Oxford Handbooks Online

Judgments of Learning: Methods, Data, and Theory

Matthew G. Rhodes

The Oxford Handbook of Metamemory
Edited by John Dunlosky and Sarah (Uma) K. Tauber

Print Publication Date: Jun 2016 Subject: Psychology, Cognitive Psychology, Cognitive Neuroscience
Online Publication Date: Feb 2015 DOI: 10.1093/oxfordhb/9780199336746.013.4

Abstract and Keywords

Several decades of research have examined predictions of future memory performance—typically referred to as judgments of learning (JOLs). In this chapter, I first discuss the early history of research on JOLs and their fit within a leading metacognitive framework. A common methodological approach has evolved that permits the researcher to investigate the correspondence between JOLs and memory performance, as well as the degree to which JOLs distinguish between information that is or is not remembered. Factors that influence each aspect of the accuracy of JOLs are noted and considered within theoretical approaches to JOLs. Thus far, research on JOLs had yielded a number of findings and promising theoretical frameworks that will continue to be refined. Future work will benefit by considering how learners combine information to arrive at a judgment, the implications of alternative methods of measuring JOLs, and the potential for JOLs to influence memory.

Keywords: metamemory, monitoring, judgments of learning, memory predictions, self-regulation of learning

Consider the following scenario. A student is preparing for an exam on the first century CE of the Roman Empire and is attempting to understand the vacillation between peace and conflict, exemplified by the various emperors from Caesar to Trajan. After careful review of the life and role of each emperor, the student evaluates whether learning has been sufficient so as to assure some degree of successful performance on the upcoming test. Will this prediction of future memory success prove to be accurate? What information has informed the prediction?

Questions of this sort have been examined formally for nearly 50 years by soliciting predictions of future memory performance, termed *judgments of learning* (JOLs). In the present chapter, I review the fruits of this literature, focusing first on JOLs within a broader framework of metacognition and the earliest research on predictions of future memory performance. Next, I consider typical methods of soliciting JOLs and the general accuracy of JOLs, highlighting variables that do or do not influence judgment. These findings are considered within the key theoretical frameworks that have been offered to explain how individuals arrive at a prediction of future memory performance. Finally, I discuss emerging issues that may characterize future research on JOLs.

Judgments of Learning within the Nelson and Narens (1990) Monitoring and Control Framework

In their classic framework for considering metacognition (i.e., our awareness of our own cognition), Nelson and Narens (1990; see also Nelson, 1996; Thiede, Dunlosky, & Mueller, this volume) distinguished between processes related to assessing one's learning (*monitoring*) and the self-regulation of learning based on information gained from monitoring (*control*). For our example student, reflection on whether information had been mastered would comprise monitoring with control processes (p. 66) reflected by a decision on whether to engage in further study of an emperor. The link between monitoring and control is a fundamental assumption of any research in metacognition. Indeed, monitoring appears to influence control over learning even when at odds with objective indices of learning (Metcalfe & Finn, 2008; Rhodes & Castel, 2009). However, although individuals may be imperfect at monitoring the contents of cognition, "*A system that monitors itself (even imperfectly) may use its own introspections as input to alter the system's behavior*" (Nelson & Narens, 1990, p. 128, italics in original).

Accordingly, JOLs constitute one type of introspective judgment that provides input to a metacognitive system that may alter behavior. Within the Nelson and Narens (1990) framework, JOLs were listed as one of several prospective monitoring judgments possible in anticipation of future memory performance. Whereas JOLs occur during information acquisition (i.e., encoding) or even during storage, other prospective judgments take place prior to encoding (e.g., ease of learning judgments) or following unsuccessful retrieval (e.g., feeling-of-knowing judgments). These prospective judgments, considering the future state of memory, may be contrasted with retrospective judgments that occur after a memory has been retrieved and involve reflecting on the current products of memory. The most common example is the retrospective confidence judgment, whereby individuals indicate their confidence that some bit of retrieved information is accurate.

Although prospective judgments are considered distinct from retrospective judgments, relatively little research has compared the different judgments under the same encoding and retrieval conditions. The work that has been done suggests that prospective judgments rely to some extent on qualitatively different information than retrospective judgments (Dougherty, Scheck, Nelson, & Narens, 2005; see also Busey, Tunnicliff, Loftus, & Loftus, 2000). As well, different types of prospective judgments may rely on different sources of information (but see Dunlosky & Tauber, 2013). For example, Leonesio and Nelson (1990) reported that ease of learning judgments were less accurate predictors of future memory performance than JOLs.

A Brief History of Early Research on Judgments of Learning

Much like other modern studies of metacognition (e.g., Hart, 1965; see chapter 1 of this volume), systematic research on JOLs began only within the past 50 years, and was initially pursued sporadically, with gaps of years sometimes occurring between publications. Arbuckle and Cuddy (1969) reported the first investigation of predictions of future memory performance made during encoding (i.e., JOLs). Their interest was in whether participants could identify differences in the associative strength of two items and thus make memory predictions that were consistent with associative strength. Accordingly, in two experiments, participants studied sets of paired associates (word-number pairs or word-word pairs) and either made a yes/no judgment of whether a target would be recalled (Experiment 1) or made predictions on a five-point Likert scale (Experiment 2). The results from both experiments showed that predictions were generally accurate, at the very least exceeding a criterion of chance performance. For example, items given a “yes” prediction were more likely to be recalled than items given a “no” prediction, with “yes” predictions more likely for strongly than weakly related pairs. Likewise, items rated as “very likely” to be recalled were recalled more frequently than items rated as “unlikely” to be recalled. Arbuckle and Cuddy (1969) concluded their paper with an optimistic assessment of the accuracy of memory predictions and suggested an agenda for future research, including exploring factors that might influence predictions even when memory performance was unaffected.

Their suggestions for future work went largely unheeded at the time. Instead, memory predictions were explored only intermittently and often confined to a burgeoning line of work examining metacognition in children spearheaded by John Flavell (Flavell, 1979). For example, Flavell, Friedrichs, and Hoyt (1970) used a variant of the JOL procedure with young children who saw pictures of objects and were asked to predict the maximum number of pictures that could be perfectly recalled. In general, such memory span predictions exceeded performance, with children of kindergarten age or younger more likely to exhibit overconfidence (see also Levin, Yussen, Pressley, & de Rose, 1977; Yussen & Levy, 1975). These experiments with children captured only one aspect of JOLs, a general prediction of overall performance based on exposure to a subset of material, but not online predictions made during learning. Indeed, Groninger (1976) reported the first investigation after Arbuckle and Cuddy (1969) to solicit predictions of memory performance during learning. Groninger's (1976; see also Groninger, 1979) participants studied a list composed of several classes of words (concrete, abstract, nonsense, emotional) and (p. 67) made confidence judgments regarding the likelihood of later recognition for each word. In general, participants were more likely to subsequently recognize items given high confidence judgments during learning, and judgments were reported to be sensitive to the type of items studied, with participants most confident of remembering emotional items.

