Chapter 13

The subjective confidence in one's knowledge and judgements: some metatheoretical considerations

Asher Koriat

The subjective certainty in one's own knowledge

Examination of the history of early Greek philosophy reveals a shift from preoccupation with ontological questions to preoccupation with epistemological questions (Burnet 1930). Pre-Socratic Greek philosophy began by asking ontological questions—questions about the nature of the universe: what does it consist of? What is its origin? Is it infinite? When disagreements broke out, attention shifted inwards, to epistemological questions about the nature of knowledge itself: what is knowledge? How do we know? How can we be certain about our own knowledge? These questions are at the heart of present-day epistemology as well as cognitive psychology. The tension between ontological and epistemological perspectives—between asking questions about what is out there and asking questions about how we know what is out there—is not only the province of philosophy or psychology; it is today the province of modern physics as well. Some of the discussions in modern physics raise the question of whether the processes by which we acquire knowledge about what is out there will ever allow us to reach definite conclusions.

Questions about truth and its justification have also concerned statisticians who examined these questions from a normative perspective, focusing on the degree of confidence in conclusions that are based on empirical observations. These questions have been important in many applied areas as well, such as jury decisions and medical diagnosis (Dunning et al. 2004). In addition, in many real-life situations, confidence in one's judgements determines the likelihood of translating these judgements to action (Tversky and Koehler 1994; Koriat and Goldsmith 1996; Dunning 2007).

In experimental research, assessments of subjective confidence in one's own knowledge and judgements have been investigated over many years in a wide range of domains. These include perception and psychophysics, memory and metacognition, judgement and decision-making, and eyewitness testimony. Increased interest in confidence judgements can also be seen in such areas as social cognition, animal cognition, and neuroscience.

Two general issues have been addressed by researchers: the accuracy of metacognitive judgements and the bases of these judgements. With regard to the accuracy of metacognitive judgements, the observation that has attracted the attention of researchers in metacognition is that participants are generally accurate in monitoring their knowledge: They can tell when they know and when they do not know; and can judge when they are right and when they are wrong. For example, when studying a list of items, participants can predict with some accuracy which items they will recall at test (Nelson and Dunlosky 1991). During recall too, people can predict with some success which of the unrecallable memory targets they will be able to recognize among

distracters (Koriat 1993). In addition, when they are asked to answer each of several questions, participants can generally discriminate between correct and wrong answers (Goldsmith and Koriat 2008). The ability to monitor one's own knowledge was seen by Tulving and Madigan (1970) as 'one of the truly unique characteristics of human memory' (p. 477). This ability raises the question: how do people know that they know?

To answer this question, we must first examine the bases of metacognitive judgements. Understanding the bases of one's metacognitive judgements may provide a clue to both the accuracy and inaccuracy of people's knowledge of their own knowledge.

The bases of metacognitive judgements

Three general approaches to the bases of metacognitive judgements may be distinguished: the *direct-access approach*, the *information-based approach*, and the *experience-based approach* (see Koriat 2007). The direct-access view is perhaps best represented in the philosophy of knowledge, by the claims of rationalist philosophers that a priori truths (e.g. mathematical propositions) are based on intuition and deduction, and that their certainty is self-evident. In memory research, the direct-access approach assumes that metacognitive judgements are based on people's privileged access to the presence and strength of stored memory traces (see Dunlosky and Metcalfe 2009). For example, it was proposed that judgements of learning (JOLs) are based on detecting the strength of the memory trace that is formed following learning (e.g. Cohen et al. 1991). Similarly, tip-of-the-tongue (TOT) and feeling-of-knowing (FOK) judgements were claimed to monitor the actual presence of the elusive target in the memory store (Hart 1965; Burke et al. 1991; Yaniv and Meyer 1987). In the case of confidence judgements too, a direct access view generally underlies the use of such judgements in the context of strength theories of memory (see Van Zandt 2000).

