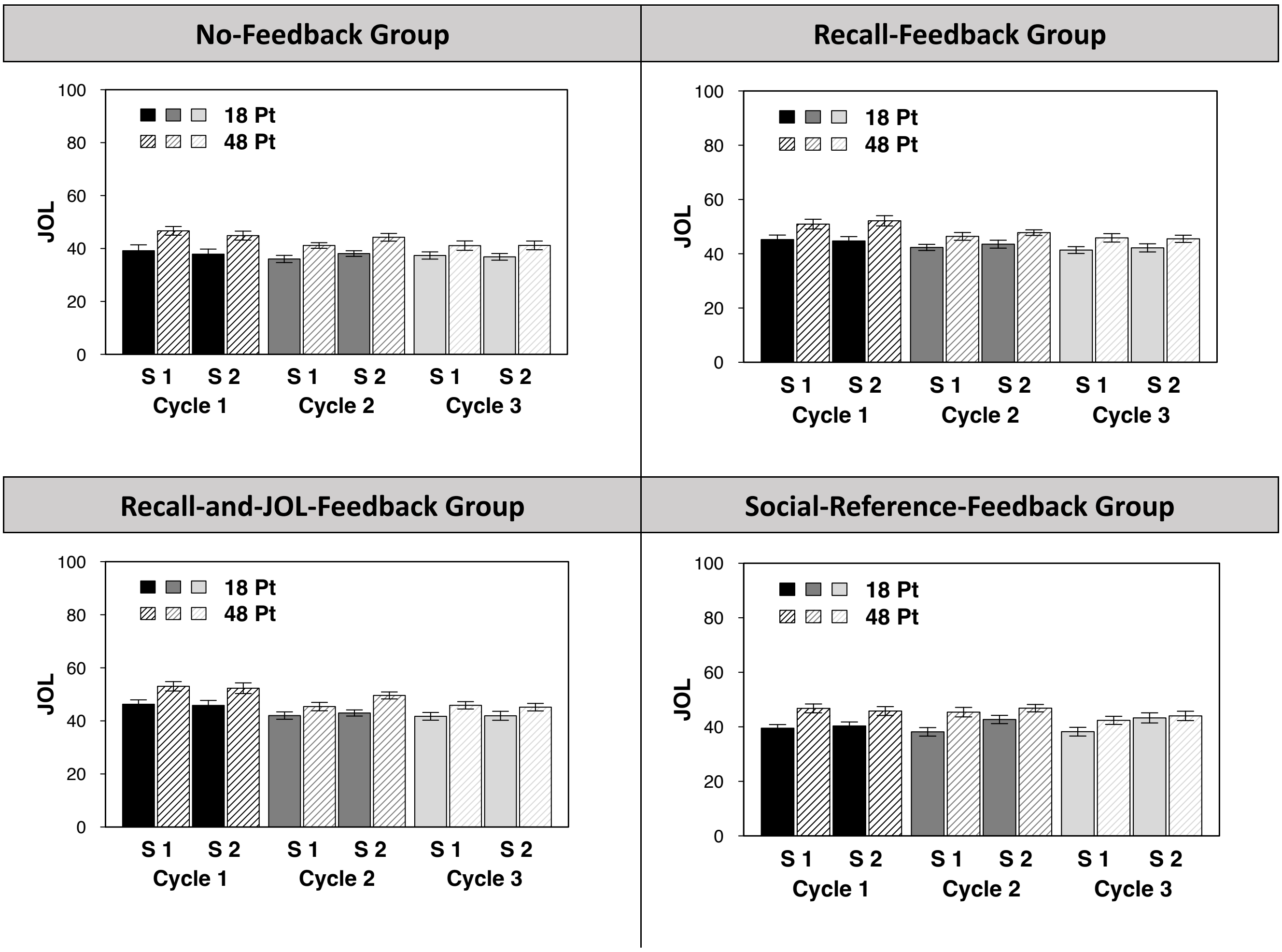
Figure 2 presents mean JOLs in each cycle by font size and number of study presentations in each group of Experiment 1. JOLs from the first study presentation were submitted to a mixed ANOVA with cycle (1, 2, 3), font size (18, 48 point), and number of study presentations (1, 2) as within-subjects factors and feedback group (no-feedback, recall-feedback, recall-and-JOL-feedback, social-reference-feedback) as a between-subjects factor.

A significant main effect of cycle revealed that JOLs decreased across cycles, $F(1.48, 225.22) = 10.33$, $p < .001$, $\eta_p^2 = .06$. Pairwise follow-up *t*-tests showed significant differences among all three cycles, $t(155) >= 2.07$, $p <= .04$, $d_z >= 0.17$. A significant main effect of font size revealed higher JOLs for words displayed in the large font than for words displayed in the small font, $F(1, 152) = 75.02$, $p < .001$, $\eta_p^2 = .33$. A significant main effect of number of study presentations revealed higher JOLs for twice-presented words than for once-presented words, $F(1, 152) = 3.96$, $p = .048$, $\eta_p^2 = .03$. The main effect of feedback group was not significant, $F(3, 152) = 1.31$, $p = .275$, $\eta_p^2 = .02$.

There were significant interactions between cycle and font size, $F(1.82, 276.24) = 6.25$, $p < .01$, $\eta_p^2 = .04$, and between cycle and number of study presentations, $F(1.84, 280.12) = 5.88$, $p < .01$, $\eta_p^2 = .04$. Follow-up *t*-tests indicated that the size of the font size effect on JOLs decreased across cycles, Cycle 1: $t(155) = 7.67$, $p < .001$, $d_z = 0.61$, Cycle 2: $t(155) = 6.78$, $p < .001$, $d_z = 0.54$, Cycle 3: $t(155) = 5.02$, $p < .001$, $d_z = 0.40$, and that JOLs were higher for twice-presented words than for once-presented words in Cycle 2, $t(155) = 3.69$, $p < .001$, $d_z = 0.30$, but not in Cycles 1 or 3, Cycle 1: $t < 1$, Cycle 3: $t(155) = 1.09$, $p = .276$, $d_z = 0.09$. None of the other interactions were significant, $F <= 1.71$, $p >= .168$.

Figure 2

Mean Judgments of Learning (JOL) in Each Cycle for Words Presented Once (S1) or Twice (S2) in a Small (18 pt) or a Large (48 pt) Font Size in Each Group of Experiment 1



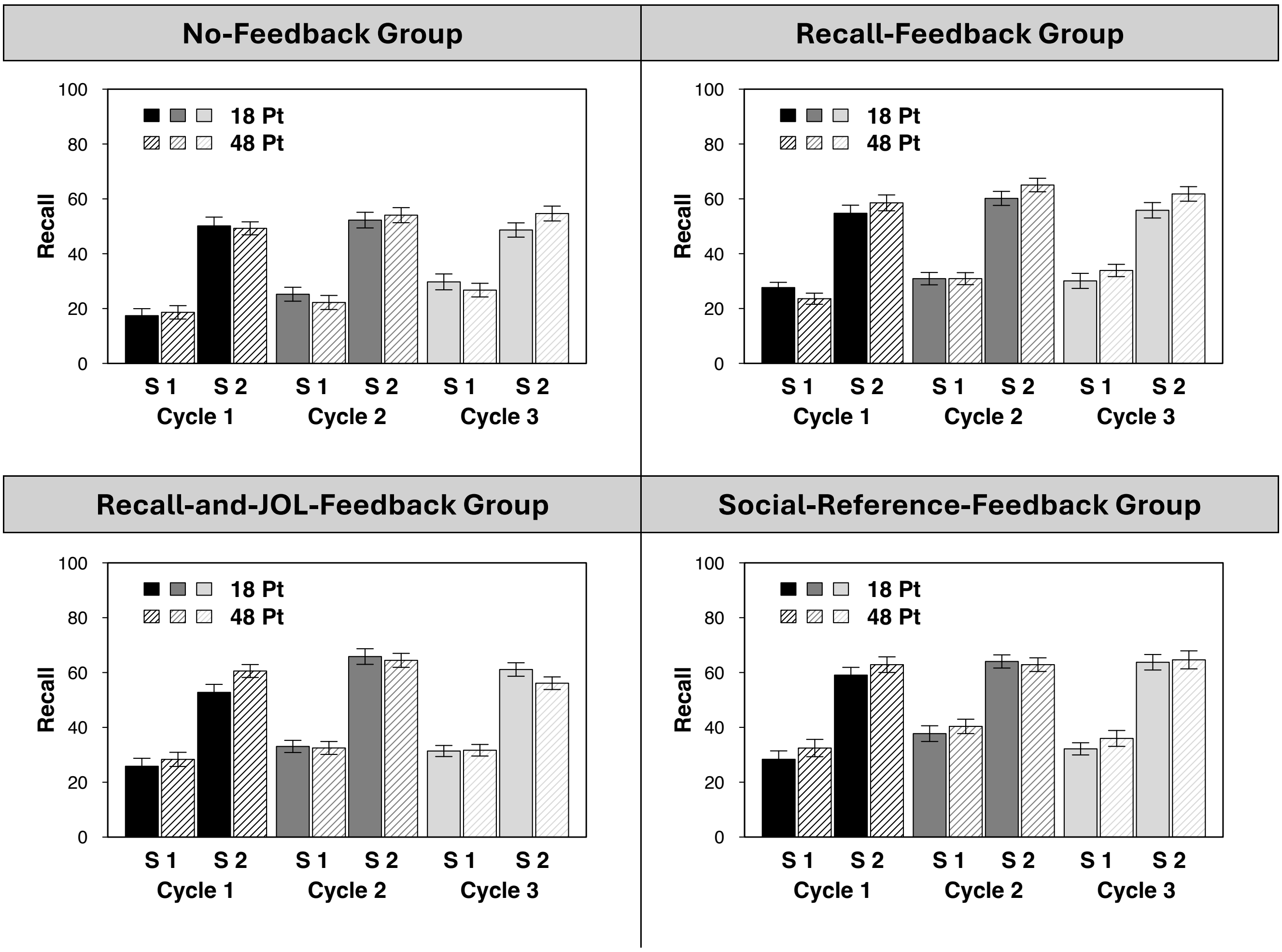
Note. Error bars represent one standard error of the mean.

Recall Performance

Figure 3 presents the mean percentage of recalled words in each cycle by font size and number of study presentations in each group of Experiment 1. A 3 (cycle: 1, 2, 3) x 2 (font size: 18, 48 point) x 2 (number of study presentations: 1,2) x 4 (feedback group: no-feedback, recall-feedback, recall-and-JOL-feedback, social-reference-feedback) mixed ANOVA on recall performance revealed that memory performance varied with cycle, *F*(1.81, 275.36) = 12.85, $p < .001$, $\eta_p^2 = .08$. Follow-up *t* tests showed worse memory performance in Cycle 1 than in Cycles 2 and 3, Cycle 1 vs. 2: $t(155) = 4.90$, $p < .001$, $d_z = 0.39$, Cycle 1 vs. 3: $t(155) = 3.23$, $p < .01$, $d_z = 0.26$, but no differences between Cycles 2 and 3, $t(155) = 1.53$, $p = .128$, $d_z = 0.12$. A significant main effect of font size revealed better memory performance for words displayed in the large font than in the small font, $F(1, 152) = 5.51$, $p = .020$, $\eta_p^2 = .04$. A significant main effect of number of study presentations revealed better memory performance for twice-presented words than for once-presented words, $F(1, 152) = 942.72$, $p < .001$, $\eta_p^2 = .86$. None of the other main effects or interactions were significant *F* <= 2.88, *p* >= .058.

Figure 3

Mean Percentage of Correctly Recalled Words (Recall) in Each Cycle for Words Presented Once (S1) or Twice (S2) in a Small (18 pt) or a Large (48 pt) Font Size in Each Group of Experiment 1



Note. Error bars represent one standard error of the mean.

Resolution and Calibration

Gamma correlations could not be computed for three participants in one cycle due to perfect recall performance.

Table 1 presents Gamma correlations in each cycle for each group of Experiment 1. Gamma correlations between JOLs and recall performance were significantly positive in each study-test cycle and feedback group, $t >= 4.21$, $p < .001$, $d >= 0.69$, indicating that participants from all feedback groups made moderately accurate JOLs in all study-test cycles. A 3 (cycle: 1, 2, 3) x 4 (feedback group: no-feedback, recall-feedback, recall-and-JOL-feedback, social-reference-feedback) mixed ANOVA revealed no main effects of cycle, $F(1.99, 296.38) = 1.26$, $p = .286$, $\eta_p^2 = .01$, or group, $F(3, 149) = 0.90$, $p = .443$, $\eta_p^2 = .02$, and also no interaction, $F(5.97, 296.38) = 1.37$, $p = .228$, $\eta_p^2 = .03$.

A similar mixed ANOVA on calibration revealed that the difference between JOLs and recall performance (i.e., bias) varied with cycle, $F(1.77, 268.83) = 23.57$, $p < .001$, $\eta_p^2 = .13$, with pairwise comparisons indicating a switch from overconfidence in Cycle 1 ($M = 5.12$, $SD = 21.8$) to underconfidence in Cycle 2 ($M = -3.08$, $SD = 20.8$) and Cycle 3 ($M = -2.74$, $SD = 19.1$). Cycle 1 vs. 2: $t(155) = 5.93$, $p < .001$, $d = 0.47$, Cycle 1 vs. 3, $t(157) = 5.20$, $p < .001$, $d = 0.42$, Cycle 2 vs. 3: $t < 1$. Neither the main effect of feedback group, $F(3, 152) = 1.77$, $p = .155$, $\eta_p^2 = .03$, nor the interaction were significant, $F < 1$.

Table 1

Means (SDs) of the Gamma Correlation between JOLs and Recall Performance in Each Cycle and Group of Experiments 1, 2, and 3

Experiment and group	Cycle		
	1	2	3
Experiment 1			
No-feedback	.33 (.23)	.30 (.34)	.26 (.37)
Recall-feedback	.21 (.27)	.36 (.26)	.29 (.28)
Recall-and JOL-feedback	.20 (.27)	.22 (.33)	.25 (.30)
Social-reference-feedback	.29 (.27)	.31 (.29)	.31 (.37)
Experiment 2			
No-feedback	.31 (.32)	.31 (.41)	.41 (.26)
Catch-all-cognitive-feedback	.26 (.26)	.30 (.29)	.23 (.37)
Experiment 3			
No-feedback	.21 (.25)	.36 (.23)	.30 (.28)
Metacognitive-feedback	.28 (.22)	.30 (.31)	.44 (.26)

Discussion

Experiment 1 showed that neither type of feedback improved the cue basis of JOLs or JOL accuracy. Participants continued overweighting font size and underweighting number of study opportunities in their JOLs even with the opportunity to relate individual JOLs to actual memory performance in the recall-and-JOL-feedback group. At the same time, there were no improvements in relative or absolute accuracy. We found, however, that the weight given to font size decreased across cycles in all groups. Since study-test experience is provided to all groups, participants may have learned from experience to rely less on font size. However, this finding was unexpected, and it remains to be seen whether it would replicate. Finally, we found that JOLs switched from overconfidence to underconfidence after the first cycle, a pattern that it is well established for repeated study-test cycles using the same materials repeatedly and known as the underconfidence-with-practice effect (Koriat et al., 2002).