The decade from 1980 to 1990 similarly yielded few papers on JOLs, but was characterized by some significant contributions. King, Zechmeister, and Shaughnessy (1980) had participants make JOLs for paired associates presented multiple times across several blocks. JOLs were made either in the context of multiple study opportunities or amidst alternating study and test opportunities. Testing enhanced performance on a final test of recall and also yielded more accurate predictions than studying alone. Specifically, King et al. (1980) demonstrated that JOLs were markedly higher for previously tested items that had been recalled compared with items that had not been recalled (see also Lovelace, 1984), presaging later accounts of how participants arrive at memory predictions during multi-trial learning (e.g., Finn & Metcalfe, 2007). Nevertheless, by 1990, the first two decades of work on JOLs had yielded some important insights (see also Mazzoni, Cornoldi, & Marchitelli, 1990; Vesonder & Voss, 1985) but few publications and data. Indeed, in their comprehensive review and theoretical treatise on metacognition, Nelson and Narens (1990) cite only three published papers that collected data on JOLs. However, this was merely a prelude to an explosion of work on JOLs that is the primary focus of the remainder of the chapter.

Methodology and Notable Findings

The methodology for soliciting JOLs has changed minimally since Arbuckle and Cuddy's (1969) original paper. For example, in a typical experiment, participants are presented with memoranda, such as paired associates (e.g., *Table-Spoon*), one at a time. Either immediately after the presentation of a pair or at some delay, participants are prompted to make a JOL of the likelihood of later recalling the target (e.g., *Spoon*), often cued by the stimulus (e.g., *Table-?*). Finally, after making JOLs for each item, participants are given a memory test for the studied items.[1] Several variants on this approach have been reported. For example, Nelson, Narens, and Dunlosky (2004) created the *prejudgment recall and monitoring* (PRAM) procedure to assess the contents of memory that inform predictions by having participants engage in retrieval just prior to providing a JOL. Also, Castel (2008) introduced the prestudy JOL procedure, whereby participants are given information about the nature of a study item (e.g., its serial position within a list) and make their JOL prior to actually viewing this item.

Unlike Arbuckle and Cuddy (1969), who solicited judgments on a 5-point Likert scale, most recent work on JOLs has asked participants to make their judgment as a probability, or percentage likelihood, of future successful memory performance[2] (but see the Pending Issues and Future Directions section for other judgments). Using a percentage scale permits investigators to calculate measures of *absolute accuracy* (i.e., *calibration*), the overall correspondence between judgment and performance (see Higham, Zawadzka, & Hanczakowski, this volume, for additional details). Consider a participant who has studied 10 faces and made a JOL for each face prior to a recognition test. If the average JOL was 70%, then absolute accuracy would be perfect if 70% of the studied faces were subsequently recognized. Judgment would be characterized by overconfidence if JOLs exceeded performance and underconfidence if performance exceeded JOL magnitude. There is some debate as to whether participants can faithfully translate memory predictions to a probability scale (e.g., Hanczakowski, Zawadzka, Pasek, & Higham, 2013), but this approach remains the most common given that the judgment (probability of recall/recognition) can be made on the same scale as the measure of memory.

Whereas absolute accuracy refers to the overall correspondence between judgment and performance, *relative accuracy* (i.e., *resolution*) refers to the degree to which an individual's JOLs distinguish between what is and what is not remembered. Relative accuracy is typically measured via the Kruskal-Goodman gamma correlation, a nonparametric index of association that ranges from –1.0 to +1.0 and quantifies the association between JOLs and memory performance (Nelson, 1984; Gonzalez & Nelson, 1996). To the degree that subsequently remembered items are given high JOLs and items that are less likely to be remembered are given lower JOLs, gamma will be positive. Likewise, to the degree that subsequently remembered items are given low JOLs and items that are less likely to be remembered are given higher JOLs, gamma will be negative. Gamma correlations at or approaching zero suggest that there is no relationship between JOLs and subsequent memory performance.

Accordingly, the ideal learner's memory predictions would be considered accurate to the extent (p. 68) that JOL magnitude matched memory performance and to the extent that JOLs discriminated between information that was or was not remembered. Given that absolute and relative accuracy reflect different components of accuracy, major findings should be considered in light of those measures (Koriat, 1997).

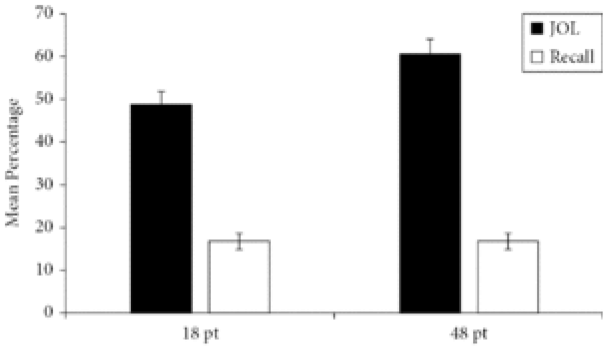
Factors That Influence the Absolute Accuracy of Judgments of Learning

Absolute accuracy is affected if a given factor influences (a) the magnitude of JOLs and/or (b) the likelihood that target information is remembered. An exhaustive account of all factors that affect absolute accuracy is beyond the scope of this review (see Schwartz & Efklides, 2012, for a partial list). However, I consider influences on JOL magnitude in terms of factors that characterize the item or information to be remembered (e.g., association strength, concreteness, etc.) and factors that characterize the conditions of study and testing (e.g., presentation rate, encoding operations, test format, retention interval, etc.).[3]

Item-Based Influences on JOL Magnitude.

The original work on JOLs (Arbuckle & Cuddy, 1969) demonstrated that learners deemed related pairs of items (e.g., *Table-Chair*) to be more memorable than unrelated pairs of items (e.g., *Horse-Rugby*). The increase in JOL magnitude for related items has been replicated consistently (e.g., Castel, McCabe, & Roediger, 2007; Koriat, 1997) and is evident regardless of whether relatedness is manipulated between or within subjects (Dunlosky & Matvey, 2001). Relatedness appears to have such a substantial influence on JOLs that participants regard related items to be more memorable than unrelated items even when the opposite is true. For example, Caroll, Nelson, and Kirwan (1997) had participants overlearn unrelated pairs (studied to a criterion of eight correct recalls) relative to related pairs (studied to a criterion of two correct recalls). When tested after a 2 or 6 week retention interval, cued recall was superior for the unrelated compared with the related pairs. However, despite memory performance favoring unrelated items, JOLs were far greater for related items. Indeed, at the 6-week interval, JOLs were vastly more overconfident with respect to memory performance for related compared with unrelated pairs. Koriat and Bjork (2005; 2006a) have further shown that participants' JOLs are sensitive to relatedness even under circumstances that undermine memory. More subtle variations, such as the potential set of items related to a target, have less pronounced effects on JOLs (e.g., Eakin & Hertzog, 2012).

While such manipulations rely on the associative strength between items, JOL magnitude is also influenced by a variety of factors that are inherent to an item and not contingent on its relationship with other information. For example, participants provide higher JOLs for concrete (*donkey*) relative to abstract (*truth*) items, consistent with memory performance (Tauber & Rhodes, 2012a; see also Hertzog, Dunlosky, Robinson, & Kidder, 2003). Likewise, JOLs are generally higher for emotional words (e.g., Groninger, 1976; Tauber & Dunlosky, 2012; Zimmerman & Kelley, 2010) and faces (Nomi, Rhodes, & Cleary, 2013). Indeed, emotion appears to elevate JOLs independently of memory performance under some circumstances. For example, Zimmerman and Kelley (2010) had participants study pairs with negative target words (e.g., *prison-cancer*), positive target words (e.g., *table-fame*) and neutral target words (e.g., *violin-avenue*). Participants made a JOL for each item, predicting the likelihood of recalling the target (e.g., *avenue*) given the cue word (e.g., *violin*) and later attempted to recall the target given the cue. JOLs were similar for positive and negative word pairs and exceeded those for neutral word pairs. However, whereas JOLs for positive word pairs accurately predicted memory performance, JOLs for negative words were characterized by high levels of overconfidence (but see Tauber & Dunlosky, 2012, and Zimmerman and Kelley, 2010, for different results with free recall).