In contrast to the direct-access view, a cue-utilization view has been gaining popularity in metacognition research (see Koriat 1997). According to this view, metacognitive judgements are inferential in nature, relying on a variety of beliefs and heuristics. A distinction is drawn, however, between information-based and experience-based judgements (Koriat et al. 2008). In information-based approaches, metacognitive judgements are assumed to rely on an analytic inference in which various considerations retrieved from long-term memory are consulted and weighed to reach an educated metacognitive judgement. For example, JOLs have been claimed to rely on the person's theories about how various characteristics of the study material or the conditions of learning influence memory performance (Koriat 1997; Benjamin 2003). Learners may also rely on their beliefs about their own skills and competence (Bandura 1997). Similarly, FOK judgements have been said to rest on deliberate inferences from one's own beliefs and knowledge (Nelson et al. 1984; Costermans et al. 1992). Discussions of subjective confidence also emphasize information-driven processes: confidence in two-alternative forced-choice (2AFC) general-knowledge question was claimed to rest on the balance of evidence in favour of the two answers (e.g. Koriat et al. 1980; Griffin and Tversky 1992; McKenzie 1997).

Unlike information-based approaches, which emphasize the content of domain-specific beliefs and knowledge retrieved from memory, experience-based approaches focus on the contribution of mnemonic cues that derive on-line from task performance. These cues are assumed to give rise automatically and unconsciously to a sheer metacognitive feeling (Koriat 2000; see Proust 2007, for a philosophical discussion). Indeed, extensive research has testified to the effects of internal cues on a variety of metacognitive judgements. Results suggest that JOLs made during study rest on the ease with which to-be remembered items are encoded or retrieved during learning (Nelson et al. 2004; Koriat and Ma'ayan 2005; Koriat et al. 2006; Karpicke 2009). FOK judgements have

been claimed to rely on the familiarity of the pointer that serves to probe memory (Reder 1988; Schwartz and Metcalfe 1992), or on the amount of partial clues that come to mind during the search for the memory target, and the ease with which they come to mind (Koriat 1993, 1995).

Confidence judgements seem also to rest on the fluency of selecting or retrieving an answer. Of particular relevance to the present work are findings indicating that participants express stronger confidence in the answers that they retrieve more quickly, whether those answers are correct or not (e.g. Kelley and Lindsay 1993; Robinson et al. 1997; Koriat et al. 2006). Largely, however, response speed is diagnostic of the correctness of the answer, so that the accuracy of confidence judgements is mediated in part by reliance on response latency (Costermans et al. 1992; Koriat and Ackerman 2010).

The processes underlying confidence judgements

Using the distinction between the three bases of metacognitive judgements, I will now outline several propositions regarding the processes underlying confidence judgements. To illustrate some of these propositions, I will use several informal observations regarding the reasons that people use to support some of the beliefs that they hold with strong conviction. For example, I would ask a student: 'What is your name?'. I would then ask: 'How confident are you that this is indeed your name?'. Generally, after an initial embarrassment, the answer is: 'Of course, one hundred percent'. When I then ask 'Why are you so confident?' the student would typically pause, and sometimes the immediate response is 'I just know'. Some students simply insist on a 'just know' response, perhaps implying a direct-access basis. Others venture to provide reasons, and these reasons seem often quite weak ('I remember that my girlfriend calls me Daniel. Actually she calls me Danny, but you know that Danny and Daniel are the same'; 'I can see my name printed on my driver's licence', etc.). Are these indeed the actual bases of one's strong conviction in one's own name? These and similar observations can help illustrate the following propositions regarding confidence judgements:

1. I propose that, in general, the *immediate* bases of feelings of confidence, as well as of other metacognitive feelings, lie primarily in mnemonic cues that derive from task performance rather than in the content of domain-specific declarative information retrieved from long-term memory. This proposal is based on observations in metacognition, which suggest that participants hardly apply their declarative knowledge and theories in making metacognitive judgements.