Experiment 2

Since Experiment 1 was not successful at improving JOLs, Experiment 2 combined all forms of feedback into a single ‘catch-all-cognitive-feedback’ group to provide participants with maximum information. Participants in the catch-all-cognitive-feedback group saw a list of the words they had recalled and not recalled accompanied by the JOL they had given to each word during study, organized by the two cues (see recall-and-JOL-feedback group in Figure 1). This was followed by information about each cue’s effectiveness and average performance of other participants in this same task (see social-reference-feedback group in Figure 1). If this extreme catch-all-cognitive-feedback does not improve cue use in JOLs and JOL accuracy, this will demonstrate the resistance of metacognition to change. In Experiment 2, we manipulated font format (standard, aLtErnAtiNg) in addition to font size. Standard-font words typically elicit higher JOLs than alternating-font words, while there are typically no differences in recall performance between the two font formats (Rhodes & Castel, 2008; but see Mueller et al., 2013). We hoped that cue learning would be facilitated by manipulating two perceptual cues.

Method

Design

The design was a mixed design with font size (18 point, 48 point), font format (standard, aLtErnAtiNg), and study-test cycle (1, 2, 3) as within-subjects factors and group (no-feedback, catch-all-cognitive-feedback) as between-subjects factor.

Materials

Stimuli were identical to Experiment 1. The only exception was that for each participant one randomly selected half of the large- and small font words were displayed in standard or aLtErNaTiNg format instead of being presented once or twice for study. The first four items represented each combination of font size and font format and were used as buffer items that were not included in the analysis.

Participants

We aimed at recruiting at least *N* = 80 participants. Power calculations were identical to those reported in Experiment 1. We recruited 43 University of Mannheim undergraduates, 37 of which completed the study in the laboratory and 6 of which completed the study online. Due to the Covid-19 pandemic, we recruited 43 additional participants from the Prolific online pool (https://www.prolific.co). These participants were native-German speakers who were located in Germany, 18 to 35 years old, and mostly students (93.02%). Participants were randomly allocated to the control (*n* = 40), and feedback (*n* = 40) groups. We used the same exclusion criteria as in Experiment 1 and excluded participants who assigned the same JOL to all items in one or more study phases (*n* = 5) or who had zero recall performance in one or more test phases (*n* = 1). The final sample included 80 participants with a mean age of 23.3 years (*SD* = 5.19), *n* = 40 in the control group, and *n* = 40 in the feedback group.

Procedure

The procedure was identical to that of Experiment 1 with the following exceptions. All words were presented only once for study. Participants in the catch-all-cognitive-feedback group received the same information as the social-reference-feedback group in Experiment 1. However, before being presented with the average memory performance of other participants, they additionally were presented with the same feedback as the recall-and-JOL-feedback group in Experiment 1. This feedback included an overview of the words they had remembered and had not remembered with JOLs organized by font size and font format.

Results

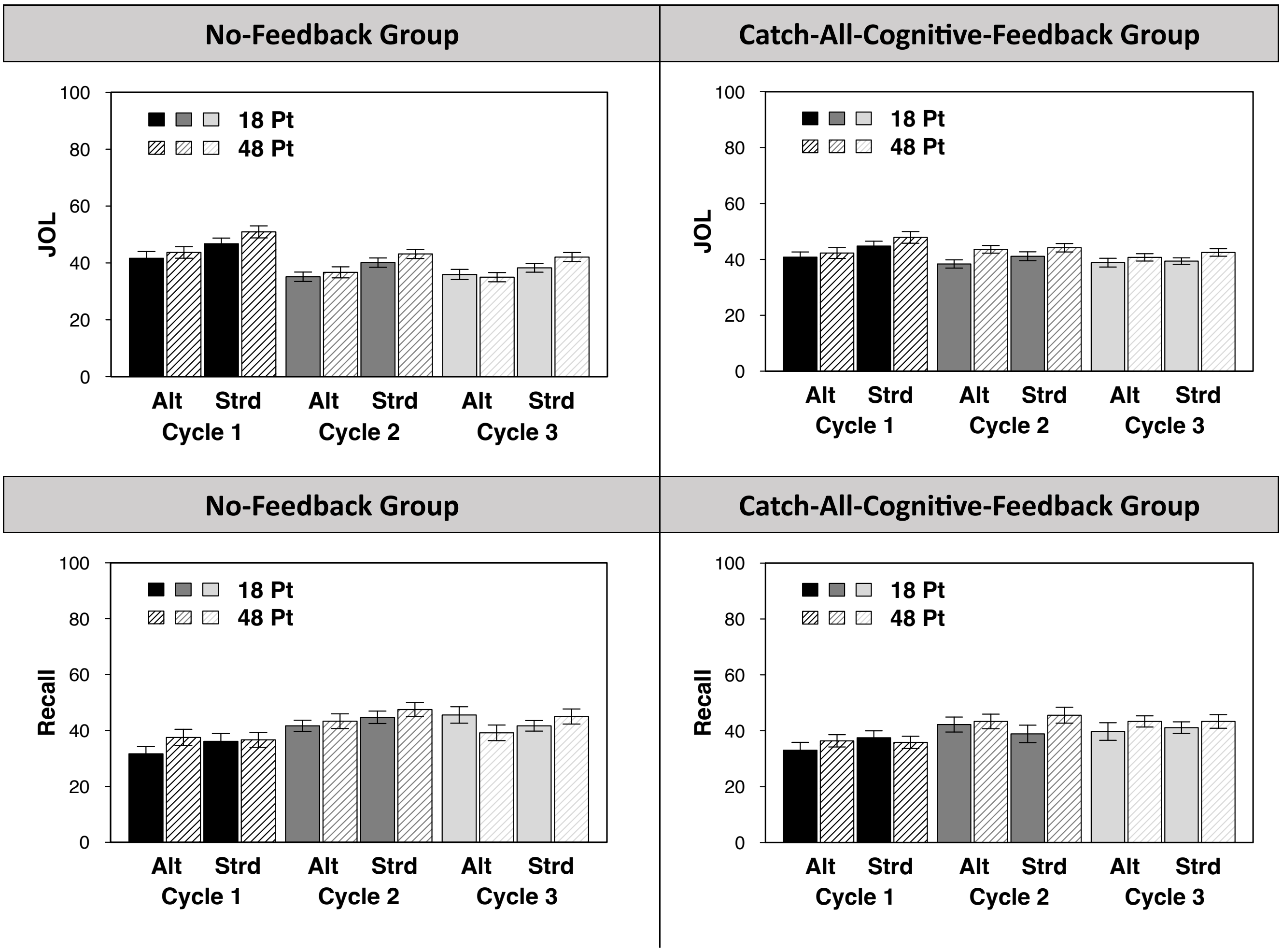
JOLs

Figure 4 presents mean JOLs in each cycle by font format and size in each group of Experiment 2 in the upper row. JOLs were submitted to a mixed ANOVA with cycle (1 vs 2 vs 3), font size (18 vs 48 point), and font format (standard vs aLtErNaTiNg) as within-subjects factors and feedback group (no-feedback, catch-all-cognitive-feedback) as between-subjects factor.

A significant main effect of cycle revealed that JOLs decreased across cycles, *F* (1.42, 110.64) = 9.82, $p < .001$, $\eta_p^2 = .11$. Follow-up *t* tests showed larger JOLs in Cycle 1 than in Cycles 2 and 3, Cycle 1 vs. 2: *t*(79) = 2.95, $p < .01$, $d_z = 0.33$, Cycle 1 vs. 3: *t*(79) = 3.54, $p < .001$, $d = 0.40$, but no differences between Cycles 2 and 3, *t*(79) = 1.47, $p = .146$, $d_z = 0.16$. A significant main effect of font size revealed higher JOLs for words displayed in large font than for words displayed in small font, *F* (1, 78) = 13.78, $p < .001$, $\eta_p^2 = .15$, and a significant main effect of font format revealed higher JOLs for words displayed in standard format than for words displayed in alternating format, F (1, 78) = 23.60, p < .001, η_p^2 = .23. None of the other main effects or interactions were significant, F <= 3.17, p >= .08.

Figure 4

Mean Judgments of Learning (JOL) and Percentage of Correctly Recalled Words (Recall) in Each Cycle for Words Presented in Alternating (Alt) or Standard (Strd) Font and in a Small (18 pt) or a Large (48 pt) Font Size in Each Group of Experiment 2



Note. Error bars represent one standard error of the mean.

Recall Performance

Figure 4 (lower row) presents the mean percentage of recalled words in each cycle by font format and font size in each group of Experiment 2. A 3 (cycle: 1, 2, 3) x 2 (font size: 18, 48 point) x 2 (font format: standard, aLtErNaTiNg) x 4 (feedback group: no-feedback, catch-all-cognitive-feedback) mixed ANOVA on recall performance revealed that memory performance varied with cycle, F (1.88, 146.85) = 19.35, $p < .001$, $\eta_p^2 = .20$. Follow-up t tests showed worse memory performance in Cycle 1 than in Cycles 2 and 3, Cycle 1 vs. 2: $t(79) = 5.29$, $p < .001$, $d_z = 0.59$, Cycle 1 vs. 3: $t(79) = 4.84$, $p < .001$, $d_z = 0.54$, but did not differ between Cycles 2 and 3, $t < 1$. Neither the main effect of font size, $F(1, 78) = 2.98$, $p = .09$, $\eta_p^2 = .04$, nor the main effect of font format were significant, $F(1, 78) = 1.84$, $p = .18$, $\eta_p^2 = .02$. None of the other main effects or interactions were significant, $F <= 2.18$, $p >= .12$.

Resolution and Calibration

Table 1 presents Gamma correlations in each cycle for each group of Experiment 2. As in Experiment 1, all Gamma correlations by group and cycle were significantly positive $t >= 4.03$, $p < .001$, $d >= .64$, indicating that participants from both groups made moderately accurate JOLs in all study-test cycles. A 3 (cycle: 1, 2, 3) x 2 (group: no-feedback, catch-all-cognitive-feedback) mixed ANOVA revealed no main effects of cycle, $F < 1$, or group, $F(1,78) = 2.45$, $p = .121$, $\eta_p^2 = .03$, and also no interaction, $F(1.98, 154.76) = 1.85$, $p = .161$, $\eta_p^2 = .02$.

A similar mixed ANOVA on calibration revealed that bias varied with cycle, $F(1.64, 127.67) = 30.13$, $p < .001$, $\eta_p^2 = .28$, with pairwise comparisons indicating a switch from overconfidence in Cycle 1 ($M = 9.25$, $SD = 23.7$) to underconfidence in Cycle 2 ($M = -3.11$, $SD = 18.2$) and Cycle 3 ($M = -3.27$, $SD = 17.5$). Cycle 1 vs. Cycle 2: $t(79) = 5.97$, $p < .001$, $d_z = 0.67$, Cycle 1 vs. 3: $t(79) = 6.02$, $p < .001$, $d_z = 0.67$, Cycle 2 vs. 3: $t < 1$. Neither the main effect of group, $F < 1$, nor the interaction were significant, $F(1.64, 127.67) = 1.94, p = .156$, $\eta_p^2 = .02$.

Discussion

Experiment 2 results showed no improvements in cue use or JOL accuracy in the catch-all-cognitive-feedback group. Participants in this group continued overweighting font size and font format despite receiving maximum information. Unlike Experiment 1, this experiment showed that the illusory effect of font size was stable across cycles. Relative accuracy did not improve across cycles, and JOLs switched again from overconfidence to underconfidence after Cycle 1. Overall, this demonstrates that metamemory judgments are resistant to change. This contrasts with the improvement in judgments about external criteria found when participants receive cognitive feedback (Balzer et al., 1989; Karlsson et al., 2004; Little & Lewandowsky, 2009; Newell et al., 2009).

One reason why participants do not change the cue basis of their JOLs can be that experiential cues dominate during learning (e.g., fluency). However, this is at odds with evidence showing that people can base their JOLs on multiple cues rather than on a unified feeling of ease (Undorf et al., 2018; Undorf & Bröder, 2020). Further, the font size effect on JOLs has been mostly explained by beliefs rather than fluency (Luna et al., 2019; Mueller et al., 2014; Undorf & Zimdahl, 2019). Yan et al. (2016) argue that metacognitive judgments are hard to change because of a) pre-existing beliefs about learning and memory, b) experiences of fluency during learning, and c) the belief of being unique as a learner. Thus, it might be that participants' pre-existing beliefs influence how they interpret and store the cognitive feedback provided. In some situations, participants might even disregard the cognitive feedback because they consider themselves experts on their own cognitions or think that the average performance of other participants is irrelevant for them as unique learners, as suggested by Yan et al. (2016). Experiment 3 aims to dismantle possible erroneous metacognitive beliefs.

Experiment 3

In Experiment 3, we designed a new form of feedback to remedy the possibility that participants might be misperceiving the cognitive feedback due to pre-existing beliefs. For this, we followed Fiedler et al.'s (2020) recommendations of effective forms of feedback. These recommendations were made by 10 scientists and state that an effective debiasing treatment should not only provide information about judgments that are "correct" versus "incorrect", but to also relate this feedback to a) the representation of the stimuli, and b) explicit instructions about how to make accurate judgments. Based on these recommendations, we designed an informative 'metacognitive-feedback' in which we adopted a first-person perspective at explaining the metacognitions that likely occur during learning (e.g., *large-font words stand out more and therefore must be more memorable*), their biased nature, and instructions about which cues to consider when making judgments. Our aim was to specifically tackle participants' metacognitions and dismantle any erroneous pre-existing beliefs.

Further, Experiment 3 aimed to ensure that participants fully attended the cognitive feedback and achieved a deep understanding of cue validities and biased metacognitions during learning. To achieve this goal, we used a similar approach as Pan and Rivers (2023) and asked participants to describe how each of the cues affected their memory and their JOLs after receiving feedback.

Method

Design

The design was a mixed design with font size (18 point, 48 point), number of study presentations (1,2), and study-test cycle (1, 2, 3) as within-subjects factors and group (no-feedback, metacognitive-feedback) as between-subjects factor.