Click to view larger

Figure 4.1 Mean JOLs (*black bars*) and mean percentage recalled (*white bars*) by participants exposed to small (18 pt) and large (48 pt) words (adapted from Rhodes & Castel, 2008). Participants provided significantly higher JOLs for larger compared with small words. However, there was no difference in the percentage of words recalled as a function of the size of the word.

Note: JOL = judgment of learning.

Other research has demonstrated that changes in the appearance of to-be-remembered information may have a significant influence on memory predictions (Busey et al., 2000; Rhodes & Castel, 2008; Rhodes & Castel, 2009; Sungkhasettee, Friedman, & Castel, 2011; Yue, Castel, & Bjork, 2013). For example, Rhodes and Castel (2008) had participants study and make JOLs for a list words that varied in the size of the type used to present each item. Namely, half of the words were presented in a large type size (48 pt) and half were presented in a smaller type size (18 pt). The results are shown in Figure 4.1. Overall, participants consistently provided higher JOLs for large compared with small words, although no differences in memory performance were detected. Thus, large words results in greater levels of overconfidence regarding recall than small words. Rhodes and Castel (2009) reported an auditory analog to this finding, demonstrating that loud words were given higher JOLs than quieter (p. 69) words, even when memory performance was essentially identical. One method of linking these findings is to suggest that loud or large words are more easily perceived (more *fluent*) and thus, more fluent materials are accorded higher JOLs. However, this account is contradicted by findings that clarity does not always engender higher JOLs (Yue et al., 2013). Regardless, it is apparent that JOLs may be driven by factors that are specific to an item.

Encoding and Retrieval Influences on JOL Magnitude.

Ideally, an individual's memory predictions should be sensitive to conditions that bear on memory performance and largely immune to conditions that have minimal effects on memory. That is, if condition X is a boon to memory and condition Y a bane to memory, then JOLs should conform to this pattern, with the magnitude of JOLs for condition X exceeding those for condition Y. In general, the literature suggests a modest concordance with this ideal pattern but is also characterized by a number of instances in which individuals' JOLs appear insensitive or inadequately sensitive to factors that have considerable effects on memory.

As a case example of this mixture of acuity and insensitivity to conditions that influence memory performance, consider work by Shaw and Craik (1989). They had participants study nouns paired with one of three types of cues: a letter cue (e.g., "starts with ic: ice"), a rhyme cue (e.g., "rhymes with dice: ice"), or a category/descriptive cue (e.g., "something slippery: ice"). After studying each item, participants provided a JOL of the likelihood of later remembering the target when it was paired with the studied cue (e.g., "rhymes with dice: ?"). Several decades of work on depth of processing have indicated that the degree to which learners consider meaning during encoding is positively related to retention (Craik & Lockhart, 1972). Shaw and Craik's (1989) results were no different, as cued recall was reliably better for items associated with category information than rhyming information, which in turn resulted in superior recall than items cued based on letters. By extension, participants' memory predictions should, ideally, reflect this relationship, with JOL magnitude greatest for items considered in terms of categorical information and lowest for items considered in terms of orthographic information. To some extent this pattern of memory predictions was evident: JOLs were greater for category and rhyming items than items cued by the specification of letters. However, memory predictions did not differ for category versus rhyming items, despite a robust memory advantage for category items. Thus, participants exhibited partial sensitivity to conditions that influenced encoding (but see Bieman-Copland & Charness, 1994; Dunlosky & Nelson, 1994).

The broader literature is likewise a boom and bust cycle indicating that, with some notable exceptions, the magnitude of participants' JOLs often fails to distinguish between effective versus ineffective encoding activities. For example, two of the most powerful methods of enhancing memory are spacing and testing. Spacing (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for a review) refers to the memorial advantage that accrues when information (p. 70) is studied multiple times, but not consecutively (i.e., presentations of the same item are separated by at least one other item), compared with massing information (i.e., consecutive presentations of the same item). The testing effect (see Roediger & Butler, 2011, for a review) refers to the finding that retention is superior when learners engage in some form of retrieval of previously studied information relative to simply restudying (rereading) this information.

Both methods are highly effective mnemonic strategies but JOL magnitude does not consistently or adequately reflect these benefits. For example, Logan, Castel, Haber, and Viehman (2012; see also Zechmeister & Shaughnessy, 1980) had participants study a list of words in which some words were repeated immediately (massed study) and some were repeated after a lag of at least three items (spaced study). Participants made a JOL after each presentation and were later administered a free recall test. Logan et al. (2012) observed that participants' JOLs were marginally higher for spaced compared with massed items. However, this difference in JOLs (approximately 2 percentage points) was dwarfed by the memory advantage apparent for spaced items (approximately 16 percentage points). Indeed, participants' JOLs generally underestimated the robust benefits of spacing. A similar pattern is apparent for studies of the testing effect, whereby participants provide comparable JOLs for tested compared with restudied information despite testing benefits (King et al., 1980; Kornell & Rhodes, 2013; Roediger & Karpicke, 2006; but see also Tullis, Finley, & Benjamin, 2013). Other work has also demonstrated that the magnitude of participants' memory predictions may ignore factors that have a significant influence on memory. For example, participants' JOLs often fail to account for the ameliorative effect of additional study opportunities after an initial test (Kornell & Bjork, 2009; Kornell, Rhodes, Castel, & Tauber, 2011) and may not distinguish between long versus short retention intervals (Koriat, Bjork, Sheffer, & Bar, 2004), particularly if the interval is manipulated between-subjects.

However, it would be a gross mischaracterization to suggest that participant's are uniformly oblivious to factors that enhance or hinder memory. For example, the magnitude of participants' JOLs differentiates between recognition versus recall tests (Groninger, 1979; Thiede, 1996; Thiede & Dunlosky, 1994) and is sensitive to variations in the number of items to be learned (Tauber & Rhodes, 2010a). Participants' JOLs also often favor effective study techniques such as generation (e.g., Begg, Vinski, Frankovich, & Holgate, 1991; Castel, Rhodes, & Friedman, 2013) or interactive imagery (Dunlosky & Nelson, 1994) over less effective techniques. More important, JOL magnitude may be altered by the updated knowledge that accumulates from a recent prior learning experience (Bieman-Copland & Charness, 1994; Castel, 2008; Hertzog & Dunlosky, 2000; Koriat & Bjork, 2006a; Tauber & Rhodes, 2010b). For example, Castel (2008) observed that participants' JOLs closely corresponded with the effects of order (i.e., serial position) on recall following practice, coupled with salient information about order during encoding. Likewise, Tauber and Rhodes (2010b) reported that participants were better able to predict differences in memory for names versus occupations after a specific prior experience with learning names. Such practice does not guarantee that JOL magnitude will be adjusted appropriately (see e.g., Koriat, Sheffer, & Ma'ayan, 2002; Logan et al., 2012) but suggests that participants can update their knowledge of factors that influence memory under some circumstances.