   For example, Koriat et al. (2004) found that JOLs made during learning were entirely indifferent to the expected retention interval, although actual recall exhibited the typical forgetting function. Thus, participants gave similar recall predictions whether they expected to be tested immediately after study, after a week, or even after a year. Koriat et al. proposed that JOLs rely primarily on encoding fluency, and that the fluency with which an item is encoded during study is not affected by when testing is expected. In addition, Kornell and Bjork (2009) found that JOLs fail to take into account the effects of number of study trials on memory (see also Kornell 2011; Kornell et al. 2011). Thus, learners do not apply spontaneously some of the most basic beliefs about learning and remembering in making recall predictions. They do so only under some specific conditions. For example, in Koriat et al.'s study, participants exhibited sensitivity to retention interval when they were asked to predict forgetting ('how many words will you forget') rather than remembering ('how many words will you recall'; see also Finn 2008).

   Furthermore, Koriat et al. (2008) had participants choose an answer to general-information questions, list reasons in support of their choice, and then indicate their confidence in the

correctness of the answer. When participants were required to list four supporting reasons, their confidence was lower than when they were required to list only one supporting reason. Thus, the effects of ease of retrieval (four reasons are more difficult to retrieve than one reason) can override the effects of the declarative content of the supporting reasons in affecting confidence judgements (Jacoby et al. 1989).

2. Information-driven processes, however, do play an important role in choice and confidence. It is proposed that when participants are presented with a 2AFC general-knowledge question, they engage typically in an analytic-like process, retrieving information from memory, and evaluating its implications before choosing the answer (see Koriat et al. 1980; Gigerenzer et al. 1991; Shafir et al. 1993). Often the pieces of information that come to mind consist of associations, hunches, and images that are not readily expressed in the form of declarative statements, but they can nevertheless tip the balance in one direction or the other. When participants have then to assess their confidence in their choice, they do not go over the entire protocol underlying their decision but rely primarily on the 'gist' of that protocol. They base their confidence on contentless mnemonic cues, such as the amount of deliberation and conflict that they had experienced in reaching the decision, and the speed with which the decision had been reached. These non-analytic cues (see Jacoby and Brooks 1984) represent the feedback from the *process* underlying the decision. Although these cues differ in quality from the considerations that were made in making the decision, they mirror significant aspects of the process that had determined the decision itself, primarily the balance of evidence in favour of the two options.

   As an analogy, we can think of a decision-making body that selects one of two alternatives based on majority rule. Once all the arguments have been heard and a vote has been cast, this vote is what finally matters. Likewise, confidence judgements would seem to rely primarily on the final vote—the overall impression formed after a deliberation regarding the relative support for each alternative. This overall impression is reflected in immediately available mnemonic cues, such as the amount of time it took to reach the decision. Perhaps, then, people are convinced about their own names not so much because of the content of individual considerations, but because of the 'unanimous vote'—the consensus among the variety of pieces of information that come to mind, and the ease and persistence with which they come to mind. Thus, it is proposed that as participants move from choosing an answer to assessing their confidence in that answer, the contribution of information-driven processes decreases and that of mnemonic cues increases.

3. The accuracy of metacognitive judgements depends largely on the extent to which the considerations and associations that come to mind lean towards the correct answer. Because these considerations and associations reflect the effects of learning and experience, they tend to support the correct answer. Proponents of the ecological probability approach (Brunswik 1956; Gigerenzer et al. 1991; Juslin and Olsson 1997; Fiedler 2007) have stressed the idea that people internalize the associations between cues and events in the world, and use the internalized knowledge when making metacognitive judgements. It is important to add that learning not only makes available declarative knowledge but also helps educate subjective experience itself. Information that is better learned, tends to be more readily retrievable, and tends to come to mind with greater consistency and persistence (Benjamin and Bjork 1996). Indeed, in a large number of studies, primitive subjective attributes, such as recognition, familiarity, fluency, and accessibility have been shown to provide valuable diagnostic information that can be used by the person as a basis for judgements (e.g. Kelley and Lindsay 1993; Koriat 1993; Goldstein and Gigerenzer 2002; Hertwig et al. 2008).

4. The processes underlying mnemonic-based metacognitive judgements occur largely outside of awareness (Proust 2008). This assumption contrasts with the spirit of information-based accounts of metacognitive judgements. For example, according to the theory of Probabilistic Mental Models (PMM; Gigerenzer et al. 1991) people choose between two answers by retrieving a cue that discriminates between the two answers. Associated with each cue is also a cue validity that describes how well that cue predicts the criterion. When the cue determines the choice, its cue validity is then reported as the confidence in the choice.

