Introspective observations have been criticized as being unverifiable, and for a period in the middle of the twentieth century were largely eschewed by the field of psychology. Since the 1970s, experimental paradigms have been developed to examine the accuracy of introspections. Unlike previous introspective methods, these new methods of examining metacognitive accuracy called for specific judgments about performance that could later be compared with criterion measures.

Metacognitive judgments can be tested with regard to their accuracy at predicting memory performance. Research on metacognitive judgments is of interest because it informs us of when the monitoring of memory is accurate and of when the output from monitoring serves as input to control processes (e.g. rehearsal) that affect subsequent memory.

INTROSPECTION

Introspective judgments were investigated in the early years of psychology by researchers such as Wundt and Titchner. Their style of investigation involved observations and reports about tiny ‘slices’ of consciousness, with extended reports about a single introspective slice.

In the early twentieth century, introspection was attacked on two fronts. Firstly, Freud’s theory of the unconscious postulated that not all mental activity was available to introspection, and that unconscious activity affected behavior. Secondly, radical behaviorists such as Watson and Skinner considered introspection to be irrelevant to the understanding of behavior, arguing that there were no reliably valid introspections. Because of these (and other) problems, the study of introspection fell out of favor, and (except in the field of perception) had little influence in psychology until the 1960s.

With the advent of cognitive psychology in the 1960s, introspection was revived. However, instead of assuming that introspections were accurate, cognitive psychologists compared them to relevant behavior. Such comparisons revealed situations in which metacognitive judgments have above-chance accuracy. For instance, investigators became interested in situations in which metacognitive judgments were very accurate (e.g. Nelson and Dunlosky, 1991), as compared to other situations in which they were mostly inaccurate, and an attempt was made to treat those findings as clues about the underlying mechanisms of metacognition.

FEELINGS OF KNOWING
Hart (1965) was one of the first researchers to investigate the accuracy of introspections about human memory. The ‘feeling of knowing’ (FOK) is an experience in which one has a feeling that a currently unretrieved item is nevertheless in memory. Hart investigated whether FOKs are accurate at predicting subsequent memory performance. He used the recall-judgment-recognition paradigm in which subjects receive a recall test, often consisting of general knowledge questions such as ‘What is the capital of Australia?’ or ‘What star is called the North Star?’ For answers recalled incorrectly, or when no answer is produced, subjects are asked to report their FOK by indicating the likelihood that they would recognize the unrecalled answer.

Hart’s experiments showed that a positive FOK was associated with higher probability of recognizing the correct answer than was a negative FOK, indicating that FOK judgments are at least somewhat accurate. Hart showed that subjects should be encouraged to guess on the recall test, even when they do not think they know the answer. This is important because when subjects are unsure of an answer, they may be reluctant to guess. Imagine a subject who retrieves an answer for a particular question, but fails to give the answer because he or she fears that it is incorrect (i.e. has low confidence in it). That item is scored as an omission, and thus will be given an FOK rating and included on the recognition test. The subject will probably give the item a high FOK rating, because although no answer was reported, one did come to mind, indicating that the subject has at least some (perhaps correct) knowledge about the question. Subsequently, if the unreported answer was correct, it will appear on the recognition test, and the subject will almost certainly recognize it as the correct answer. This will exaggerate FOK accuracy, because accuracy depends on the ability of subjects to use FOK judgments to discriminate between items that will be correctly recognized and those that will not. The item in question will probably be recognized, adding to the number of items given high FOKs that were subsequently correctly recognized.

Procedures for Investigating FOK Judgments

General knowledge questions are not the only kind of questions that have been used to study FOKs. Noun–noun paired associates (e.g. ‘OCEAN – TREE’) have also been used to investigate FOK judgments. In this procedure, subjects study a list of items, and during the recall test are asked to recall the second word when prompted with the first. As with general knowledge questions, subjects are asked to make FOK judgments for items that are not correctly recalled, and then a recognition test occurs for those items.

The difference between general knowledge questions and paired-associate learning is relevant. People are able not only to accurately monitor items that have been stored in memory for a relatively long time, but under some circumstances (Nelson and Narens, 1990) can also monitor items that are recently acquired.