Materials

Stimuli were identical to Experiment 1.

Participants

We aimed at recruiting at least *N* = 80 participants. Power calculations were identical to those reported in Experiment 1. We recruited 23 University of Mannheim and 57 Technical University of Darmstadt undergraduates. Participants were randomly allocated to the no-feedback (*n* = 40), and metacognitive-feedback (*n* = 40) groups. We used the same exclusion criteria as in Experiment 1 and excluded participants who assigned the same JOL to all items in one or more study phases (*n* =1) or who had zero recall performance in one or more test phases (*n* = 0). Additionally, we excluded incomplete data due to a technical PC error (*n* = 2). The final sample included 77 participants with a mean age of 21.92 years (*SD* = 2.92), *n* = 39 in the control group, and *n* = 38 in the feedback group.

Procedure

The procedure was identical to that of Experiment 2 with the following exceptions. All words were in a standard format for study. For each participant, one randomly selected half of the small-font (18-pt Arial font) and large-font (48-pt Arial font) words were presented once or twice for study. To ensure that participants understood the feedback screen that presented an overview of the words they had remembered and had not remembered with JOLs, they were asked two questions about their recall performance and memory predictions; Question 1: "*Which statement best describes your actual memory performance in Part 1 [2, 3]? In the test, I was able to…"* Answer Option 1: *remember more words learned twice [in large font size] than words learned once [in small font size].* Answer Option 2: *remember* *fewer words learned twice [in large font size] than words learned once [in small font size].* Answer Option 3: *remember the same number of words learned twice [in large font size] and once [in small font size].* Question 2: "*Which statement best describes your memory predictions in Part 1 [2, 3]?"* Answer Option 1: *I underestimated the impact of an additional learning opportunity [of a large font size] on my memory.* Answer Option 2: *I correctly assessed the influence of an additional learning opportunity [of a large font size] on my memory.* Answer Option 3: *I overestimated the influence of an additional learning opportunity [of large font size] on my memory.* After being presented with the average memory performance of other participants, participants read textual information about the font size illusion and the stability bias. The text informed about possible perceptions during learning, the validities of the cues, font size and number of study opportunities, and recommendations on which factors to focus when making JOLs (see Appendix).

Results

JOLs

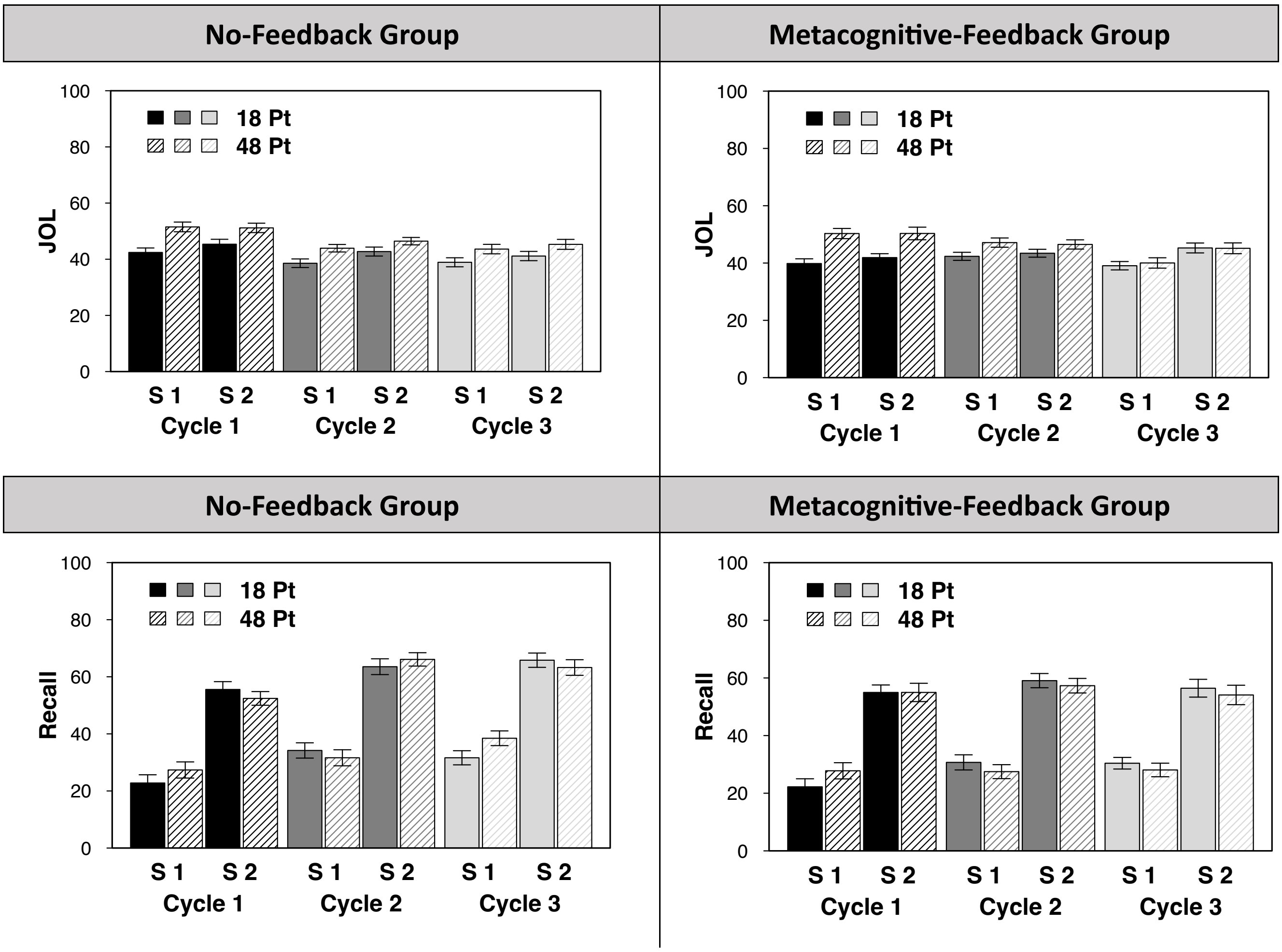
Figure 5 presents mean JOLs in each cycle by font size and number of study presentations in each group of Experiment 3. JOLs from the first study presentation were submitted to a mixed ANOVA with cycle (1, 2, 3), font size (18, 48 point), and number of study presentations (1, 2) as within-subjects factors and feedback group (no-feedback, metacognitive-feedback) as a between-subjects factor.

A significant main effect of cycle revealed differences in JOLs across cycles, $F(1.57, 117.99) = 6.90$, $p < .01$, $\eta_p^2 = .08$. Specifically, JOLs were lower in Cycles 2 and 3 than in Cycle 1, Cycle 1 vs. Cycle 2: $t(76) = 2.51$, $p = .01$, $d_z = 0.29$, Cycle 1 vs. 3: $t(76) = 3.01$, $p < .01$, $d_z = 0.34$, but did not differ between Cycle 2 and 3, $t(76) = 1.66$, $p = .10$, $d_z = 0.19$. A significant main effect of font size revealed higher JOLs for words displayed in the large font than for words displayed in the small font, $F(1, 75) = 39.05$, $p < .001$, $\eta_p^2 = .34$. A significant main effect of number of study presentations revealed higher JOLs for twice-presented words than for once-presented words, $F(1, 75) = 8.40, p < .01, \eta_p^2 = .10$. The main effect of group was not significant, $F < 1$.

As in Experiment 1, there was a significant interaction between cycle and font size, $F(1.68, 125.65) = 10.06, p < .001, \eta_p^2 = .12$. Follow-up *t* tests indicated that the size of the font size effect on JOLs decreased across cycles, Cycle 1: $t(76) = 6.47, p < .001, d_z = 0.74$, Cycle 2: $t(76) = 4.20, p < .001, d_z = 0.48$, Cycle 3: $t(76) = 2.29, p = .02, d_z = 0.26$. Importantly, the three-way interaction between group and cycle and number of study presentations was significant, $F(1.99, 149.17) = 3.07, p = .0495, \eta_p^2 = .04$. Follow-up analyses showed that the interaction between cycle and number of study presentations was not significant in the control group, $F < 1$, but was significant in the metacognitive-feedback group, $F(1.93, 71.48) = 4.07, p = .02, \eta_p^2 = .10$, with *t*-tests showing that JOLs were higher for twice- than for once-studied words in Cycle 3, $t(37) = 2.91, p < .01, d_z = 0.47$, but not in Cycle 1 and 2, Cycle 1: $t < 1$, Cycle 2: $t < 1$. The three-way interaction between group, cycle and font size was not significant, $F(1.68, 125.65) = 2.39, p = .10, \eta_p^2 = .03$. None of the other interactions were significant, $F <= 3.69, p >= .06$.

Figure 5

Mean Judgments of Learning (JOL) and Percentage of Correctly Recalled Words (Recall) in Each Cycle for Words Presented Once (S1) or Twice (S2) in a Small (18 pt) or a Large (48 pt) Font Size in Each Group of Experiment 3



Note. Error bars represent one standard error of the mean.

Recall Performance

Figure 5 presents the mean percentage of recalled words in each cycle by font size and number of study presentations in each group of Experiment 3. A 3 (cycle: 1, 2, 3) x 2 (font size: 18, 48 point) x 2 (number of study presentations: 1,2) x 2 (no-feedback, metacognitive-feedback) mixed ANOVA on recall performance revealed that memory performance varied with cycle, $F(1.91, 143.50) = 8.84$, $p < .001$, $\eta_p^2 = .11$. Follow-up t tests showed worse memory performance in Cycle 1 than in Cycles 2 and 3, Cycle 1 vs. 2: $t(76) = 3.59$, $p < .001$, $d_z = 0.41$, Cycle 1 vs. 3: $t(76) = 3.29$, $p < .01$, $d_z = 0.38$, but no differences between Cycles 2 and 3, $t < 1$. A significant main effect of number of study presentations revealed better memory performance for twice-presented words than for once-presented words, $F(1, 75) = 666.88$, $p < .001$, $\eta_p^2 = .90$. None of the other main effects or interactions were significant $F <= 2.62$, $p >= .080$.

Resolution and Calibration

Table 1 presents Gamma correlations in each cycle for each group of Experiment 3. As in Experiments 1 and 2, all Gamma correlations by group and cycle were significantly positive, $t >= 5.24$, $p < .001$, $d >= .84$, indicating that participants from both groups made moderately accurate JOLs in all study-test cycles. A 3 (cycle: 1, 2, 3) x 2 (group: no-feedback, metacognitive-feedback) mixed ANOVA revealed a main effect of cycle, $F(1.94, 145.76) = 5.51$, $p < .01$, $\eta_p^2 = .07$, no main effect of group, $F(1,75) = 1.46$, $p = .230$, $\eta_p^2 = .02$, and a significant interaction, $F(1.94, 145.76) = 3.57$, $p = .032$, $\eta_p^2 = .05$. In the metacognitive-feedback group, resolution was markedly improved in Cycle 3, but it did not differ between Cycle 1 and 2, Cycle 1 vs. 2: $t < 1$, Cycle 1 vs. 3: $t(37) = 2.84$, $p < .001$, $d_z = 0.46$, Cycle 2 vs. 3: $t(37) = 2.45$, $p = .019$, $d_z = 0.40$. In contrast, in the control group, resolution improved only between Cycle 1 and 2, $t(38) = 2.90$, $p < .01$, $d_z = 0.46$, but it remained similar in Cycle 1 and

3 and Cycle 2 and 3, Cycle 1 vs. 3: $t(38) = 1.51$, $p = .140$, $d_z = 0.24$, Cycle 2 vs. 3: $t(38) = 1.23$, $p = .224$, $d_z = 0.20$.

A similar mixed ANOVA on calibration revealed a main effect of cycle, $F(1.82, 136.34) = 17.69$, $p < .001$, $\eta_p^2 = .19$, no main effect of group, $F(1, 75) = 1.13$, $p = .290$, $\eta_p^2 = .02$, and a significant interaction, $F(1.82, 136.34) = 4.38$, $p = .017$, $\eta_p^2 = .06$. In the control group, there was a switch from overconfidence in Cycle 1 ($M = 8.08$, $SD = 21.3$) to underconfidence in Cycle 2 ($M = -5.94$, $SD = 16.5$) and 3 ($M = -7.54$, $SD = 18.7$). Cycle 1 vs. 2: $t(38) = 4.82$, $p < .001$, $d_z = 0.77$, Cycle 1 vs. 3: $t(38) = 4.87$, $p < .001$, $d_z = 0.78$, Cycle 2 vs. 3: $t < 1$. In the metacognitive-feedback group, calibration did not differ across cycles, but there was a trend towards a reduction in bias across cycles (Cycle 1: $M = 5.60$, $SD = 16.1$, Cycle 2: $M = 1.21$, $SD = 22.7$, Cycle 3: $M = 0.12$, $SD = 21.2$), $t(37) >= 1.99$, $p >= .053$, $d_z <= 0.32$.

Questionnaire Data

Regarding the number of study opportunities, most of the 38 participants in the feedback group reported remembering more twice-learned than once-learned words; Cycle 1: 78.95% (with 73.68% out of the 38 participants being correct), Cycle 2: 81.58% (73.68% were correct), Cycle 3: 76.31% (57.89% were correct). The remaining participants either said that they remembered more words learned once than twice; Cycle 1: 13.16% (all were incorrect), Cycle 2: 7.89% (2.63% were correct), Cycle 3: 13.16% (2.63% were correct), or that they remembered the same number of words learned twice and once (2.63% were on average correct across cycles). The percentage of participants who indicated that they had accurately predicted the effect of number of study repetitions in their JOLs increased from around one third in Cycle 1 to half in Cycle 2 and 3; Cycle 1: 31.58% (18.42% were correct), Cycle 2: 55.26% (26.32% were correct), Cycle 3: 52.63% (28.95% were correct). The remaining participants said that they had underestimated the impact of a second repetition on their memory; Cycle 1: 42.10%

(26.32% were correct), Cycle 2: 23.68% (21.05% were correct), Cycle 3: 31.58% (10.53% were correct), or that they had overestimated it (11.32% were on average correct across cycles).