Factors That Influence the Relative Accuracy of Judgments of Learning

As noted previously, relative accuracy refers to the degree to which an individual's JOLs distinguish between information that will or will not be remembered. In other words, relative accuracy describes the rank-ordering of JOLs with respect to memory performance. Accordingly, the ideal learner is presumed to provide higher JOLs for remembered information compared with information that will not be remembered. There remains some controversy over Nelson's (1984) proposal that the Kruskal-Goodman gamma correlation is the appropriate measure of relative accuracy (see e.g., Benjamin & Diaz, 2008; Masson & Rotello, 2009, for criticisms and alternative measures). However, given its predominance in research on JOLs, I discuss relative accuracy in reference to the ubiquitous gamma correlation. In particular, three factors with a common mechanism have a substantial influence on the relative accuracy of memory predictions.

Testing.

In addition to being a potent method of enhancing retention, the act of retrieval also produce strong levels of relative accuracy. King and colleagues (1980) first documented the benefits of testing for memory predictions, reporting that (p. 71) JOLs were far more accurate when participants experienced some test trials during learning compared with conditions that only involved studying. Subsequent work has confirmed that testing (e.g., Kornell & Rhodes, 2013) or conditions that foster retrieval (e.g., Dunlosky & Nelson, 1992; see JOL Timing for additional details) provide a significant boost to relative accuracy compared with conditions that only require individuals to review materials or discourage testing. One account of this finding is that participants use retrieval success as an index of future memory performance, allocating high JOLs to retrieved items and low JOLs to items that were not retrieved (Finn & Metcalfe, 2007; King et al., 1980). Given that current retrieval will often remain stable on a future test (but see Schmidt & Bjork, 1992), retrieval success leads to highly accurate predictions (cf. Dougherty et al., 2005).

Practice.

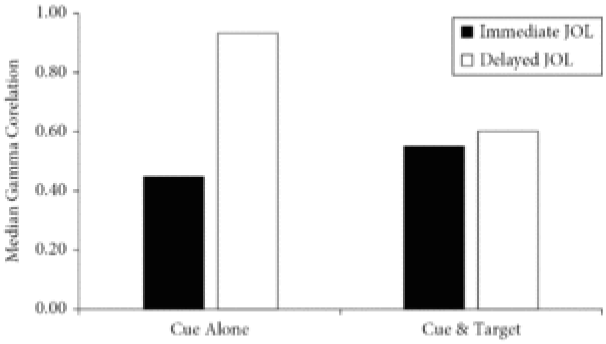
Relative accuracy appears to improve markedly when the same information is presented across multiple study-test cycles. The most striking example of this is the underconfidence-with-practice (UWP) effect, characterized by decrements to absolute accuracy and benefits to relative accuracy (Koriat et al., 2002). Specifically, across repeated study-test cycles with the same information, participants often exhibit overconfidence (i.e., memory predictions exceed performance) on a first study-test trial followed by underconfidence (i.e., memory performance exceeds predictions) on later study-test trials. In contrast, relative accuracy generally increases across trials. Tauber and Rhodes (2012b) report data emblematic of this pattern. They had participants make JOLs and receive a memory test for pairs of unrelated words in three study-test trials. Recall increased across trials, going from a mean of approximately 35% on Trial 1 to 72% by Trial 2, and 83% by Trial 3. JOLs initially exceeded memory performance on Trial 1 (50%) and then were unconfident for Trials 2 (58%) and 3 (76%). However, relative accuracy increased steadily from Trial 1 (Gamma = 0.23) to Trial 2 (Gamma = 0.64) to Trial 3 (Gamma = 0.82).

Such benefits to relative accuracy as a function of practice have been widely observed (e.g., Finn & Metcalfe, 2007, 2008; Koriat, 1997; Koriat et al., 2002; Koriat, Ma'ayan, Sheffer, & Bjork, 2006). Much like explanations of the metacognitive advantages of testing, the predominant account of this finding is that participants use success or failure of retrieval as a basis for JOLs (Finn & Metcalfe, 2007, 2008). For example, during a second study opportunity, low JOLs are likely to be assigned to items identified as forgotten on the first test, whereas higher JOLs are likely to be assigned to items identified as recalled on the first test. This will generally beget JOLs that are high in relative accuracy but may also underestimate new learning that occurs (e.g., some previously forgotten information may now be learned), leading to the underconfidence evident on later study-test cycles.

Thus, although other factors certainly inform multi-trial learning (Ariel & Dunlosky, 2011; England & Serra, 2012; Tauber & Rhodes, 2012b), retrieval is the predominant factor driving judgment accuracy. This also suggests an important caveat when considering the influence of practice on relative accuracy. Namely, when each study-test cycle contains new information, gamma correlations appear to remain largely consistent and do not improve across trials (e.g., Koriat & Bjork, 2006b; Matvey, Dunlosky, Shaw, Parks, & Hertzog, 2002; Tauber & Rhodes, 2010a). Given that introducing new items in each study-test cycle entails that participants cannot rely on past retrieval success or failure, their JOLs are less effective at differentiating items that will or will not be remembered. As a consequence, practice has little impact on relative accuracy under these circumstances.

Timing.

Following the procedure outlined by Arbuckle and Cuddy (1969), the first two decades of research on JOLs generally involved soliciting judgments immediately after an item was studied. However, Nelson and Dunlosky (1991; see also Begg, Duft, Lalonde, Melnick, & Sanvito, 1989) reported that varying the timing of JOLs had a dramatic influence on relative accuracy. They had participants study a list of 66 unrelated paired associates (e.g., *Table-Spoon*). For half of these items, JOLs were made immediately after an item's presentation, whereas for the remaining half of the items JOLs were delayed such that at least ten items intervened prior to making a JOL. Following this study phase participants received a cued recall test. Delaying JOLs led to a striking pattern, with gamma correlations far greater for delayed JOLs (Gamma = 0.90) compared with immediate JOLs (G = 0.38). The benefit of delaying JOLs was so substantial that Nelson and Dunlosky (1991) noted that, "Every subjects' accuracy on delayed JOL was greater than the mean of those same subjects' accuracy on immediate JOL" (p. 269). Nelson and Dunlosky coined (p. 72) their finding the *delayed JOL effect*. Subsequent work has generally confirmed the benefits of delaying judgment (e.g., Dunlosky & Nelson, 1992; 1994; 1997; Kelemen & Weaver, 1997; Koriat & Ma'ayan, 2005; Nelson, et al., 2004). For example, in their meta-analysis of 45 studies (112 effect sizes) of the delayed JOL effect, Rhodes and Tauber (2011a) observed that delaying JOLs was characterized by a nearly one standard deviation (g = 0.93) increase in gamma correlations compared with immediate JOLs.



Click to view larger

Figure 4.2 Median gamma correlations for immediate (*black bars*) and delayed (*white bars*) JOLs solicited by the cue alone (e.g., *Table-?*) and the cue and target (e.g., *Table-Spoon*). The timing of JOLs did not influence relative accuracy when JOLs were solicited via the cue and target. In contrast, relative accuracy was much greater for delayed compared with immediate JOLs when solicited via the cue alone (adapted from Dunlosky & Nelson, 1992).

Note: JOL = judgment of learning.

The predominant account of the delayed JOL effect suggests that delaying judgment encourages participants to attempt retrieval from long-term memory, with that information informing judgment (Nelson & Dunlosky, 1991; see also Dunlosky & Nelson, 1992, 1994; Nelson et al., 2004), akin to the previously discussed benefits of testing for relative accuracy. Accordingly, delayed JOLs are guided by the products of retrieval (Rhodes & Tauber, 2011b). In contrast, immediate JOLs are presumed to reflect information accessible in memory right after study that will be less diagnostic of future memory performance. Dunlosky and Nelson (1992) report compelling evidence regarding the importance of retrieval for the delayed JOL effect. They varied the cue used to solicit JOLs such that half of their participants provided JOLs solicited via a cue and target (e.g., *Table-Spoon*) and half provided JOLs prompted with the cue alone (e.g., *Table-?*). As can be seen in Figure 4.2, a robust delayed JOL advantage was evident for JOLs solicited by the cue alone but not with the cue and target. This presumably occurred because soliciting a JOL with the cue and target eliminates the opportunity to interrogate long-term memory and thus robs the learner of diagnostic information.