Basis of FOK Judgments

Some researchers have investigated the factors on which FOK judgments are based. For example, Nelson et al. (1982) investigated whether the degree of learning affects the magnitude of FOK. Subjects learned paired associates to a criterion of one, two, or four correct recalls, and received a recall test four weeks later. Items not recalled during the test were ranked by the subject according to the likelihood that they would be correctly recognized. The results showed that subjects’ FOK ranks increased with degree of learning, indicating that this is one factor on which FOK judgments may be based.
Reder and Ritter (1992) investigated the basis of very rapid FOK judgments that occurred prior to extended attempts at recall. Subjects were asked to solve mathematics problems. Some of the previously solved problems were presented again, while other problems consisted of the components of previously solved problems but with at least one critical change. For example, the subject may have previously solved the problem ‘6×19’. Subsequently, the 6 and 19 might be presented again, but with a different operand (e.g. ‘6 +19’). Upon seeing the problem, subjects were asked to make an FOK judgment, estimating whether they could recall the answer or whether they would have to compute it anew. The results showed that higher FOKs were associated with familiar components of a problem even when the operand was changed, indicating that FOK judgments may be based partially on familiarity with the cue rather than on the subject’s retrieval of the answer.

It is important to discover the factors that affect FOK accuracy in order to determine whether FOK judgments are monitoring unrecalled answers directly or whether they are monitoring information available from external cues and recalled portions of answers that are diagnostic of subsequent memory performance.

**JUDGMENTS OF LEARNING AND ALLOCATION OF STUDY**

Students studying for an examination make judgments about whether various facts have been learned sufficiently to be recalled in the exam or whether further study is required. This phenomenon is known as ‘judgment of learning’ (JOL). Judgment of learning is a kind of metacognitive monitoring process. The students also make decisions about which items to continue studying. This is known as ‘allocation of study’ and is a kind of metacognitive control process.

**Paired Associates and JOL**

Participants in a typical JOL experiment learn a list of items. The experimenter is able to control the duration of study, number of repetitions, order of presentation, and other factors that may affect performance. Paired associates are particularly well suited to the study of JOLs because they provide stimuli for a cued-recall test, which is useful for controlling effects caused by the order of recall.

Only the stimulus item of the pair should be present at the time of the judgment. Presence of the partial or entire response word at the time of the judgment has been shown to reduce the accuracy of the JOL in predicting recall (Dunlosky and Nelson, 1992).

**Procedures for Investigating JOLs**

After studying a given item, subjects may be prompted to make a JOL by giving a percentage confidence judgment that in about ten minutes they will be able to recall the second word of the pair when prompted with the first. Subjects make this judgment, known as an ‘individual-item’ JOL, for each pair.

An ‘aggregate’ JOL may also be made for the entire list. Here, subjects are asked to estimate how many of the items they will be able to recall in the test.

**Kinds of Accuracy**

Item-by-item JOLs can be used to investigate ‘relative accuracy’, which is the ability of people to distinguish items that will be correctly recalled at test from those that will not. For example, imagine that a person has studied two pairs, ‘OCEAN – TREE’ and ‘DAFFODIL – BLOOD’, and then made JOLs for both pairs, assigning the first pair a JOL of 80% and the second pair a JOL of 20%. If at test the
person recalls ‘TREE’ when prompted with ‘OCEAN’ and fails to recall ‘BLOOD’ when prompted with ‘DAFFODIL’, the person can be said to have been accurate insofar as the item that received the greater JOL was also the item that had the better outcome during recall.

Another type of accuracy is ‘absolute accuracy’, which can be measured both for individual item JOLs and for aggregate JOLs. Absolute accuracy refers to the extent to which the cardinal value of the JOL corresponds to the percentage of correct recall. For example, if the person has studied and made item-by-item JOLs, and then, at test, recalls none of the items that had received a JOL of 0%, 20% of the items that had received a JOL of 20%, and 40% of the items that had received a JOL of 40%, then this person can be said to have perfect absolute accuracy. Likewise, for aggregate JOLs, absolute accuracy is the degree to which the aggregate JOL matches the percentage of recall.