Regarding the font size of words, the percentage of participants who reported remembering more large-font words than small-font words decreased after Cycle 1; Cycle 1: 47.37% (42.10% were correct), Cycle 2: 21.05% (13.16% were correct), Cycle 3: 28.95% (15.79% were correct), with just below half of the participants reporting equal memory for large and small words in Cycles 2 and 3; Cycle 1: 36.84% (7.89% were correct), Cycle 2: 47.37% (7.89% were correct), Cycle 3: 44.74% (7.89% were correct), and a minority saying that memory was better for small-font than large-font words (18.42% were on average correct across cycle). Regarding the accuracy of their JOLs, more than half of the participants reported that they correctly predicted the impact of font size on memory after Cycle 1: Cycle 1: 42.10% (15.79% were correct), Cycle 2: 63.16% (34.21% were correct), Cycle 3: 57.89% (36.84% were correct). The other participants said that they had underestimated the impact of font size decreased after Cycle 1: Cycle 1: 34.21% (2.63% were correct), Cycle 2: 13.16% (5.26% were correct), Cycle 3: 15.79% (5.26% were correct), or that they had overestimated it (15.94% were on average correct across cycles).

Discussion

Experiment 3 results showed that metacognitive feedback informing not only about cognitive processes but also about how metacognition operates and how metacognition can be improved was successful at increasing JOL reliance on valid cues. Specifically, the metacognitive feedback was effective at increasing the reliance of JOLs on number of study opportunities, improving resolution, and, descriptively, improving calibration. Participants learned that number of study opportunities is a valid predictor of their memory and therefore relied on this cue for their predictions in Cycle 3. At the same time, improved resolution demonstrated that relying on valid cues fostered JOL accuracy. Also, in the metacognitive-feedback group, JOLs were well-calibrated after Cycle 1. The questionnaire data clearly shows that most participants correctly identified better memory performance for words learned twice than once, and the percentage of participants who accurately reported correctly estimating the number of study opportunities cue increased after Cycle 1 to nearly half. Further, the interaction between font size and cycle from Experiment 1 was again obtained, unlike in Experiment 2, in which the invalid cue font format was manipulated as a second cue. It might be that a valid cue needs to be present in the environment for participants to learn to rely less on font size from experience. However, this is only a speculation that would need further testing. Although font size descriptively affected JOLs less strongly in Cycle 2 and even less so in Cycle 3 in the metacognitive-feedback group but not in the no-feedback group, the three-way interaction was not significant. This might be because JOLs are easier to correct for those cues with robust predictive validity, and not for those without validity. Furthermore, it could be that individual differences in the effect of font size on memory performance (i.e., some participants showing better recall for large than small words) did not enable that all participants updated their JOLs which was reflected in the non-significant three-way interaction.

General discussion

This study tested the effectiveness of different forms of feedback for improving the cue basis and accuracy of JOLs in the context of metacognitive illusions. In all three experiments, feedback and multiple study-test-cycles for different lists were provided to remedy the font size illusion (Rhodes & Castel, 2008). Additionally, in Experiment 1 and 3, feedback was directed to remedy the stability bias (Kornell & Bjork, 2009), and, in Experiment 2, the font format illusion (Rhodes & Castel, 2008). We found that cognitive feedback presenting individual task performance (recall only, recall and JOL) for each studied item and/or aggregated recall performance from previous participants was not effective at correcting either illusion (Experiments 1 and 2). In contrast, additional *metacognitive* feedback informing participants about possible metacognitions during the task, their biased nature, and ways to enhance the accuracy of their JOLs was effective at remedying the stability bias and improving the relative accuracy of JOLs (Experiment 3). This study thus shows that people do not learn to appropriately use cues from mere task experience and/or cognitive feedback on their own memory and JOLs, but rather that information about how metacognition operates and how it may be improved is effective for mending metacognitive illusions.

The finding that metacognitive feedback is effective is in line with studies showing that detailed information about metacognition improve metacognitive judgments (Koriat & Bjork, 2006b; Miller & Geraci, 2011). Importantly, however, this is the first demonstration that information about metacognition is effective even when provided as part of computerized experiment instructions rather than in personal discussions with an experimenter.

Further, this is the first study to show that providing participants with the JOL and recall status of each studied item organized by cues is not effective for improving the cue basis of JOLs in a subsequent cycle. This contrasts with studies about external world criteria (Karlsson et al., 2004; Little & Lewandowsky, 2009; Newell et al., 2009; Seong & Bisantz, 2008). It also disconfirms the *inferential deficit hypothesis* in metamemory research according to which JOLs do not improve across cycles because there is a failure to monitor test performance and make correct cue attributions (Dunlosky & Hertzog, 2000; Matvey et al., 2002). Experiment 3 showed that most participants can correctly understand the feedback on how the number of study opportunities and, to a somewhat lesser extent, how font size affect recall performance. So, one reason for why the illusions were not remedied by feedback in Experiments 1 and 2 might be that participants failed to apply the declarative knowledge gained about the cue effects on recall performance when making JOLs for a new study list (see Dunlosky & Hertzog, 2000). This might be because beliefs must be activated to impact JOLs (Ariel et al., 2014; Undorf & Erdfelder, 2015). Another reason might be that other experiential cues during learning such as fluency or idiosyncratic cues overshadow the cue knowledge recently acquired (Koriat & Ackerman, 2010; Undorf, et al. 2022; Undorf & Erdfelder, 2011). However, the finding that JOLs are based on multiple cues instead of a unified feeling of ease speak against the latter possibility (Undorf & Bröder, 2020). A more compelling reason presumably is that more in depth-knowledge about metacognition than acquired when receiving cognitive feedback about one's JOLs and recall performance is needed to improve the judgment cue basis. This is demonstrated by our finding that JOLs relied on the number of study opportunities after metacognitive feedback in Experiment 3. We think that the declarative knowledge must be more substantial to be correctly applied when making JOLs.

Our finding that metamemory feedback is effective may rise the question why warnings are not (Kornell & Bjork, 2009; Rhodes & Castel, 2008; Yan et al., 2016). The crucial difference between the metacognitive feedback used in this study and the warnings used in other studies is that only metacognitive feedback provides information about how metacognitive illusions arise from a first-person perspective (e.g., by explaining the impressions during learning that might bias metacognition). In contrast, typical warnings only provide information about the normative correctness of metacognitive judgments but do not entail any information about the experiences during learning or metacognitive beliefs that may bias metacognition (e.g., Kornell & Bjork, 2009; Rhodes & Castel, 2008).

This study is not without limitations. First, while metacognitive feedback reduced the stability bias, it did not alleviate the font size illusion. One possible reason for this is that because the font size illusion decreased across study-test cycles in all conditions, additional beneficial effects of metacognitive feedback were harder to substantiate, and doing so would have required more statistical power. Alternatively, it might be that decreasing the reliance on invalid cues is more resistant to metacognitive feedback than increasing the reliance on valid cues. This is because experiences of high fluency produced by invalid cues may still contribute despite metacognitive beliefs being correct. Second, the improvement in the cue basis of JOLs was observed only in the third cycle whereas the cue basis of JOLs remained unaffected in the second cycle. One would need to replicate this finding to know whether it is robust. If so, it might be related to participants wanting to verify that the metacognitive feedback is correct. Hence, presenting the metacognitive feedback immediately before or after the cognitive feedback would lead to improved JOLs. This is because participants can then confirm that the information in the metacognitive feedback indeed aligns with their task performance in the first cycle.

In conclusion, this study shows that cognitive feedback on memory performance and JOLs on the item level is not sufficient for improving the cue basis and resolution of JOLs. In contrast, metacognitive feedback that provides an in-depth explanation of biased metacognitions and how to evade their detrimental impact is effective for correcting the cue basis of JOLs and, in turn, improves JOL resolution. This is a very promising direction to mending metacognitive illusions, which has proved very challenging in many prior studies.

Acknowledgments

This work was supported by grants UN 345/2-1 (Monika Undorf) and BR 2130/14-1 (Arndt Bröder) from the Deutsche Forschungsgemeinschaft and by the University of Mannheim's Graduate School of Economic and Social Sciences (Sofia Navarro-Báez). We thank Elena Huber, Maike Klingemann, Leila Plohmann, Franziska Schäfer, Charlotte Smid, Anna Fee Wefelmeier, and Clara Wiebel for help with data collection.

References

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition*, *39*(1), 171–184. https://doi.org/10.3758/s13421-010-0002-y

Ariel, R., Hines, J. C., & Hertzog, C. (2014). Test framing generates a stability bias for predictions of learning by causing people to discount their learning beliefs. *Journal of Memory and Language*, *75*, 181–198. https://doi.org/10.1016/j.jml.2014.06.003

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of Cognitive Feedback on Performance. *Psychological Bulletin*, *106*(3), 410–433.

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica*, *45*(1–3), 223–241. https://doi.org/10.1016/0001-6918(80)90034-7

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, *36*(2), 429–437. https://doi.org/10.3758/MC.36.2.429

Chang, M., & Brainerd, C. J. (2022). Association and dissociation between judgments of learning and memory: A Meta-analysis of the font size effect. *Metacognition and Learning*, *17*(2), 443–476. https://doi.org/10.1007/s11409-021-09287-3

Dunlosky, J., & Hertzog, C. (2000). Updating Knowledge About Encoding Strategies: A Componential Analysis of Learning About Strategy Effectiveness From Task Experience. *Psychology and Aging*, *15*(3), 462–474.

Dunlosky, J., Mueller, M. L., Morehead, K., Tauber, S. K., Thiede, K. W., & Metcalfe, J. (2021). Why Does Excellent Monitoring Accuracy Not Always Produce Gains in Memory Performance? *Zeitschrift Für Psychologie*, *229*(2), 104–119. https://doi.org/10.1027/2151-2604/a000441

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

Fiedler, K., Schott, M., Kareev, Y., Avrahami, J., Ackerman, R., Goldsmith, M., Mata, A., Ferreira, M. B., Newell, B. R., & Pantazi, M. (2020). Metacognitive myopia in change detection: A collective approach to overcome a persistent anomaly. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 649–668. https://doi.org/10.1037/xlm0000751

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*(1), 19–34. https://doi.org/10.1016/j.jml.2007.03.006

Hertzog, C., Price, J., Burpee, A., Frentzel, W. J., Feldstein, S., & Dunlosky, J. (2009). Why do people show minimal knowledge updating with task experience: Inferential deficit or experimental artifact? *Quarterly Journal of Experimental Psychology*, *62*(1), 155–173. https://doi.org/10.1080/17470210701855520

Karlsson, L., Juslin, P., & Olsson, H. (2004). Representational Shifts in a Multiple-Cue Judgment Task with Continuous Cues. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*(26), 648–653.

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

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for Self and Other during self-paced study. *Consciousness and Cognition*, *19*(1), 251–264. https://doi.org/10.1016/j.concog.2009.12.010

Koriat, A., & Bjork, R. A. (2006a). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959–972. https://doi.org/10.3758/BF03193244

Koriat, A., & Bjork, R. A. (2006b). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1133–1145. https://doi.org/10.1037/0278-7393.32.5.1133

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing Objective and Subjective Learning Curves: Judgments of Learning Exhibit Increased Underconfidence With Practice. *Journal of Experimental Psychology: General*, *131*(2), 147–162.

Kornell, N., & Bjork, R. A. (2009). A Stability Bias in Human Memory: Overestimating Remembering and Underestimating Learning. *Journal of Experimental Psychology: General*, *138*(4), 449–468. https://doi.org/10.1037/a0017350

Little, D. R., & Lewandowsky, S. (2009). Better learning with more error: Probabilistic feedback increases sensitivity to correlated cues in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1041–1061. https://doi.org/10.1037/a0015902

Luna, K., Martín-Luengo, B., & Albuquerque, P. B. (2018). Do delayed judgements of learning reduce metamemory illusions? A meta-analysis. *Quarterly Journal of Experimental Psychology*, *71*(7), 1626–1636.

Luna, K., Nogueira, M., & Albuquerque, P. B. (2019). Words in larger font are perceived as more important: Explaining the belief that font size affects memory. *Memory*, *27*(4), 555–560. https://doi.org/10.1080/09658211.2018.1529797

Matvey, G., Dunlosky, J., Shaw, R. J., Parks, C., & Hertzog, C. (2002). Age-related equivalence and deficit in knowledge updating of cue effectiveness. *Psychology and Aging*, *17*(4), 589–597. https://doi.org/10.1037/0882-7974.17.4.589

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning*, *6*(3), 303–314. https://doi.org/10.1007/s11409-011-9083-7

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2015). Why is knowledge updating after task experience incomplete? Contributions of encoding experience, scaling artifact, and inferential deficit. *Memory & Cognition*, *43*(2), 180–192. https://doi.org/10.3758/s13421-014-0474-2

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, *20*(2), 378–384. https://doi.org/10.3758/s13423-012-0343-6

Nelson, T. O., & Narens. (1990). Metamemory: A Theoretical Framework and New Findings. In *Psychology of Learning and Motivation* (Vol. 26, pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5

Newell, B. R., Weston, N. J., Tunney, R. J., & Shanks, D. R. (2009). The Effectiveness of Feedback in Multiple-Cue Probability Learning. *Quarterly Journal of Experimental Psychology*, *62*(5), 890–908. https://doi.org/10.1080/17470210802351411

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance.

Metacognition and Learning, *13*(2), 179–212. https://doi.org/10.1007/s11409-018-9183-8

Pan, S. C., & Rivers, M. L. (2023). Metacognitive awareness of the pretesting effect improves with self-regulation support. *Memory & Cognition*, *51*(6), 1461–1480. https://doi.org/10.3758/s13421-022-01392-1

Rhodes, M. G. (2016). *Judgments of Learning: Methods, Data, and Theory* (J. Dunlosky & S. (Uma) K. Tauber, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.4

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615–625. https://doi.org/10.1037/a0013684

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, *38*(7–8), 608–625. https://doi.org/10.1016/j.ergon.2008.01.007

Smithson, C. J. R., Eichbaum, Q. G., & Gauthier, I. (2023). Object recognition ability predicts category learning with medical images. *Cognitive Research: Principles and Implications*, *8*(1), 9. https://doi.org/10.1186/s41235-022-00456-9

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin & Review*, *18*(5), 973–978. https://doi.org/10.3758/s13423-011-0114-9

Tauber, S. K., & Rhodes, M. G. (2010). Metacognitive errors contribute to the difficulty in remembering proper names. *Memory*, *18*(5), 522–532. https://doi.org/10.1080/09658211.2010.481818

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging*, *27*(2), 474–483. https://doi.org/10.1037/a0025246

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Tullis, J. G., & Benjamin, A. S. (2012). Consequences of restudy choices in younger and older learners. *Psychonomic Bulletin & Review*, *19*(4), 743–749. https://doi.org/10.3758/s13423-012-0266-2

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429–442. https://doi.org/10.3758/s13421-012-0274-5

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629–642. https://doi.org/10.1177/1747021819882308

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264–1269. https://doi.org/10.1037/a0023719

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition*, *43*(4), 647–658. https://doi.org/10.3758/s13421-014-0479-x

Undorf, M., Navarro-Báez, S., & Bröder, A. (2022). "You don't know what this means to me" – Uncovering idiosyncratic influences on metamemory judgments. *Cognition*, *222*(105011), 1–9. https://doi.org/10.1016/j.cognition.2021.105011

Undorf, M., Navarro-Báez, S., & Zimdahl, M. F. (2022). Metacognitive illusions. In R. F. Pohl, *Cognitive Illusions* (3rd ed., pp. 307–323). Routledge. https://doi.org/10.4324/9781003154730-22

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., & Zimdahl, M. F. (2019). Metamemory and memory for a wide range of font sizes: What is the contribution of perceptual fluency? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 97–109. https://doi.org/10.1037/xlm0000571

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*(7), 918–933. https://doi.org/10.1037/xge0000177

Appendix

As you have seen, many students overestimated the influence of font size on their memory in this part of the experiment.