It should be noted that other theories dispute this account. For example, an alternative perspective is that when participants successfully retrieve a target during a delayed JOL they will likely ascribe a higher JOL to that item than to items that were not retrieved. Because successful retrieval provides a boost to memory, the act of making a delayed JOL enhances memory and ensures that the high JOL is accurate (Spellman & Bjork, 1992; see the section Influence of Judgment on Memory for additional details). Nevertheless, regardless of the explanation, delaying judgment appears to be one of the most robust methods of enhancing relative accuracy.

Theoretical Accounts of the Bases of Judgments of Learning

As has been documented, a vast array of data has been collected on the factors that do or do not influence JOL. But an overriding question remains: How do individuals make JOLs? Several theoretical approaches have been offered in answer to this question. These approaches can be broadly classified into two categories of theory comprising direct access accounts and inferential accounts (King et al., 1980; Koriat, 1997). The distinction reflects the amount and type of information the learner has available to make predictions.

(p. 73) Direct Access Accounts

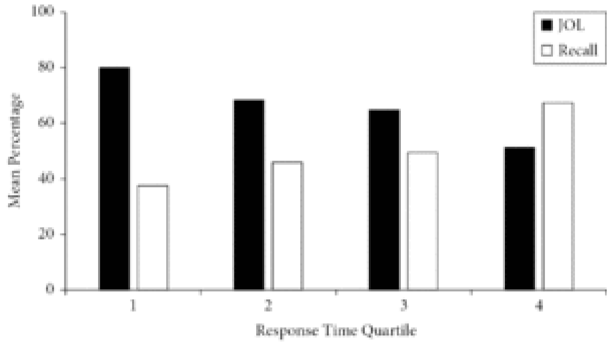
Arbuckle and Cuddy (1969) conducted their original experiment on JOLs under the premise that if "items differed in associative strengths immediately following presentation, [participants] should be able to detect these differences just as they can detect differences in strength of any other form of input signal" (p. 126). That is, participants should be able to make judgments by assessing the strength of a memory trace. A corollary of this account is that individuals have privileged access to the contents of memory that allows one to directly assess the efficacy of encoding and differentiate between items that have or have not been well-learned. Accordingly, direct access accounts propose that JOLs are a translation of the very contents (i.e., strength) of memory onto the scale used for judgment (Cohen, Sandler, & Keglevich, 1991).

The direct access view appears to have been met with skepticism even among early investigators. For example, King et al. (1980) noted that "Although this hypothesis [direct access] is intuitively appealing, we know of no evidence providing direct support for the hypothesis" (p. 340). King et al. did not articulate what standard of evidence would be necessary to evaluate a direct-access hypothesis, but a means of evaluation was suggested by Koriat (1997). In particular, Koriat notes that a direct-access hypothesis predicts a strong correspondence between memory and JOLs because both are putatively based on the same factor (trace strength). By this standard, it is difficult to sustain a direct-access account of JOLs given the sheer number of demonstrations of stark discrepancies between JOLs and memory performance (see e.g., the summary of research on absolute accuracy). As well, any correspondence between JOLs and memory performance is just as amenable to alternative explanations (e.g., individuals' knowledge about memory) as an account predicated on access to the strength of memory traces. Thus, despite several sophisticated attempts at understanding the interplay of memory strength and JOLs (e.g., Jang & Nelson, 2005), direct access accounts have rarely been favored.

Inferential Accounts

Rather than reflecting direct access to the contents of memory, inferential accounts hold that JOLs are based on a variety of information available during learning. This could include the type of information being studied (e.g., related vs. unrelated items), whether an item was recalled previously, the type of test to be administered, the rate at which information is presented, and so on. Koriat (1997) has proposed a *cue-utilization framework* to organize the many cues that might influence JOLs. His framework starts with the assumption that "JOLs are based on the implicit application of rules or heuristics in order to achieve a reasonable assessment of the probability that the information in question will be recalled or recognized at some later time" (p. 350). As such, "JOLs are accurate as long as the cues used at the time of making the judgments are consistent with the factors that affect subsequent performance on the criterion memory test" (p. 350). Thus, JOLs are inferences that are made based on the cues available during learning.

Koriat (1997) has specified three classes of cues that inform JOLs. *Intrinsic cues* refer to characteristics of the items to be learned that may influence (or are deemed to influence) learning. This includes perceptual characteristics of the items (e.g., size, clarity), associative relatedness, concreteness, emotional qualities of the stimuli, and essentially any other characteristics inherent to the item. *Extrinsic cues* refer to the conditions of encoding or testing and the processes applied by the learner. Examples include presentation rate (Groninger, 1976), spacing vs. massing study (Logan et al., 2012), recall vs. recognition tests (Thiede & Dunlosky, 1994), or the depth of processing applied to study items (Craik & Shaw, 1989; Dunlosky & Nelson, 1994). Whereas intrinsic and extrinsic cues refer to information external to the learner (e.g., the type of item and type of processing), *mnemonic cues* refer to internal indices that reflect the learner's memorial experience of an item. This might include how easily an answer comes to mind in response to a cue (Benjamin, Bjork, & Schwartz, 1998), memory for a prior test (e.g., Finn & Metcalfe, 2007), and the familiarity of a cue (Maki, 1999), among many other possibilities.



Click to view larger

Figure 4.3 Mean JOLs (*black bars*) and the mean percentage of previous answers recalled (*white bars*) as a function of response time quartile (adapted from Benjamin, Bjork, & Schwartz, 1998). Latencies to answer each question during the study phase were grouped into quartiles ranging from the fastest responses (Quartile 1) to the slowest responses (Quartile 4). The magnitude of JOLs was negatively related to response time, such that the highest JOLs were provided to the questions answered most quickly (Quartile 1). In contrast, the percentage of answers recalled was positively related to response latencies, such that the highest levels of recall were evident for the slowest quartile (Quartile 4).

Note: JOL = judgment of learning.

The cue-utilization framework can readily explain the many discrepancies between JOLs and memory performance that have been documented (e.g., Benjamin et al., 1998; Carroll et al., 1997; Kornell & Bjork, 2009; Rhodes & Castel, 2008). From this view, discrepancies arise because the cues used by the learner to inform JOLs are unrelated to actual memory performance. As illustrative of this approach, consider findings reported by Benjamin et al. (1998). They had participants answer a series of general knowledge questions (e.g., *What is the largest desert on earth?*). Benjamin et al. indexed the difficulty of each question as the latency of (p. 74) indicating that an answer was available, grouping each participant’s responses into quartiles based on the speed of responding. After answering the question, participants made a JOL of the likelihood of successfully engaging in free recall of the answer on a later test. The results are shown in Figure 4.3. As can be seen, the answers retrieved most quickly (Quartile 1) were given the highest JOLs; however, the opposite pattern was apparent for recall, as items with the longest latencies (Quartile 4) were most likely to be recalled. Thus, participants used a mnemonic cue (retrieval latency) as a basis for JOLs but inferred the wrong relationship between retrieval latency and free recall success. That is, although retrieval latency is likely to be a strong marker of success given a cue such as the original question posed, it is a poor basis for JOLs when those cues will be unavailable.