Production of Accurate JOLs

The student studying for an examination will be interested in any strategy that will make JOLs more accurate, in order to know which items need further study and which are already learned well enough to be recalled later. Although individual-item JOLs made immediately after study are generally above-chance at predicting subsequent retention, they are far from perfectly accurate. JOLs have been shown to be very accurate when they are made at least 30 seconds after study (Nelson and Dunlosky, 1991). This is known as the ‘delayed-JOL effect’. Although 30 seconds of filled activity is sufficient to produce a substantial increase in JOL accuracy, judgments made after a longer delay (e.g. five minutes) may be even more accurate (Kelemen and Weaver, 1997).

Monitoring and Control Processes

It is important to understand the way in which information acquired during metacognitive monitoring processes is used for metacognitive control. For instance, some researchers have investigated the interplay between JOLs and the allocation of subsequent study. In an experiment by Nelson et al. (1994), subjects studied, and made JOLs for, Swahili-English equivalents (e.g. ‘ARDHI - SOIL’). Each subject then studied again either the items receiving the highest JOLs (from that subject) or those receiving the lowest JOLs (from that subject). Further study improved multi-trial learning more when it was devoted to the items that had received the lowest JOLs (but for boundary conditions, see Metcalfe and Son, 2000).

RETROSPECTIVE CONFIDENCE JUDGMENTS

Students leaving an examination may think about the questions, judging which of them were answered correctly. Such judgments are called ‘retrospective confidence judgments’ (RCJs). They may relate to individual items or to the aggregate (for example, an estimate of the percentage of answers that were correct).

Procedures for Investigating RCJs

In the laboratory, the procedures for investigating RCJs are similar to those for investigating other metacognitive judgments. Subjects are shown the cue for the judgment, which consists of a question (in the case of general knowledge questions) or the first word of a pair (in the case of paired associates). Sometimes the subject’s answer is displayed along with the cue. Treadwell and Nelson (1996) showed that the accuracy of aggregate RCJs is not affected by the amount of information provided in the prompt for the judgment (cue alone, cue with subject’s answer, or cue with answer and the list of choices). This is an important difference from JOLs, whose accuracy is affected by the
content of the cue (as discussed above). Individual-item or aggregate judgments are elicited as described previously.

**Accuracy of RCJs**

Koriat et al. (1980) asked subjects to make RCJs for general knowledge questions. They found that when subjects were asked to give reasons for why their answer may be incorrect, the subjects gave more accurate RCJs than when they were asked to give reasons for why their answer was correct or when they were not asked to give any reasons. This finding suggests that subjects may search automatically for reasons for why their answers are correct – a phenomenon known as ‘confirmation bias’ – whereas searching for reasons why the answer may be wrong is not automatic, and must be done deliberately in order to increase the accuracy of RCJs.

**FEELING OF WARMTH IN PROBLEM SOLVING**

In an investigation of the feeling-of-warmth phenomenon, Metcalfe and Wiebe (1987) proposed two broad categories of problems, namely incremental problems and insight problems. Incremental problems are solved by degrees, with each step taking the person a little closer to the solution. Metcalfe and Wiebe predicted that people will make incremental feeling-of-warmth judgments for these types of problems, with ratings increasing as steps in the problem are completed. By contrast, insight problems are typically solved suddenly. Feeling-of-warmth judgments for insight problems were predicted to be fairly low and constant until the problem was solved, at which point they would increase suddenly. Metcalfe and Wiebe had subjects solve either incremental or insight problems, making feeling-of-warmth judgments every 15 seconds. The results confirmed the predicted patterns. These results suggest that the feeling of warmth is accurate only for incremental problems, and not for insight problems. With insight problems, a low feeling of warmth does not necessarily indicate that the problem is difficult or even that the person is far from the solution.

**SUMMARY**

Metacognitive judgments reflect the monitoring of one’s own memory when they occur as predictions about subsequent performance on studied items (as in JOLs), as predictions about subsequent retrieval of currently unretrieved items (as in FOKs), as predictions of the accuracy of answers given on a test (as in RCJs), and as predictions of the imminence of solving problems (as in feeling-of-warmth judgments). In some circumstances, as in the case of delayed JOLs, metacognitive judgments are highly accurate; in other circumstances, as in the cases of immediate JOLs and feelings of warmth about insight problems, they are less accurate. It is important to know when metacognitive judgments are accurate, so that we may use them to control our cognitive processing effectively, and to know when they are inaccurate, so that we may discover ways to improve them.

**References**

Psychology: Learning, Memory and Cognition 6: 107-118.


Further Reading