This is because words written in a large font size are particularly conspicuous during learning and are perceived as particularly easy to read and learn. However, these perceptions during learning say little about actual test performance: The font size of words does not usually affect memory performance.

You have also seen that many students underestimated the influence of an additional learning opportunity on their memory.

This is because the sensations during learning are very similar for words with an additional learning opportunity and for words without an additional learning opportunity. Despite the similarity in the sensations during learning, the following applies: An additional learning opportunity greatly improves memory performance in the test.

The following therefore applies in this experiment:

The perceptions about the ease of learning triggered by the font size say little about how well the words can actually be learnt. These perceptions should therefore not play a role in your assessments. An additional learning opportunity, on the other hand, has a stronger influence on memory.

Detecting structure: Cues for metacognitive judgments are acquired via statistical learning

Sofia Navarro-Báez[12], Arndt Bröder[1], & Monika Undorf[2]

Word Count: 8,182

Authors Note

[1]Department of Psychology, Technical University of Darmstadt.

[2]Department of Psychology, School of Social Sciences, University of Mannheim.

All data, materials, and analysis code are available at

https://osf.io/pze9b/?view_only=1f1450dc4ed4476390ea20135456daa5

Abstract

Since there is no direct monitoring of cognition during activities such as learning and memory, metacognition is inferential in nature and informed by cues. Much is known about the many cues that underlie predictions of future memory performance (judgments of learning, JOLs); however, little is known about specific mechanisms for learning cues. To address this gap, we examined statistical learning as a mechanism to extract regularities from the environment and use this knowledge to inform JOLs. Across two experiments, participants were exposed to a continuous auditory stream of artificial words with fixed transitional probabilities between adjacent syllables. Afterwards, they studied and made JOLs for (1) word items that were presented in the stream (2) phantom items that were not presented in the stream but followed transitional probabilities, and (3) non-word items that did not follow transitional probabilities. Results showed that JOLs were based on the wordlikeness cue arising from transitional probabilities: JOLs were higher for word and phantom than for non-word items. In Experiment 1, using an old-new recognition memory test, discrimination was worse for word and phantom than for non-word items. In Experiment 2, using a 2-alternative-forced-choice recognition memory test, performance did not vary across trial types. This study shows that statistical learning is one method through which metacognitive cues are acquired even when they are invalid.

Keywords: metacognition; metamemory; statistical learning; cue learning; judgments of learning

Metacognition defined as the ability to monitor and control cognitive performance is crucial for many cognitive activities such as learning and memory. Hence, it is often studied in the domain of memory by obtaining metamemory judgments like judgments of learning (JOLs). JOLs are predictions about the likelihood of remembering a recently studied item on an upcoming test (Dunlosky & Thiede, 2013). For example, a student learning for an exam predicts whether she will remember a definition at test. In metamemory research, it is well known that people cannot directly monitor the varying strengths of memory traces when making their predictions but rely on cues to make inferences about memory performance (Koriat, 1997). Many cues have been found to underlie JOLs in laboratory studies such as concreteness (Begg et al., 1989; Witherby & Tauber, 2017), word frequency (Begg et al., 1989; Benjamin, 2003; Mendes et al., 2020), word pair relatedness (Mueller et al., 2013; Undorf & Erdfelder, 2015), or font size (Mueller et al., 2014; Rhodes & Castel, 2008). However, little attention has been paid to how people learn cues for their JOLs.

Understanding the potential mechanisms of cue acquisition is essential for gaining a deeper understanding of metacognition. This study aims to examine whether statistical learning is a mechanism for learning cues for JOLs. *Statistical learning*, defined as the extraction of regularities from the environment, is a powerful form of learning (Saffran, Aslin & Newport, 1996). It was originally examined in the context of language acquisition as an experiential mechanism for segmenting fluent speech into words (Saffran et al., 1996, 1997; Seidenberg, 1997), but it has also been demonstrated in visual search (Jones & Kaschak, 2012), sequence learning (Stadler, 1992), causal learning (Sobel & Kirkham, 2007), and category learning (Brady & Oliva, 2008). Given the ubiquity of statistical learning in various areas of cognition, the idea that metacognitive cues are acquired in this fashion is a viable hypothesis deserving investigation.

Cue learning in metamemory

Research on the cue basis of metamemory judgments has mostly focused on identifying and understanding which and how cues are used rather than how they are learned. There is ample evidence that cues are used directly via beliefs about memory and/or indirectly via experiences of 'ease' during learning (e.g., Frank & Kuhlmann, 2017; Mueller et al., 2014; Undorf & Zimdahl, 2019, Undorf et al., 2017). It is only in situations where JOLs are based on invalid cues (i.e., not predictive of memory performance) or fail to rely on valid cues (i.e., predictive of memory performance) that research has addressed the question of cue learning. Typically, this is done by attempts to correct people's beliefs about memory or raising awareness about valid cues (e.g., Castel, 2008; Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Koriat & Bjork, 2006; Kornell & Bjork, 2009, Experiment 8; Mueller et al., 2015; Rhodes & Castel, 2008, Experiment 4).

For example, a set of studies on knowledge updating tested whether people learn to use study strategy effectiveness as a cue for their metamemory judgments from task experience across multiple study-test cycles (Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Mueller et al., 2015). Results show that task experience alone is insufficient for improving metamemory judgments. An explanation is that people do not keep track of their test performance and therefore cannot identify effective study strategies. In line with this explanation, providing test performance feedback or presenting study strategies during test allowed participants to learn the differential effects of study strategies and use study strategy effectiveness as a cue for their judgments (Pan & Rivers, 2023; Price et al., 2008; Yan et al., 2016). However, there are also findings showing that even participants who have learned the differential effectiveness of study strategies fail to rely on this information when making metamemory judgments (Hertzog et al., 2009; Mueller et al., 2015; Yan et al., 2016). Moreover, there is evidence that recognizing valid cues can be challenging in itself (e.g., Castel, 2008; Koriat & Bjork, 2006; Kornell & Bjork, 2009, Experiment 8; Pan & Rivers, 2023; Yan et al., 2016).

Another study by Koriat & Bjork (2006) tested whether people learn to correct their metamemory judgments from either task experience or educational training in the context of a metamemory illusion. *Foresight bias* is the overconfidence in recalling the second word of a word pair that seems obviously associated with the first word at study, but it is difficult to remember at test because the association is backward. For example, *rain – umbrella* appears semantically associated and highly memorable during study, but many other words are likely to come to mind when *rain* is presented alone at test. In one condition, participants were provided with experience in a first study-test cycle. In another condition, participants were additionally educated after the first study-test cycle by being informed about the nature of the bias and asked to estimate the likelihood of recalling ten words pairs, and then presented with the actual recall probabilities. In this study, task experience was successful in reducing foresight bias, but only when the same materials were used in a second cycle. The educational feedback, however, did not only reduce foresight bias with old materials but also with new materials, thus showing learning transfer effects.

In general, research on cue learning in metamemory shows that it is not easy for people to learn valid cues and use them for making metamemory judgments. Task experience has proven to be insufficient for learning cues (Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Mueller et al., 2015). If anything, experience with the task is helpful when using the same materials but there is no transfer to new materials (Koriat & Bjork, 2006). In cases when valid cues are learned through instructions, performance feedback, or hints at test (Hertzog et al., 2009; Mueller et al., 2015; Yan et al., 2016), the learned cues are not always used for memory predictions. Providing detailed information to participants was the most successful intervention in achieving that participants learned the predictive validity of a cue and used it consistently in their judgments (Koriat & Bjork, 2006; but see Rhodes & Castel, 2008, Experiment 4; Yan et al., 2026, Experiments 3, 4, and 5). Importantly, studies on cue learning in metamemory have not focused on the mechanisms by which cues could be spontaneously acquired from the environment.

Learning cues via statistical learning

Statistical learning refers to a set of processes through which regularities or patterns in the environment are discovered by repeated experience. One type of regularity often used by statistical learning studies is transitional probabilities. Transitional probabilities describe the predictive relationship between two elements such as syllables. For example, in the sequence "prettybaby", assuming sufficient experience with English, "pre" is more predictive of "ty" than "ty" is of "ba". In the typical auditory statistical learning paradigm from Saffran, Aslin & Newport (1996), participants are exposed to a continuous auditory stream of repeating three-syllable artificial words. Critically, the continuous stream does not contain any acoustic information about word boundaries, such as pauses or stress differences, and all syllables are repeated equally often. The only cues to word boundaries are transitional probabilities between syllables, which are high within words and low between words. These co-occurrences of syllables described by transitional probabilities are the target of statistical learning. To assess statistical learning, studies use a forced-choice-task which each artificial word from the stream is presented together with a word foil that has the same syllables as those in the stream but does not follow the transitional probabilities. Above-chance performance in this task indicates statistical learning, which has been found in many studies (e.g., Batterink et al., 2015; Endress & Mehler, 2009; Ordin & Polyanskaya, 2021; Perruchet & Poulin-Charronnat, 2012).

A key process underlying statistical learning is *extraction* which refers to the identification of statistically related clusters and their storage in memory (Thiessen et al., 2013). Two models have been proposed to explain extraction. While boundary-finding models place words boundaries where there is a low likelihood that two syllables follow each other (Endress & Mehler, 2009), clustering models group syllables that are highly likely to co-occur (Frank et al., 2010; Perruchet & Poulin-Charronnat, 2012; Perruchet & Vinter, 1998). Consequently, boundary finding models predict the extraction of statistical relationships between syllables rather than the extraction of word units as predicted by clustering models. Evidence regarding the two models is mixed. However, studies have shown that artificial words not presented in the stream that do, however, preserve the transitional probabilities (phantom items) are equally chosen as word items from the stream in the 2-alternative-forced choice test (Endress & Mehler, 2009; Ordin et al., 2020; Ordin & Polyanskaya, 2021). These findings confirm boundary finding models in auditory statistical learning.

Since phantoms are perceived as part of the language in the auditory stream (Endress & Mehler, 2009; Ordin et al., 2020; Ordin & Polyanskaya, 2021), it is likely that they acquire a wordlike quality. In fact, learners show no explicit knowledge about the regularities in the stimuli when assessed via verbal reports (e.g., Brady & Oliva, 2008; Conway & Christiansen, 2005; Turk-Browne et al., 2005). When closely examining implicit and explicit knowledge, Batterink et al. (2015) found that explicit recognition-based knowledge can be developed in parallel to implicit representations measured by a reaction time-based task. Importantly, participants who did not show explicit recognition still showed facilitation effects on the implicit measure based on reaction times. A study by Ordin and Polyanskaya (2021) suggests that confidence judgments about performance in the forced-choice-task reflect a conscious feeling of familiarity with respect to statistically learned content. Crucially, this finding does not imply that people are consciously aware of their knowledge about transitional probabilities. Overall, the literature on statistical learning strongly suggests that items with transitional probabilities (i.e., phantoms and words) are chosen as part of the language learned in the familiarization phase based on their wordlike quality rather than statistical rules.

This study

In this study, we tested statistical learning as a set of mechanisms for learning metacognitive cues to predict memory performance with JOLs. Based on prior statistical learning findings (Endress & Mehler, 2009; Ordin et al., 2020; Ordin & Polyanskaya, 2021), we expected that if statistical learning takes place, words and phantoms will be perceived as wordlike because they follow transitional probabilities. Further, we expected that participants will use wordlikeness as a cue for their JOLs.

Across two experiments, we combined the auditory statistical learning paradigm (Saffran et al., 1997; Saffran, Newport, et al., 1996) and a metamemory task with JOLs (Koriat, 1997; Undorf et al., 2018). In a familiarization phase, participants were exposed to a language that consisted of a continuous auditory stream of artificial words with fixed transitional probabilities between adjacent syllables. Afterwards, they studied and made JOLs for items that were presented in the familiarization phase and follow the transitional probabilities ('word'), for items that were not presented in the familiarization phase but follow the transitional probabilities ('phantom'), and for items that were neither presented in the familiarization phase nor followed the transitional probabilities ('non-word'). Finally, all participants completed a recognition memory test (Experiment 1) or a 2-alternative-forced-choice memory test (Experiment 2). At the end of the study, they were asked about their knowledge of syllable patterns in the studied artificial words.

To verify whether participants learned the statistical structure of the language in the familiarization phase, we asked one group of participants to indicate whether each study item belonged to the language of the auditory stream before making their JOL ("SL-assessment" group). To rule out the possibility that statistical learning effects on JOLs were only due to prompting participants to think about the language, another group made JOLs only ("no SL-assessment" group).

We predicted that if statistical learning takes place, items that follow the transitional probabilities (words and phantoms) would be more likely to be classified as belonging to the language than items that do not follow the transitional probabilities (non-words). Further, we predicted that if JOLs rely on wordlikeness acquired via statistical learning, JOLs would be higher for words and phantoms than for items that do not follow the transitional probabilities in both groups. Notably, phantoms were not presented during the familiarization phase and were therefore unfamiliar to the participants. Thus, if participants still assigned higher JOLs to phantoms than to non-words, this can only be attributed to phantoms being perceived as words of the artificial language, indicating that wordlikeness serves as a cue to inform JOLs.

Experiment 1

Experiment 1 was pre-registered (https://osf.io/836y2) and aimed to test whether metacognitive cues to predict memory performance can be acquired through statistical learning. All experiments were conducted in line with international and local ethics guidelines; they were exempt from review by the local ethics committees.

Method

Design

The design was a 3 (item type: word, phantom, non-word) x 2 (group: SL-assessment, no SL-assessment) mixed design with item type as within-subjects factor and group as between-subjects factor. Half of the participants were randomly allocated to the SL-assessment group (*n* = 45). The other half of the participants were randomly allocated to the no SL-assessment group (*n* = 42).