Such a framework is appealing in its explanatory power: Any pattern of JOLs can be explained as a match or mismatch of the cues available and the factors that drive memory. However, its status as a “framework” necessarily calls for greater precision in delineating and testing those cues that inform JOLs (see Dunlosky & Matvey, 2001).

Moreover, the mechanism of influence of any single cue is often ambiguous. For example, as described previously, Rhodes and Castel (2008) reported that participants deemed large words to be more memorable than small words (see Figure 4.1). They noted that one candidate explanation is that participants perceived larger words to be processed more easily (more fluently) than small words and thus misattributed this ease of processing to indicate a greater ease of retrieval on a later test. Thus, one might propose that a mnemonic cue (experienced ease of processing) drove judgment. However, there is a viable alternative. Specifically, participants may have a general belief that larger items are easier to remember than small items and applied that belief when making JOLs (Mueller, Dunlosky, Tauber, & Rhodes, 2014). This distinction captures a second important element of an inferential account. That is, in addition to using a variety of cues available, participants' JOLs may rely on their own mnemonic experience while processing an item (*experience-based judgments*) or use their general knowledge about learning and memory (*theory-based judgments*) to inform judgment (Kelley & Jacoby, 1996;[4] Koriat, 1997; Koriat et al., 2004).

With regard to experience-based judgments, a number of researchers have suggested ease processing at encoding or retrieval as a central cue that influences JOLs (Begg et al., 1989; Benjamin et al., 1998; Hertzog et al., 2000; Koriat & Ma'ayan, 2005; Rhodes & Castel, 2008; Undorf & Erfelder, 2011). For example, Hertzog et al. (2000) reported that JOLs were negatively related to the latency to form an image during encoding (i.e., higher JOLs were accorded items encoded quickly) whereas the (p. 75) ease of forming an image was unrelated to recall. Likewise, Benjamin et al. (1998) also observed a negative relationship between retrieval latency and JOLs (see Figure 4.3). Theory-based judgments are evident in demonstrations that JOLs frequently conform to general beliefs about memory. For example, participants' JOLs appear to reflect the belief that generated items are more memorable than items that have been read (Begg et al., 1991; Castel et al., 2013), that diagrams are easier to remember than text (Serra & Dunlosky, 2010), and that important information is highly memorable (Soderstrom & McCabe, 2011).

There is some evidence that experience may sometimes usurp belief (theory) even when a strong belief is in place. For example, Koriat et al. (2004) had participants study paired associates and asked them to make JOLs anticipating either a 5 minute, 1 day, or 1 week retention interval. Although individuals have a strong belief that the length of a retention interval is negatively related to memory performance, JOLs did not reflect this pattern when the retention interval was manipulated between subjects. Rather, participants provided similar JOLs regardless of the length of the retention interval (but see Tauber & Rhodes, 2012a). Kornell has similarly demonstrated that participants will often ignore the benefits of additional study-test opportunities when predicting future learning (e.g., Kornell & Bjork, 2009; Kornell et al., 2011).

At present, the circumstances under which JOLs are driven by experience, belief, or some combination of those influences remains poorly understood. Part of this reflects an often-untested assumption about whether JOLs are driven by belief or experience. As an example, consider the influence of relatedness on JOLs. This may indicate a general belief that related items (e.g., *Nurse-Doctor*) are easier to remember than unrelated items or it could reflect enhanced processing fluency emanating from two items that have a strong association. Mueller, Tauber, and Dunlosky (2013) attempted to address this issue by (a) evaluating beliefs about relatedness and (b) competitively assessing the influence of relatedness and a measure of processing fluency. To assess belief, Mueller et al. (2013) used Castel's (2008) prestudy JOL procedure, whereby participants make a JOL prior to seeing an item, rendering it impossible for the fluency of a particular item to drive judgment. Accordingly, participants made prestudy JOLs, cued only with an indication of whether the item was related or unrelated. Under those circumstances, participants' prestudy JOLs were far greater for related compared with unrelated items, suggesting a strong belief that related items are easier to remember than unrelated items. In another experiment, Mueller et al. had participants make lexical decision judgments (to assess ease of processing) on pairs of items and also solicited JOLs. Overall, JOLs were associated with relatedness (r = 0.66) and lexical decision times (r = –0.21). However, controlling for response latencies had no impact on the magnitude of the correlation between JOLs and relatedness. Thus, these data indicate that the influence of relatedness on JOLs is predominantly a function of belief rather than experience. Additional investigations of this nature will be necessary to more firmly understand how belief and experience contribute to JOLs.

Pending Issues and Future Directions

Although work on monitoring and JOLs is well beyond its nascent stage as a domain of research in metacognition, a number of issues remain to be resolved and necessitate more attention. For example, as noted in the previous section, the role of belief and experience in driving JOLs has still to be elucidated. In addition, JOLs have largely been confined to verbal materials, particularly the paired associate. In order to advance understanding of memory predictions, a larger variety of materials, tests, and encoding conditions remain to be explored (see e.g., Rhodes, Sitzman, & Rowland, 2013; Simon & Bjork, 2001; Tauber, Dunlosky, Rawson, Wahlheim, & Jacoby, 2013, for efforts in this vein). These are just two of a number of issues that merit attention and will drive future research in metacognition.

Methods of Soliciting Predictions

The modal method of soliciting predictions of future memory performance is to simply have participants rate, on a Likert or percentage scale, the likelihood that some bit of information will be remembered in the future. Recent work has sought to determine whether altering the method of judging future remembering influences the accuracy of predictions (e.g., Finn, 2008; Hanczakowski et al., 2013; McGillivray & Castel, 2011; Tauber & Rhodes, 2012a; McCabe & Soderstrom, 2011). For example, Finn (2008) compared typical JOLs with instructions to indicate the likelihood that information would be forgotten on a future test (termed *judgments of forgetting*). Judging forgetting did not affect relative accuracy but did influence absolute accuracy, making participants' judgments (p. 76) generally more conservative compared with JOLs (see also Koriat et al., 2004). McCabe and Soderstrom (2011) reported that asking participants to predict whether an item would be accompanied by contextual details at test improved relative accuracy compared with JOLs. Further, Tauber and Rhodes (2012a) asked participants to indicate how long (in minutes) information would be remembered. In contrast to work indicating that participants may deem information equally likely to be remembered in 1 week or 5 minutes (Koriat et al., 2004), Tauber and Rhodes' (2012a) participants provided modest predictions, indicating that information would be remembered, on average, for approximately 15 minutes.

Although intriguing, these data do not engender a clear interpretation of how the scale used affects judgment. For example, the method of soliciting judgment may alter response distributions without changing the underlying information that informs judgment (cf. Nelson, Leonesio, Landwehr, & Narens, 1986). Indeed, it is not apparent whether these alternative judgments induce participants to consider different kinds of information about memory or reflect a variation in judgment based on the same information (Dunlosky & Tauber, 2013). For example, Serra and England (2012) demonstrated that soliciting judgments of forgetting may lead participants to use a different anchor (i.e., intial value for considering judgments) than JOLs but does not represent a qualitative change in the type of information considered. Similarly, asking for estimates of the amount of time information will be remembered has no impact on relative accuracy compared with standard JOLs, suggesting that discrimination is not affected by that judgment (Tauber & Rhodes, 2012a). However, there is some indication that altering the method of soliciting judgment may change the information under consideration, potentially ameliorating predictions. As noted, McCabe and Soderstrom (2011) reported that asking about future memory states improved relative accuracy. Further, Soderstrom and Rhodes (2014) have shown that such predictions may diminish a potent metacognitive illusion. Future work should continue to refine methods of soliciting JOLs in light of whether or not the framing of a judgment changes the information under consideration.