Materials

The stimuli were obtained from Ordin et al. (2020). The stimulus set consisted of 18 syllables that created three types of tri-syllabic nonsense items: 1) 12 words, 2) 12 phantoms, and 3) 12 non-words. In each item type subset, each syllable contributed to two items only and no pair of adjacent syllables appeared in more than one item. In words, transitional probabilities between adjacent syllables were 0.5. For example, the transitional probability that the syllable *ro* is followed by the syllable *se* is 0.5 because there were two words with *ro* (*rosenu* and *rokafa*) and *ro* was followed by *se* in one of these words. Phantoms had the same pairs of adjacent syllables as words, and thus followed the same transitional probabilities as words (e.g., *roseti, rokati*). Syllables in non-words were randomly combined with the restriction that they did not follow transitional probabilities between adjacent syllables as words. We synthesized the stimuli using eSpeak NG with the MBROLA de2 voice (Dutoit et al., 1996). Each syllable was 240 ms in duration.

For the familiarization phase, we randomly concatenated words into a continuous auditory stream with no pauses between them (18 min in total), with the restriction that each word was repeated 125 times and the same word never occurred consecutively. Because the auditory stream contained no pauses or prosodic cues indicative of word units, the only linguistic cues to word units were statistical in nature (i.e., transitional probabilities were higher between adjacent syllables within words than between words). Transitional probabilities between word boundaries were around 0.15 and varied because of the random order of syllables in the continuous auditory stream. The auditory stream was divided into 3 equal blocks of 6 minutes in length. To prevent that the first and last word unit in each block was extracted, the stream was faded in and out at the start and end of each block.

To ensure attention to the auditory stream, we asked participants to complete a pitch detection task. For this, we introduced 10 pitch changes in each block. Pitch changes (high or low) started at a random syllable of the stream (first, second, or third syllable of a word) and spanned four consecutive syllables (960 ms in duration). We created 10 additional syllable sequences for practice trials of 25 syllables from the stimulus set that were randomly concatenated without pauses and included one high and one low pitch change.

For the study phase and the recognition memory test, each word, phantom, and non-word was synthesized as a separate file that was 720 ms in duration.

Participants

We conducted an a priori power analyses using G*Power (Faul et al., 2007) with a focus on assessing a medium-sized main effect of item type on SL classifications in the SL-assessment group ($f = .25$, equivalent to $\eta_p^2 = .06$). We aimed at a sample of 43 participants per group to obtain a statistical power of $(1 - \beta) = .95$ with a $\alpha = .05$ when assuming a correlation of .50 between repeated measures.

We recruited 90 participants from the Prolific online subject pool (https://www.prolific.com) who were 18 to 30 years old, reported German as their first language, reported no language-related disorders, and had at least a high school diploma. The experiment took approximately 45 minutes and participants were paid £6.75. We excluded participants who reported technical problems ($n = 0$), admitted having completed the study without headphones ($n = 1$), admitted having used helping tools during the study ($n = 0$), admitted completing the study with the help of someone else ($n = 1$), or failed the seriousness check by admitting having just clicked through the study without taking part seriously ($n = 0$). We also excluded participants who had no variability in JOLs ($n = 1$). The final sample included 87 participants (40 female, 45 male, 2 non-binary) with a mean age of 24.38 (*SD* = 3.05), $n = 45$ in the SL-assessment group and $n = 42$ in the No SL-assessment group.

Procedure

The experiment was programmed in lab.js (Henninger, et al., 2022). At the start of the experiment, participants were informed about the study goal and procedure, and signed an online consent form. As mentioned above, the experiment consisted of a familiarization phase, study phase, distraction task, and recognition memory test. All auditory material in the experiment was played through headphones. Before the familiarization phase, participants were asked to complete a headphone screening task in which they had to judge which of three tones was the quietest (Woods et al., 2017). If participants successfully completed the headphone screening task, they could proceed to the familiarization phase.

Familiarization phase

In the familiarization phase, participants were informed that they would hear an alien language in which words were strung together without pauses in three blocks of 6 minutes each. They were asked to try to familiarize themselves with the language and get a feeling for where each alien word begins and ends. To ensure attention to the language during the familiarization phase, participants were asked to detect at least 70% of pitch changes (i.e., 7 pitch changes) in each block by pressing the upwards arrow for high pitch changes and the downwards arrow for low pitch changes. A similar cover task on pitch change detection was used by Batterink et al. (2015). Before starting with the first block of the familiarization phase, participants completed three practice trials and were given the choice to continue with up to five additional practice trials. Each practice trial took 6 s. Participants had to correctly identify one high and one low pitch change in each practice trial.

Metamemory task

Immediately after the familiarization phase, participants studied 8 randomly selected items per type (word, phantom, non-word). They were told that their task was to study 24 artificial words for a later memory test in which they would have to recognize the studied words among new artificial words. They were also asked to estimate the likelihood of recognizing each artificial word at test. Participants in the SL-assessment group were additionally informed that some of the artificial words belonged to the alien language and they would be asked to indicate whether each artificial word is an alien language word immediately after studying it. Nothing was said about the artificial words being part of the alien language to participants in the no-SL-assessment group. During the study phase, each item was displayed at the center of the screen for 3 seconds (preceded by a 500-ms fixation cross that appeared in the same location) and presented auditorily at 500 ms after the onset of the screen for a duration of 720 ms. Immediately after each item presentation, participants in the SL-assessment group classified the item as belonging to the alien language or not and then made their JOL, participants in no-SL-assessment group made their JOL immediately after each item presentation. Both SL classifications and JOLs were self-paced. SL classifications were prompted by the question “Is this word part of the alien language you got to know during the familiarization phase?” and participants typed 1 for ‘yes’ and 0 for ‘no’. JOLs were prompted by “Chances of recognition?” and participants entered any whole number between 0-100%. Participants were told that 0% indicates that the item will not be recognized and 100% indicates that the item will be absolutely recognized. Following the study phase, participants performed a 1.5-minute distraction task consisting of abstract reasoning ability items from Chierchia et al. (2019). Then, participants completed a recognition memory test that included all 24 studied items (8 words, 8 phantoms, and 8 non-words) and the 12 remaining items (4 words, 4 phantoms, and 4 non-words). At test, each item was displayed at the center of the screen and presented auditorily at 500 ms after the onset of the screen. Participants indicated whether they had studied it by typing 1 for ‘yes’ and 0 for ‘no’. After the test, participants were asked 1) to describe their approach to learning the artificial words for the memory test, and 2) to describe whether they had the impression that the syllables within artificial words followed certain patterns and, if so, at what point in time they had noticed the patterns. Finally, participants completed questions about demographics and instructions, then they were debriefed.

Results

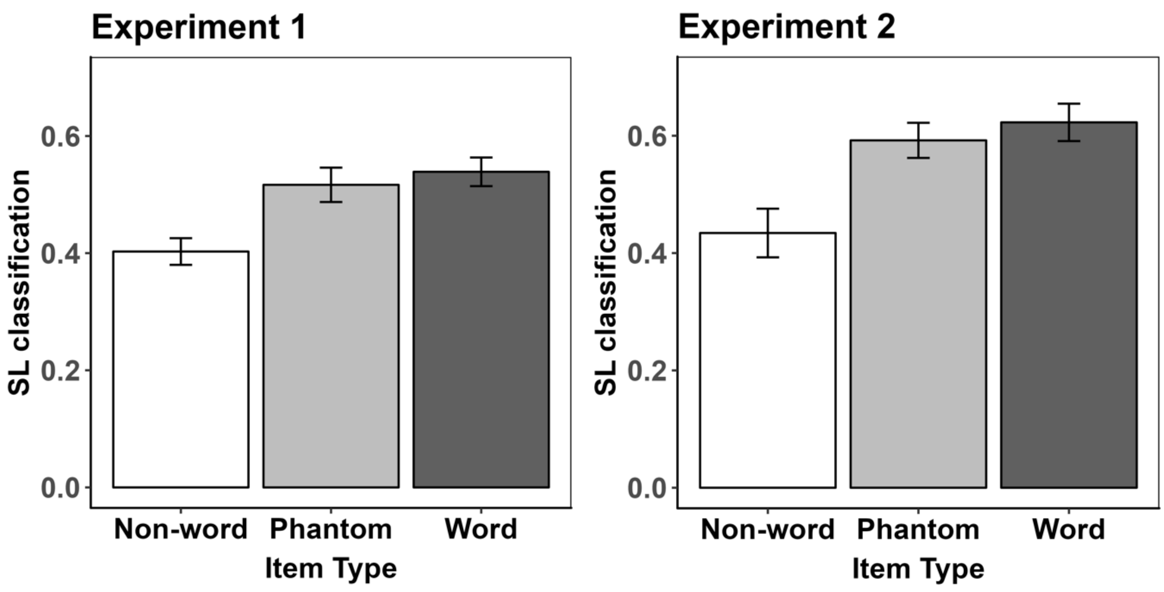
The Greenhouse-Geisser correction was used when the assumption of sphericity was violated.

SL classifications

Figure 1 presents the mean proportion of SL classifications for each item type in the SL-assessment group. A one-way ANOVA on SL classifications with item type (word, phantom, non-word) as within-subjects factor revealed that SL classifications varied with item type, $F(1.93, 84.87) = 8.55$, $p < .001$, $\eta_p^2 = .16$. Pairwise comparisons (one-tailed) showed that there was a higher proportion of SL classifications for words than for non-words, $t(44) = 4.29$, $p < .001$, $d_z = 0.64$. Similarly, the proportion of SL classifications was higher for phantoms than for non-words, $t(44) = 3.09$, $p < .01$, $d_z = 0.46$. SL classifications did not differ between words and phantoms, $t < 1$. These results demonstrate that statistical learning took place since participants classified items according to their transitional probabilities.

Figure 1

Mean proportion of SL classifications for each type of item in the SL-assessment groups in Experiment 1 and Experiment 2



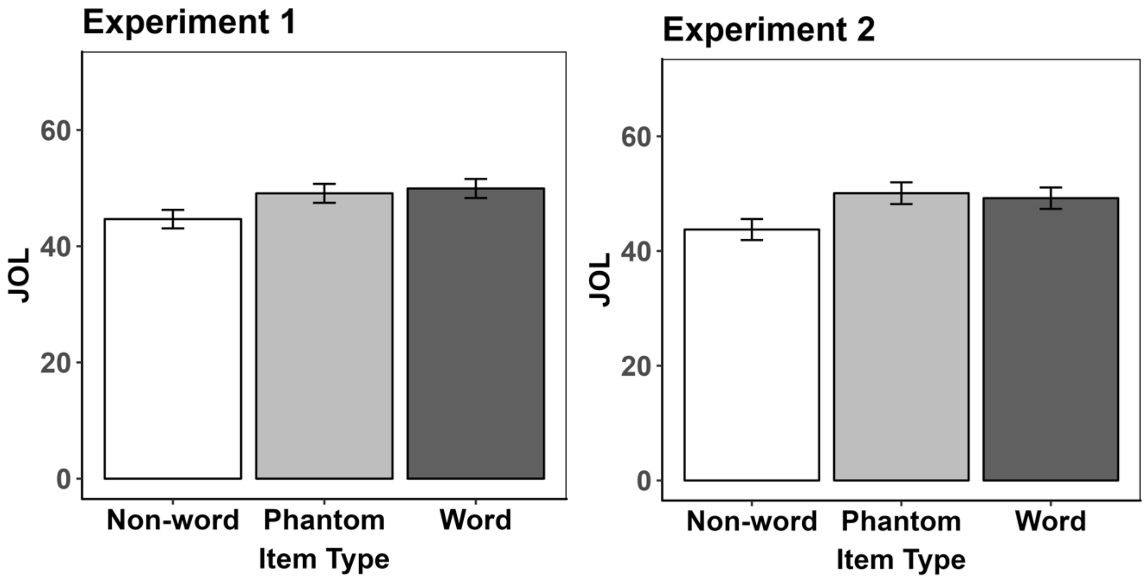
Note. Error bars represent one standard error of the mean.

JOLs

Figure 2 presents mean JOLs for each type of item collapsed across the SL-assessment and no-SL-assessment groups (see Table A1 in the appendix for mean JOLs per item type and group). A 3 (item type: word, phantom, non-word) x 2 (group: SL-assessment, no-SL-assessment) mixed ANOVA on JOLs revealed a main effect of item type, *F*(1.97, 167.51) = 12.54, $p <$ .001, η_p^2 = .13, no main effect of group, $F < 1$, and no interaction between item type and group, *F*(1.97, 167.51) = 1.95, p = .15, η_p^2 = .02. Pairwise comparisons (one-tailed) showed that JOLs were higher for words than non-words, t(86) = 4.50, p < .001, d_z = 0.67, higher for phantoms than non-words, t(86) = 3.86, p < .001, d_z = 0.58, and did not differ between words and phantoms, $t < 1$. Thus, JOLs relied on wordlikeness arising from transitional probabilities within items learned by statistical learning.

Figure 2

Mean JOL for each type of item collapsed across the no-SL-assessment and SL-assessment groups in Experiment 1 and Experiment 2



Note. Error bars represent one standard error of the mean.

SL classifications and JOLs

In addition to the pre-registered analyses, we explored the relationship between SL classifications and JOLs in the SL-assessment group by conducting a multilevel regression analysis with the R *lme4* package (Bates et al., 2015). We included SL classifications (0 = item does not belong to the alien language, 1 = item belongs to the alien language) as a fixed-effects predictor in the model and used random intercepts for participants. This model revealed a significantly positive coefficient for SL classifications, $b = 21.45$ ($SE = 1.24$), $t = 17.37$, $p < .001$, indicating that JOLs for items perceived as part of the alien language were higher by 21.45 points.

Mediation analysis

To directly test whether statistical learning is the driving mechanism for basing JOLs on the perception of wordlikeness in items following transitional probabilities, we conducted a Bayesian multilevel mediation analysis with the R *bmlm* package (Vuorre, 2017).[1] In a first model, we examined whether SL classifications mediated JOL differences between non-words and words and thus included item type (non-word = 0, word = 1) as the independent variable, SL classifications as the mediator variable, and JOLs as the outcome variable. In a second model, we examined whether SL classifications mediated JOL differences between non-words and phantoms and thus included item type (non-word = 0, phantom = 1) as the independent variable. Both models estimated the mediation effect with four Markov Chain Monte Carlo (MCMC) chains and 10,000 iterations for each chain. Results from the first model showed an indirect effect of item type (words vs. non-words) on JOLs through SL classifications, *M* = 3.04, 95% CI = [1.34, 4.99], indicating that transitional probabilities within words indirectly increased JOLs via statistical learning. However, a direct effect of item type (words vs. non-words) on JOLs, M = 3.26, 95% CI = [0.07, 6.44] indicated that statistical learning only partially mediated the JOL difference between words and nonwords. Results from the second model showed an indirect effect of item type (phantoms vs. non-words) on JOLs through SL classifications, *M* = 2.65, 95% CI = [0.84, 4.77], indicating that the transitional probabilities within phantoms indirectly increased JOLs via statistical learning. A non-reliable direct effect of item type (phantoms vs. non-words) on JOLs, *M* = 0.64, 95% CI = [-2.29, 3.56], indicated that statistical learning fully mediated the JOL difference between phantoms and non-words.

Recognition memory test

Table 1 present hits, false alarms (FAs), discrimination index (*Pr*), and bias index (*Br*) for each item type in the two groups. The discrimination index (*Pr*) was calculated by subtracting the false alarm rate from the hit rate (Snodgrass & Corwin, 1988). 3 (item type: word, phantom, non-word) x 2 (group: SL-assessment, no SL-assessment) mixed ANOVAs

[1] This analysis was not pre-registered. We are grateful to David Shanks for suggesting this analysis.

on hits and FAs revealed that both varied with item type, $F(1.94, 164.81) = 19.56$, $p < .001$, $\eta_p^2 = .19$ (hits), $F(1.87, 159.22) = 37.16$, $p < .001$, $\eta_p^2 = .30$ (FAs). The ANOVAs also revealed differences in memory performance between groups; there were more hits, $F(1, 85) = 5.89$, $p < .05$, $\eta_p^2 = .07$, and fewer FAs, $F(1, 85) = 8.95$, $p < .01$, $\eta_p^2 = .10$, in the no-SL-assessment than in the SL-assessment group. The interactions between group and item type on hits and FAs were insignificant, both $F < 1$. Pairwise comparisons showed that there were more hits and FAs for words than non-words, $t(86) = 5.16$, $p < .001$, $d_z = 0.77$ (hits), $t(86) = 7.30$, $p < .001$, $d_z = 1.09$ (FAs), and for phantoms than non-words, $t(86) = 5.28$, p < .001, $d_z = 0.79$ (Hits), $t(86) = 7.79$, $p < .001$, $d_z = 1.16$ (FAs). In contrast, hits and FAs did not differ between words and phantoms, $t < 1$ (Hits), $t(86) = 1.26$, $p = 0.21$, $d_z = 0.19$ (FAs).

A 3 x 2 mixed ANOVA revealed that *Pr* varied with item type, $F(1.91, 162.32) = 7.08$, $p < .01$, $\eta_p^2 = .07$, that *Pr* was higher in the no-SL-assessment group than in the SL-assessment group, $F(1, 85) = 16.51$, $p < .05$, $\eta_p^2 = .16$, and no interaction between item type and group, $F < 1$. Unlike SL classifications and JOLs, discrimination *Pr* was better for non-words than words, $t(86) = 3.53$, p < .001, $d_z = 0.53$, and better for non-words than phantoms, $t(86) = 2.84$, p < .001, $d_z = 0.42$, and did not differ between words and phantoms, $t(86) = 1.16$, p = .25, $d_z = 0.17$. Further, *Br* indicated a more liberal response bias for words than non-words, $t(86) = 8.33$, p < .001, $d_z = 1.24$, and for phantoms than non-words, $t(86) = 8.67$, p < .001, $d_z = 1.29$, but no difference between words and phantoms, $t < 1$. These results suggest that recognition memory was better for non-words than for words and phantoms, and that participants tended to classify words and phantoms as ‘old’ rather than ‘new’ (lenient response bias).

Table 1

Means (SDs) of Hits, FAs, Pr, and Br for each item type in the no-SL-assessment and SL-assessment groups in Experiment 1

Group and measure	Item Type		
	Non-word	Phantom	Word
No-SL assessment			
Hits	.65 (.17)	.78 (.12)	.75 (.14)
FAs	.12 (.18)	.35 (.22)	.36 (.24)
Pr	.53 (.18)	.43 (.22)	.40 (.24)
Br	.19 (.28)	.58 (.27)	.53 (.28)
SL assessment			
Hits	.59 (.19)	.71 (.19)	.72 (.19)
FAs	.19 (.21)	.42 (.25)	.50 (.30)
Pr	.40 (.21)	.29 (.25)	.22 (.30)
Br	.25 (.25)	.57 (.27)	.60 (.25)

Verbal reports

When asked at the end of the study whether they had the impression that the syllables within artificial words of the alien language followed certain patterns, around half of the participants (*n* = 39, 44.83%) exclusively said that they noticed, either during the familiarization phase or during the learning phase, that artificial words consisted of 3 syllables and that each syllable consisted of a consonant and a vowel. Other participants (*n* = 28, 32.18%) said that they had feelings of syllable repetitions or similarity among artificial words, but they did not articulate any specific sequence of syllables. The remaining participants (*n* = 20, 22.99%) said that they had not detected any syllable pattern. Thus, not a single response indicated that participants had explicit knowledge about transitional probabilities.

We also asked participants to describe their approach to learning the artificial words for the memory test (responses could be classified into multiple categories). Around half of the participants said that they had tried to associate the artificial words with real words or names (*n* = 52, 59.77%). Participants also reported that they had payed attention to the sound of words (*n* = 26, 29.89%), memorized the spelling of words (*n* = 22, 25.29%), used mental repetition (*n* = 16, 18.39%), used imagery *(n* = 3, 3.09%), or linked words together (*n* = 2, 2.3%).

Discussion

Experiment 1 results showed clear and marked statistical learning effects. On average, words and phantoms were more likely to be classified as part of the language than non-words. Further, SL classifications did not differ between words and phantoms indicating that SL classifications were based on the learned statistical structure of items rather than on increased familiarity of word units (see also Endress & Mehler, 2009). JOLs were also clearly influenced by the wordlikeness of items arising from the transitional probabilities acquired via statistical learning: they were higher for words and phantoms than for non-words and did not differ between words and phantoms. Additionally, mediation analyses showed that statistical learning as assessed through SL classifications mediated the relationship between item type and JOLs. Notably, the non-significant interaction between item type and group on JOLs suggests that statistical learning also occured in the no-SL-assessment group.

Recognition memory performance showed a different pattern of results than SL classifications and JOLs: It was better for non-words than for words and phantoms. This was probably the case because items coherent with the statistical structure of the language (words and phantoms) were highly familiar in the test, regardless of whether they were studied or new. This probably impaired discrimination and promoted a lenient response bias. In contrast, high familiarity of non-words could stem only from their occurrence in the study list, which allows for accurate recognition memory responses and yielded better discrimination and a stricter response bias for non-word items.

While mean JOLs accurately tracked differences in hit rates across item types, they were inaccurate when compared to the discrimination index *Pr*. This means that participants made correct predictions and were metacognitively competent when measuring memory performance in terms of hit rates. In doing so, participants followed task instructions which were to predict the probability of recognizing each study item, rather than to predict discrimination performance. Considering *Pr* as the relevant standard of comparison is thus one from researchers who know that *Pr* is a better measure of recognition memory performance than hit rates. Experiment 2 focused on resolving this ambiguity to conclusively evaluate the question if participants use the cue wordlikeness in an adaptive fashion to increase their JOL accuracy.

Experiment 2

Experiment 2 was pre-registered (https://osf.io/vqrt9) and had two aims. The first aim was to replicate Experiment 1 findings that people base their JOLs on the wordlikeness of items acquired via statistical learning. The second aim was to examine memory performance in a test where hit rates are unaffected by response bias. For this, we used a 2-alternative-forced-choice (2-AFC) memory test, in which participants chose between one target item and one distractor item. To prevent any response tendencies or biases based on the type of item, target and distractor items were always from the same type (i.e, word-word, phantom-phantom, or non-word-non-word). This allowed us to test whether JOLs are predictive of memory performance when JOL instructions (chance to recognize each study item at test) coincide with the accuracy measure (percentage of correct choices).

Method

Design

The design was identical to Experiment 1.

Materials

Materials were the same as in Experiment 1.

Participants

Power analysis was identical to that of Experiment 1. We aimed at recruiting 43 participants in each group. We recruited 50 participants from the Prolific online subject pool (https://www.prolific.com) and 39 participants from the student pool of the Technical University of Darmstadt. Participants were 18 to 35 years old, reported German as their first language, reported no language related disorders, and had at least a high school diploma as highest degree. The experiment took approximately 45 minutes. Participants from Prolific were paid £6.75 or €7.90 approximately. Participants from the Technical University of Darmstadt received course credits or €7.50. As in Experiment 1 and according to the pre-registration, we excluded participants who reported technical problems (*n* = 0), admitted having completed parts of the study without headphones (*n* = 1), admitted having used helping tools during the study (*n* = 1), admitted completing the study with the help of someone else (*n* = 2), or admitted having just clicked through the study without taking part seriously (*n* = 0). We excluded no participant for lacking variability in JOLs. The final sample included 85 participants (50 female, 34 male, 1 non-binary) with a mean age of 23.71 (*SD* = 3.41), *n* = 38 in the SL-assessment group and *n* = 47 in the no-SL-assessment group.

Procedure

The procedure was identical to Experiment 1 except that the participants studied 18 randomly selected items per type (6 words, 6 phantoms, 6 non-words) and that the memory test was a 2-AFC recognition memory test. The 36 items used in the 2-AFC test included 18 studied and 18 new items for a total of 18 trials. In the study phase, participants were informed that their task was to study artificial words for a later memory test in which each study word will be shown together with a new artificial word, and they would have to recognize the studied word. At test, two items were displayed vertically centered at the left and right side of the screen. The item at the left side was displayed first at the onset of the screen and presented auditorily at 500 ms after the onset of the screen. The item at the right side was displayed second at 1420 ms after the onset of the screen and presented auditorily at 1920 ms after the onset of the screen (i.e., with a gap of 700 ms between the two auditory presentations). Participants indicated that they had studied the item at the left side by typing 1 and that they had studied the item at the right side by typing 0. We counterbalanced item positions such that one half of the studied items were presented on each side for every participant.

Results

The Greenhouse-Geisser correction was used when the assumption of sphericity was violated.

SL classifications

Figure 1 presents the mean proportion of SL classifications for each type of item in the SL-assessment group. A one-way ANOVA revealed that SL classifications varied with item type, $F(1.87, 69.30) = 9.26$, $p < .001$, $\eta_p^2 = .20$. Pairwise comparisons (one-tailed) showed that proportions of SL classifications were higher for words than for non-words, $t(37) = 3.58$, $p < .001$, $d_z = 0.58$, for phantoms than for non-words, $t(37) = 3.71$, $p < .001$, $d_z = 0.60$, and did not differ between words and phantoms, $t < 1$. As in Experiment 1, these results show that statistical learning took place.

JOLs

Figure 2 presents mean JOLs for each type of item collapsed across the SL-assessment and no-SL-assessment groups. A 3 (item type: word, phantom, non-word) x 2 (group: SL assessment, no SL assessment) mixed ANOVA on JOLs revealed a main effect of item type, $F(1.97, 163.44) = 13.34$, $p < .001$, $\eta_p^2 = .14$, no main effect of group, $F < 1$, and no interaction between item type and group, $F < 1$ (see Table A1 in the appendix for mean JOLs per item type and group). Pairwise comparisons (one-tailed) showed that JOLs were higher for words than for non-words, $t(84) = 4.17$, p < .001, $d_z = 0.45$, for phantoms than for non-words, $t(84) = 4.96$, p < .001, $d_z = 0.54$, and did not differ between words and phantoms, $t < 1$. These findings replicate JOL results from Experiment 1.

SL classifications and JOLs

As in Experiment 1, in addition to pre-registered analyses, we examined the relationship between SL classifications and JOLs in the SL-assessment group by conducting a multilevel regression analysis with random intercepts for participants. SL classifications (0 = item does not belong to the alien language, 1 = item belongs to the alien language) were the fixed-effects predictor. The significantly positive coefficient for SL classifications, $b = 19.04$ ($SE = 1.38$), $t = 14.03$, $p < .001$, indicated that JOLs for items perceived as part of the alien language were higher by 19.04 points.

Mediation analysis

We conducted the same Bayesian multilevel mediation analysis as in Experiment 1. The first model again included item type (non-word = 0, word = 1) as independent variable, SL classifications as mediator variable, and JOLs as outcome variable and the second model was identical to the first one except that it coded item type (non-word = 0, phantom = 1) as independent variable. Results from the first model showed an indirect effect of item type (words vs. non-words) on JOLs through SL classifications, $M = 3.79$, 95% CI = [1.56, 6.27], indicating that transitional probabilities within words indirectly increased JOLs via statistical learning. The direct effect of item type (words vs. non-words) on JOLs, $M = 3.63$, 95% CI = [0.14, 7.20] again indicated that statistical learning only partially mediated the JOL difference between words and non-words. Results from the second model showed an indirect effect of item type (phantoms vs. non-words) on JOLs through SL classifications, $M = 2.78$, 95% CI = [0.92, 4.90], indicating that transitional probabilities within phantoms indirectly increased JOLs via statistical learning. Unlike in Experiment 1, the direct effect of item type (phantom vs. non-word), $M = 4.16$, 95% CI = [0.98, 7.38], indicated that statistical learning only partially mediated the JOL difference between phantoms and non-words.

Two-alternative-forced-choice (2-AFC) memory test

Table 2 presents the percentage of correct responses in the 2-AFC test for each item type in both groups. A 2 (Group: SL-assessment, no-SL-assessment) x 3 (Item type: word, phantom, non-word) mixed ANOVA[2] revealed that the percentage of correct responses did not vary with item type, $F < 1$. It also revealed higher performance in the no-SL-assessment group than in the SL-assessment group, $F(1, 83) = 4.30$, $p = .04$, $\eta_p^2 = .05$, and a significant interaction between group and item type, $F(2.00, 165.59) = 3.59$, $p = .03$, $\eta_p^2 = .04$. Comparing the percentage of correct responses for each item type across groups revealed better memory performance for non-words in the no-SL-assessment group than in the SL-assessment group, $t(83) = 3.25$, $p = .002$, $d_s = 0.71$, but no group differences for words, $t(83) = 1.49$, p = .14, $d_s = 0.32$, or phantoms, $t < 1$.

----

[2] We deviate from the pre-registration here by not reporting results for the target-order condition (left, right). When including target order as an additional control factor in the ANOVA, we found a significant interaction between item type and target-order condition, $F(1.97, 163.49) = 4.47$, $p = .01$, $\eta_p^2 = .05$, which was driven by better memory performance for words presented on the right side, $t(84) = 3.24$, $p < .01$, $d_z = 0.35$. No differences by target-order were found for phantoms or non-words, $t <= 1.19$. We did not expect the significant interaction and abstain from further interpretation. All other results of the pre-registered ANOVA were identical to the ones reported here.