Influence of Judgment on Memory

In their original report on JOLs, Arbuckle and Cuddy (1969) included a condition that did not provide JOLs during the learning phase in order to determine whether JOLs interfered with learning. Their results indicated precisely the opposite: Participants required to make JOLs exhibited significantly better recall than participants who did not make JOLs. Thus, the act of making a JOL altered the phenomenon under investigation, enhancing memory for the information being judged (Spellman & Bjork, 1992). Surprisingly, relatively few studies also report a condition that assesses the potential impact of JOLs on memory performance (e.g., Benjamin et al., 1998; Dougherty et al. 2005; Kelemen & Weaver, 1997; King et al., 1980; Sommer, Leuthold, & Schweinberger, 1995; Sundqvist, Todorov, Kubik, & Jonssön, 2012; Tauber & Rhodes, 2012a). King et al. (1980) varied whether participants received a prediction trial or an additional study trial. They found that both conditions resulted in similar levels of memory performance leading them to conclude that "performing the prediction task was comparable to having an additional study trial" (King et al., p. 336). Likewise, Sommer et al. (1995) observed superior face recognition for a condition that made JOLs compared with a condition that did not make JOLs. Others have reported no effect (Benjamin et al., 1998; Tauber & Rhodes, 2012b) or mixed effects of JOLs on memory. For example, Keleman and Weaver (1997) observed no effect of immediate JOLs on retention but did report a memorial benefit following delayed JOLs (see also Sundqvist et al., 2012).

The potential for reactivity (i.e., soliciting JOLs alters memory performance) suggests an agenda for future research (a) to include appropriate control conditions to assess the impact of prediction on memory and (b) to provide a viable explanation of such reactivity. One possible explanation is that JOLs simply increase the amount of exposure time for a particular item and have the same benefits that would accrue with additional study time (King et al., 1980). However, JOLs might also encourage additional encoding operations that provide a unique advantage even when controlling for study time. For example, Dougherty and colleagues (2005) used the PRAM procedure (Nelson et al., 2004), asking participants to make a recall attempt prior to judging either their confidence that a studied item was recalled or making a JOL concerning the likelihood of future recall. JOLs led to higher levels of recall on a later test than a no-judgment condition but also compared to participants who made retrospective confidence judgments (RCJs). (p. 77) In explaining this finding, Dougherty et al. (2005) suggested that

> … the requirement to make JOLs actually alters how well participants are learning the to-be-recalled items at study. Perhaps participants who make JOLs implement a more effective study strategy than participants making RCJs because the JOL task forces them to focus on future retrieval. Interestingly, the finding that there was no effect of making RCJs on any of the recall measures suggests that there is something special about making JOLs that improves learning.
>
> (p. 1110)

Future research will profit by examining whether there is indeed something special for learning that results from making JOLs.

Multiple Cues to Judgment

Much of the JOL literature focuses on specific cues in isolation in order to determine whether that cue influences judgment. Although this reveals important determinants of JOLs the approach falls short of enhancing understanding of the type of learning environments experienced by individuals such as the example student highlighted at the beginning of the chapter. In those settings, predictions of learning are likely driven by multiple cues. Accordingly, it is essential to achieve an understanding of how multiple cues function collectively to inform judgment (Ariel & Dunlosky, 2011; Tauber & Rhodes, 2012b).

As an example of the importance of examining multiple sources of information in forming JOLs, consider advising a student that predictions of learning will be most accurate when they are delayed. This advice is generally sound but ignores the fact that the benefits of delaying judgment hinge on the type of cue used to solicit judgment. Namely, delayed JOLs accompanied by the cue and the target are little more accurate than immediate JOLs (Dunlosky & Nelson, 1992). It is only when JOLs are solicited in a manner that permits unfettered retrieval from long-term memory (i.e., cue-only JOLs) that the advantage of delayed judgment is evident (see Figure 4.2). Other work does not suggest such interactions but points to independence among cues. For example, although participants believe that important information is more likely to be remembered, manipulating the value of information does not alter the impact of relatedness on JOLs (Soderstrom & McCabe, 2011). Further, relatedness appears to drive judgment and dwarf the impact of other cues, such as the size of words presented for study (Rhodes & Castel, 2008). Regardless of the nature of the relationship, identifying how multiple sources of information interact or sum to contribute to judgment will be necessary for a full account of the bases and nature of JOLs.

A Few Final Thoughts

The first four and a half decades of research on JOLs have yielded a number of insights into how individuals judge their future learning and have provided a rich array of data to inform theory. Advances over the past 20 years have been rapid and characterized by an increased understanding of the information that drives memory predictions but also a greater appreciation of the complexities of assessing the future state of memory. Although work on JOLs has generated myriad findings, several core themes can be extracted.

1. Individuals are unlikely to have privileged access to the contents and strength of memory. Instead, predictions rely on a host of cues that vary in their diagnosticity.
2. JOLs are the product of a combination of elements from the online experience of learning and general beliefs about memory. The proportional contribution of each to judgment remains an open question.
3. Orienting individuals toward judgments that rely on retrieval after a delay is one of the soundest means to ensure that a learner can discriminate between what has been learned and what has not been learned.

These themes will continue to be refined and future work should yield additional insights into how individuals predict learning.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, 81, 126–131.

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgments-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition*, 39, 171–184.

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, 610–632.

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition*, 19, 487–497.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, 127, 55–68.

Benjamin, A. S., & Diaz, M. (2008). Measure of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 73–94). New York, NY: Psychology Press.

Bieman-Copland, S., & Charness, N. (1994). Memory knowledge and memory monitoring in adulthood. *Psychology and Aging*, 9, 287–302.

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, 26–48.

Carroll, M., Nelson, T. O., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and long-term retention. *Acta Psychologica*, 95, 239–253.

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, 429–437.

Castel, A. D., McCabe, D. P., & Roediger, H. L. III. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review*, 14, 107–111.

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition*, 41, 28–35.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132, 354–380.

Cohen, R. L., Sandler, S. P., & Keglevich, L. (1991). The failure of memory monitoring in a free recall task. *Canadian Journal of Psychology*, 45, 523–538.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging*, 12, 50–71.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684.

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic-extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1180–1191.

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, 374–380.

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, 545–565.

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language*, 36, 34–49.

Dunlosky, J., & Tauber, S. K. (2013). Understanding people's metacognitive judgments: An isomechanism framework and its implications for applied and theoretical research. In T. Perfect & S. Lindsay (Eds.) *Handbook of Applied Memory*. Thousand Oaks, CA: Sage.

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition*, 33, 1096–1115.

Eakin, D. K., & Hertzog, C. (2012). Immediate judgments of learning are insensitive to implicit interference effects at retrieval. *Memory and Cognition*, 40, 8–18.

England, B. D., & Serra, M. J. (2012). The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review*, 19, 715–722.

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition*, 36, 813–821.

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, 238–244.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, 58, 19–34.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34, 906–911.

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive Psychology*, 1, 324–340.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in measures that contain tied scores. *Psychological Bulletin*, 119, 159–165.

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

Groninger, L. D. (1979). Predicting recall: The” feeling-that-I-will-know” phenomenon. *The American Journal of Psychology*, 92, 45–58.