Table 2

Mean (SD) percentage of correct responses in the 2-AFC test for each type of item in the No SL-assessment and SL-assessment group in Experiment 2

Item Type	% correct	
	No-SL assessment	SL-assessment
Non-word	85.82 (34.95)	75 (43.40)
Phantom	79.08 (40.75)	81.14 (39.20)
Word	81.56 (38.85)	75.88 (42.88)

Verbal reports

At the end of the study, participants (*n* = 35, 41.18%) exclusively said that they noticed that artificial words consisted of 3 syllables and that each syllable consisted of a consonant and a vowel. Other participants (*n* = 23, 27.06%) said that they had feelings of syllable repetitions or similarity among artificial words, but they did not articulate any specific sequence of syllables except for one participant who mentioned the syllable “ko” appearing together with “niko” or “riko” and therefore “iko” was memorable. The remaining participants (*n* = 27, 31.76%) said that they had not detected any syllable pattern. Importantly, none of the results reported above changed when we excluded the one participant who verbalized explicit knowledge about one transitional probability from the analysis.

As in Experiment 1, we also asked participants to describe their approach to learning artificial words for the memory test. Again, around half of the participants (*n* = 46, 54.12%) said that they tried to associate the artificial words with real words or names. Participants also reported using mental repetition (*n* = 26, 30.59%), paying attention to the sound of the words (*n* = 11, 12.94%), using imagery *(n* = 11, 12.94%), memorizing the spelling (*n* = 10, 11.76%), and linking words together (*n* = 3, 3.53%).

Discussion

Experiment 2 replicated Experiment 1 in showing that statistical learning took place and that JOLs were based on the wordlikeness of items, with higher JOLs for words and phantoms than for non-words and no JOL differences between words and phantoms. The mediation analysis again showed that SL classifications mediated the relationship between the item type and JOLs, demonstrating that participants used the wordlikeness of items acquired via statistical learning as a cue for JOLs.

Results clearly showed that JOLs did not accurately track memory differences across item types, because performance in the 2-AFC test was similar for words, phantoms, and non-words. Discarding response bias in the 2-AFC memory test thus provided clear evidence that basing JOLs on wordlikeness resulted in discrepancies between JOLs and actual memory performance.

As in Experiment 1, memory performance was worse in the SL-assessment group than in the no-SL-assessment group. A plausible explanation is that there are less resources for encoding when participants have the additional task of making an SL-classification immediately after studying each item (Mitchum et al., 2016). Unlike in Experiment 1, however, this difference in performance was restricted to the non-words because non-words are presumably the most difficult to encode and suffer the most from an additional task.

General Discussion

This study tested statistical learning as a mechanism for extracting regularities from the environment and using them as metacognitive cues to predict memory performance. Across two experiments, JOLs were based on the perceived wordlikeness of items acquired via statistical learning. An intriguing aspect of the results is that study items that had been repeatedly presented as part of the familiarization phase (words) did not yield higher JOLs than study items that followed transitional probabilities but had never been presented (phantoms). This clearly demonstrates that participants learned the statistical relations between syllables that added a wordlike quality to the items and used wordlikeness as a cue for their JOLs.

Our results also show that statistical learning of cues did not contribute to JOL accuracy but rather resulted in people basing their JOLs on an invalid cue. Experiments 1 and 2, using a recognition memory performance test and a 2-alternative forced-choice memory test, respectively, show that items following transitional probabilities from the familiarization phase were not better remembered than items not following transitional probabilities. Thus, contrary to what people's JOLs predicted, items perceived as wordlike did not have a memory advantage over items not perceived as wordlike. Examining whether this is true for metacognitive cues acquired through statistical learning in general or specific to the paradigms used in this study is an important avenue for future research. For instance, wordlikeness might be predictive of actual memory performance when using a recall memory test. In an analogous fashion, JOLs predict that high-frequency words are more memorable than low-frequency words, which is accurate in recall tests, but not in recognition tests (Benjamin, 2003).

The present study is an important first step towards understanding the mechanisms of cue acquisition in metamemory, a topic that has been largely neglected in the metacognition literature. Its results indicate that statistical learning is one way for people to extract regularities from the environment and to use them as cues for inferences about future memory. This is relevant for metamemory in real-world learning because regularities or patterns in the environment are abundant, and statistical learning processes are present in many everyday situations. For example, in many real languages, syllables with high transitional probabilities are likely to be part of the same word (Swingley, 2005). A study by Alexander et al. (2023) found that statistical learning mechanisms support the acquisition of a second language in real-world contexts. In this study, participants who were exposed to Italian for two weeks demonstrated better ability to detect Italian words than a control group. It is thus possible that lexical regularities inform metamemory monitoring when learning a second language.

In the visual and conceptual domain, a study found that participants learned co-occurrences in the categories of real-world scenes (Brady & Oliva, 2008). In particular, sequences of scene triplets (e.g., A, B, C; mountain, bathroom, and street scene) were discriminated above chance level from foil sequences that presented images from three different triplets. This statistical categorical learning occurred even when the scenes were different and only the categories remained the same, when the test used verbal labels rather than scenes (e.g., “mountain”), and when participants were not consciously aware of the categorical patterns in the stream of scenes. The authors concluded that statistical learning is a useful tool for organizing conceptual knowledge and learning relationships between visual episodes. For instance, cognitive maps can be learned by extracting sequences of places that frequently appear together, such as “playgrounds appearing in the center of residential areas”. This can influence metacognition during navigation by increasing the confidence of being near a playground when the area looks residential.

At a more general level, this study contributes to the metamemory literature by demonstrating that *any* feature or cue may be incorporated in metamemory judgments if it is connected to a belief about how memory works and/or feelings of ease or fluency. In this study, it is likely that items perceived as wordlike were processed more fluently than items not perceived as wordlike. In fact, statistical learning facilitates the processing of complex stimuli and results in structures becoming simpler (Forest et al., 2022). Notably, any fluency of processing wordlike items in our study can only be attributed to statistical learning mechanisms, because all items were composed from the same syllables that occurred for the same number of times and, thus, were similar in complexity. Since items acquired their wordlike quality with experience, our study provides support for the experiential basis of JOLs other studies have found (e.g., Besken, 2016; Frank & Kuhlmann, 2017; Undorf & Erdfelder, 2011, 2015).

In conclusion, this study is the first to demonstrate that cues for metamemory judgments can be acquired through statistical learning. This finding opens the door for further research on mechanisms of cue learning, a topic that has been largely overlooked in metacognition research. We hope that understanding how people acquire information from the environment that they then utilize for assessing their cognitions will help to develop new methods for effectively training metacognition.

Acknowledgments

This work was supported by grants UN 345/2-1 (Monika Undorf) and BR 2130/14-1 (Arndt Bröder) from the Deutsche Forschungsgemeinschaft and by the University of Mannheim's Graduate School of Economic and Social Sciences (Sofia Navarro-Báez). We thank Clara Wiebel, Jana Jung, and Anna Tieftrunk for help with material and experiment preparation.

References

Alexander, E., Van Hedger, S. C., & Batterink, L. J. (2023). Learning words without trying: Daily second language podcasts support word-form learning in adults. *Psychonomic Bulletin & Review*, *30*(2), 751–762. https://doi.org/10.3758/s13423-022-02190-1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language*, *83*, 62–78. https://doi.org/10.1016/j.jml.2015.04.004

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610–632. https://doi.org/10.1016/0749-596X(89)90016-8

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*(2), 297–305. https://doi.org/10.3758/BF03194388

Besken, M. (2016). Picture-perfect is not perfect for metamemory: Testing the perceptual fluency hypothesis with degraded images. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1417–1433. https://doi.org/10.1037/xlm0000246

Brady, T. F., & Oliva, A. (2008). Statistical Learning Using Real-World Scenes: Extracting Categorical Regularities Without Conscious Intent. *Psychological Science*, *19*(7), 678–685. https://doi.org/10.1111/j.1467-9280.2008.02142.x

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, *36*(2), 429–437. https://doi.org/10.3758/MC.36.2.429

Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S.-J. (2019). The matrix reasoning item bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society Open Science,* 6, 1–13, http://dx.doi.org/10.1098/rsos.190232

Conway, C. M., & Christiansen, M. H. (2005). Modality-Constrained Statistical Learning of Tactile, Visual, and Auditory Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 24–39. https://doi.org/10.1037/0278-7393.31.1.24

Dunlosky, J., & Hertzog, C. (2000). Updating Knowledge About Encoding Strategies: A Componential Analysis of Learning About Strategy Effectiveness From Task Experience. *Psychology and Aging*, *15*(3), 462–474.

Dunlosky, J., & Thiede, K. W. (2013). *Metamemory*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0019

Dutoit, T, Pagel, V, Pierret, N., Baraille, F., & van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. *Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP)*, *3*, 1393–1396. https://doi.org/10.1109/ICSLP.1996.607874

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351–367. https://doi.org/10.1016/j.jml.2008.10.003

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

Forest, T. A., Siegelman, N., & Finn, A. S. (2022). Attention Shifts to More Complex Structures With Experience. *Psychological Science*, *33*(12), 2059–2072. https://doi.org/10.1177/09567976221114055

Frank, D. J., & Kuhlmann, B. G. (2017). More than just beliefs: Experience and beliefs jointly contribute to volume effects on metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 680–693. https://doi.org/10.1037/xlm0000332

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. https://doi.org/10.1016/j.cognition.2010.07.005

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). lab.js: A free, open, online study builder. *Behavior Research Methods*, *54*(2), 556–573. https://doi.org/10.3758/s13428-019-01283-5

Hertzog, C., Price, J., Burpee, A., Frentzel, W. J., Feldstein, S., & Dunlosky, J. (2009). Why do people show minimal knowledge updating with task experience: Inferential deficit or experimental artifact? *Quarterly Journal of Experimental Psychology*, *62*(1), 155–173. https://doi.org/10.1080/17470210701855520

Jones, J. L., & Kaschak, M. P. (2012). Global statistical learning in a visual search task. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 152–160. https://doi.org/10.1037/a0026233

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

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *32*(5), 1133–1145. https://doi.org/10.1037/0278-7393.32.5.1133

Kornell, N., & Bjork, R. A. (2009). A Stability Bias in Human Memory: Overestimating Remembering and Underestimating Learning. *Journal of Experimental Psychology: General*, *138*(4), 449–468. https://doi.org/10.1037/a0017350

Mendes, P. S., Luna, K., & Albuquerque, P. B. (2020). Experience Matters: Effects of (In)Congruent Prompts About Word Frequency on Judgments of Learning. *Zeitschrift Für Psychologie*, *228*(4), 254–263. https://doi.org/10.1027/2151-2604/a000423

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General*, *145*(2), 200–219. https://doi.org/10.1037/a0039923

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, *20*(2), 378–384. https://doi.org/10.3758/s13423-012-0343-6

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2015). Why is knowledge updating after task experience incomplete? Contributions of encoding experience, scaling artifact, and inferential deficit. *Memory & Cognition*, *43*(2), 180–192. https://doi.org/10.3758/s13421-014-0474-2

Ordin, M., Polyanskaya, L., Soto, D., & Molinaro, N. (2020). Electrophysiology of statistical learning: Exploring the online learning process and offline learning product. *European Journal of Neuroscience*, *51*(9), 2008–2022. https://doi.org/10.1111/ejn.14657

Ordin, M., & Polyanskaya, L. (2021). The role of metacognition in recognition of the content of statistical learning. *Psychonomic Bulletin & Review*, *28*(1), 333–340. https://doi.org/10.3758/s13423-020-01800-0

Pan, S. C., & Rivers, M. L. (2023). Metacognitive awareness of the pretesting effect improves with self-regulation support. *Memory & Cognition*, *51*(6), 1461–1480. https://doi.org/10.3758/s13421-022-01392-1

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. https://doi.org/10.1016/j.jml.2012.02.010

Perruchet, P., & Vinter, A. (1998). PARSER: A Model for Word Segmentation. *Journal of Memory and Language*, *39*(2), 246–263. https://doi.org/10.1006/jmla.1998.2576

Price, J., Hertzog, C., & Dunlosky, J. (2008). Age-Related Differences in Strategy Knowledge Updating: Blocked Testing Produces Greater Improvements in Metacognitive Accuracy for Younger than Older Adults. *Aging, Neuropsychology, and Cognition*, *15*(5), 601–626. https://doi.org/10.1080/13825580801956225

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615–625. https://doi.org/10.1037/a0013684

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, *35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental Language Learning: Listening (and Learning) Out of the Corner of Your Ear. *Psychological Science*, *8*(2), 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x

Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science*, *10*(3), 298–306. https://doi.org/10.1111/j.1467-7687.2007.00589.x

Stadler, M. A. (1992). Statistical structure and implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(2), 318–327. https://doi.org/10.1037/0278-7393.18.2.318

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86–132. https://doi.org/10.1016/j.cogpsych.2004.06.001

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin*, *139*(4), 792–814. https://doi.org/10.1037/a0030801

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The Automaticity of Visual Statistical Learning. *Journal of Experimental Psychology: General*, *134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264–1269. https://doi.org/10.1037/a0023719

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition*, *43*(4), 647–658. https://doi.org/10.3758/s13421-014-0479-x

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., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language*, *92*, 293–304. https://doi.org/10.1016/j.jml.2016.07.003

Undorf, M., & Zimdahl, M. F. (2019). Metamemory and memory for a wide range of font sizes: What is the contribution of perceptual fluency? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 97–109. https://doi.org/10.1037/xlm0000571

Witherby, A. E., & Tauber, S. K. (2017). The concreteness effect on judgments of learning: Evaluating the contributions of fluency and beliefs. *Memory & Cognition*, *45*(4), 639–650. https://doi.org/10.3758/s13421-016-0681-0

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*(7), 918–933. https://doi.org/10.1037/xge0000177

Appendix

Table A1

Means (SDs) of JOLs for each type of item in the No SL-assessment and SL-assessment group in Experiment 1 and Experiment 2

Group and Item Type	JOLs	
	Experiment 1	Experiment 2
No-SL assessment		
Non-word	44.85 (24.32)	43.71 (23.36)
Phantom	50.32 (26.68)	49.89 (24.53)
Word	48.85 (26.32)	47.80 (25.31)
SL-assessment		
Non-word	44.49 (24.65)	43.79 (25.63)
Phantom	47.96 (24.90)	50.31 (26.02)
Word	50.95 (25.79)	50.96 (26.28)

This thesis was supported by the *Graduate School of Economic and Social Sciences* (GESS) at the University of Mannheim.

Cover image designed by neonfactory12/Freepik.