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language*, 69, 429–444.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*, 208-216.

Hertzog, C., & Dunlosky, J. (2000). Updating knowledge about encoding strategies: A componential analysis of learning (p. 79) about strategy effectiveness from task experience. *Psychology and Aging*, 15, 462–474.

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, 22–34.

Jang, Y., & Nelson, T. O. (2005). How many dimensions underlie judgments of learning and recall? Evidence from state-trace methodology. *Journal of Experimental Psychology: General*, 134, 308–326.

Kelemen, W. L., & Weaver, C. (1997). Enhanced memory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 1394–1409.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory & Language*, 35, 157–175.

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

Koriat, A. (1997). Monitoring one’s own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126, 349–370.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one’s knowledge during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 31, 187–194.

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, 959–972.

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, 1133–1145.

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, 643–656.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, 52, 478–492.

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*, 147-162.

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, 595–608.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, 138, 449–468.

Kornell, N., & Rhodes, M. G. (2013). Feedback reduces the metacognitive benefit of tests. *Journal of Experimental Psychology: Applied*, 19, 1–13.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and stability bias: Dissociating memory, memory beliefs, and memory judgment. *Psychological Science*, 22, 787–794.

Levin, J. R., Yussen, S. R., Pressley, M., & de Rose, T. M. (1977). Developmental changes in assessing recall and recognition memory capacity. *Developmental Psychology*, 13, 608–615.

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, 464–470.

Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: The role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacognition and Learning*, 7, 175–195.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 756–766.

Maki, R. H. (1999). The roles of competition, target accessibility, and cue familiarity in metamemory for word pairs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1011–1023.

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, 509–527.

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, 589–597.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study time allocation? *Memory & Cognition*, 18, 196–204.

McCabe, D. P., & Soderstrom, N. C. (2011). Recollection-based metamemory judgments are more accurate than those based on confidence: Judgments of Remembering and Knowing (JORKs). *Journal of Experimental Psychology: General*, 140, 605–621.

McGillivray, S., & Castel, A. D. (2011). Betting on memory leads to metacognitive improvement by younger and older adults. *Psychology and Aging*, 26, 137–142.

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

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, 378–384.

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.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 84, 93–116.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, 51, 102–116.

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, 267–270.

Nelson, T. O., Leonesio, R. J., Landwehr, R. S., & Narens, L. (1986). A comparison of three predictors of an individual's memory performance: The individual's feeling of knowing versus the normative feeling of knowing versus base-rate item difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 279–287

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.), **(p. 80)** *The psychology of learning and motivation* (pp. 125–173). New York, NY: Academic Press.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A Revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods*, 9, 53–69.

Nomi, J. S., Rhodes, M. G., & Cleary, A. M. (2013). Effect of emotional facial expressions on predictions of identity recognition. *Cognition and Emotion*, 27, 141–149

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

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review*, 16, 550–554.

Rhodes, M. G., Sitzman, D. M., & Rowland, C. A. (2013). Monitoring and control of learning own-race and other-race faces. *Applied Cognitive Psychology*, *27*, 553–563.

Rhodes, M. G., & Tauber, S. K. (2011a). The influence of delaying judgments of learning (JOLs) on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, 137, 131–148.

Rhodes, M. G., & Tauber, S. K. (2011b) Eliminating the delayed JOL effect: The influence of the veracity of retrieved information on metacognitive accuracy. *Memory*, 19, 853–870.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science*, 15, 20–27.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249–255.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science* 3, 207–217.

Schwartz, B. L., & Efklides, A. (2012). Metamemory and memory efficiency: Implications for student learning. *Journal of Applied Research in Memory and Cognition*, 1, 145–151.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgments reflect the belief that diagrams improve learning from text. *Memory*, 18, 698–711.

Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgments of remembering and forgetting. *Quarterly Journal of Experimental Psychology*, 65, 2231–2257.

Shaw, R. J., & Craik, F. I. M. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging*, 4, 131–135.

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 907–912.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, 5, 315–316.

Soderstrom, N. C., & McCabe, D. P. (2011). The interplay between value and relatedness as bases for metacognitive monitoring and control: Evidence for agenda-based monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1236–1242.

Soderstrom, N. C., & Rhodes, M. G. (2014) Metacognitive illusions can be reduced by monitoring recollection during study. *Journal of Cognitive Psychology*, 26, 118–126.

Sommer, W., Heinz, A., Leuthold, H., Matt, J., & Schweinberger, S. R. (1995). Metamemory, distinctiveness, and event-related potentials in recognition memory for faces *Memory & Cognition*, 23, 1–11.

Sundqvist, M. L., Todorov, I., Kubik, V., & Jönsson, F. U. (2012). Study for now, but judge for later: Delayed judgments of learning promote long-term retention. *Scandinavian Journal of Psychology*, 53, 450–454.

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, 973–978.

Tauber, S. K., & Dunlosky, J. (2012). Can older adults judge their learning of emotional information? *Psychology and Aging*, 27, 924–933.

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychonomic Bulletin & Review*, 20, 356–363.

Tauber, S. K., & Rhodes, M. G. (2010a). Are judgments of learning (JOLs) sensitive to the amount of material to-be-remembered? *Memory*, 18, 351–362.

Tauber, S. K., & Rhodes, M. G. (2010b). Metacognitive errors contribute to the difficulty in remembering proper names. *Memory*, 18, 522–532.

Tauber, S. K., & Rhodes M. G. (2012a). Measuring memory monitoring with judgments of retention interval (JOR). *Quarterly Journal of Experimental Psychology*, 65, 1376–1396.

Tauber, S. K., & Rhodes M. G. (2012b). Multiple bases for young and older adults' judgments-of-learning (JOLs) in multitrial learning. *Psychology and Aging*, 27, 474–483.

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance *The Quarterly Journal of Experimental Psychology*, 49A, 901–918.

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology*, 86, 290–302.

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, 429–442.

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, 1264–1269.

Vesonder, G. T., & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language*, 24, 363–376.

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition*, 41, 229–241.

Yussen, S. R., & Levy, V. M. (1975). Developmental changes in predicting one's own span of short-term memory. *Journal of Experimental Child Psychology*, 19, 502–508.

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, 41–44.

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, 240–253.

Notes:

(1.) This chapter focuses exclusively on item-by-item judgments of learning. Another approach is to solicit aggregate, or global, predictions of memory performance, such as the total number of items that will be remembered (e.g., Connor, Dunlosky, & Hertzog, 1997).

(2.) For example, in their meta-analysis of JOLs elicited following a delay compared with immediately after study, Rhodes and Tauber (2011a) observed that 103 of 112 (92%) effect sizes were for JOLs solicited on a percentage scale.

(3.) It is important to note that the distinction between item-based influences and those present in the conditions of encoding or testing is necessarily imprecise and serves only as a rough organizing framework for considering the wider literature on JOLs. For example, the manner in which an item is processed will be an interactive function of the learner and the information to-be-learned (e.g., consider how individuals with or without knowledge of English would process the word *allergy*). In addition, the conditions of encoding and/or testing may alter how an item is perceived, as when the same items are studied repeatedly. Thus, although these factors are not mutually exclusive, they are treated separately to facilitate exposition.

(4.) Kelley and Jacoby (1996) used somewhat different terms, distinguishing between subjective and analytic bases for judgment.

Matthew G. Rhodes

Matthew G. Rhodes, Department of Psychology, Colorado State University