   Experience-based approaches, in contrast, assume that the process is much less analytic, and that people have little awareness of the mnemonic cues underlying their metacognitive judgements, let alone their cue validity (see Koriat et al. 2009). For example, in the mere exposure effect, repeated exposure to stimuli, even under subliminal presentation, has been found to lead to increased liking of these stimuli, although during debriefing, most participants predict that repeated exposures would lead to boredom and decreased liking (Murphy et al. 1995).

   Because metacognitive feelings rest on unconscious inferences (Jacoby et al. 1989), the phenomenology of these feelings is most consistent with the direct-access view. Metacognitive feelings often have the quality of direct perceptions (Kahneman 2003; Kahneman and Frederick 2005). A person in a TOT state, for example, can 'sense' the elusive name or word and can monitor its emergence into consciousness (Brown and McNeill 1966; see Schwartz and Metcalfe 2011). Subjective convictions in beliefs also have the quality of direct access. Therefore, the validity of metacognitive feelings is sometimes taken for granted by the person (Epstein and Pacini 1999), although such feelings may prove illusory in retrospect (Koriat 1994; Schwartz 1998). It would seem that direct-access accounts of metacognitive feelings derive their power primarily from the phenomenology of these feelings and from their general accuracy in predicting memory performance.

5. Because the heuristics that underlie immediate metacognitive feelings operate below full consciousness (Koriat 2000), when participants are asked to explain the reasons for their metacognitive feelings, they usually refer to declarative knowledge and theories rather than to the underlying mnemonic cues that derive from task performance. Never have I heard a participant justify his or her high JOL, FOK, or confidence by referring to such factors as processing fluency or ease of retrieval. Of course, the reasons mentioned by participants to justify their metacognitive feelings often capture some of the distal ecological influences that have shaped the mnemonic cues underlying these feelings. Going back to the conviction in one's own name, it is my argument, as I noted, that the student is convinced of his name because of the simple fact that every way he thinks about his name, the same name comes consistently, insistently and quickly to mind. However, the justifications mentioned by him may reflect the historical factors that are responsible for the mnemonic qualities associated with retrieving one's name. These qualities derive from one's own experience, such as the frequent usage of the name by one's acquaintances, the many instances in which one has to say or write one's name, and so forth.

In sum, the three approaches to the basis of metacognitive judgements may reflect different aspects of the processes underlying these judgements. Although these approaches imply qualitatively different processes, there is a great deal of overlap between their predictions. The mnemonic cues assumed to underlie subjective confidence mirror the information-based cues that drive the choice of an answer. In turn, the phenomenological quality of subjective convictions is seen to derive from the unconscious nature of mnemonic-based feelings, resulting in retrospective justifications of these feelings that stress declarative semantic and episodic considerations.

These propositions depart from what might be concluded from the preponderance of experimental findings demonstrating misleading effects of mnemonic cues (e.g. Chandler 1994; Koriat 1995; Benjamin et al. 1998; Brewer and Sampaio 2006). These demonstrations, which were intended to bring to the fore the contribution of mnemonic cues, have resulted in overemphasis on situations in which mnemonic cues drive judgements away from what would be implied by analytic considerations, resulting in faulty judgements. Under natural conditions, however, mnemonic cues tend to be valid, and their validity derives from the effects of learning and past experience.

Subjective confidence: the motivation for the present proposal

I will now describe some of the work that has led to the self-consistency model of subjective confidence. Some of the tasks that I used to study subjective confidence were intended to tap 'intuitive' judgements. These tasks were inspired by the idea of some philosophers that universally shared notions that are grasped by intuition, have the quality of self-evidence: they strike you as being right. One such task that I used was based on the well-known demonstration by Köhler (1947): 'There is a language that has names for different shapes. Guess which of these shapes is called Maluma and which is Takete'. Two observations were noteworthy: first, practically all participants matched the rounded shape with Maluma, but when I asked them to state the reasons for their choice, their reasons differed greatly across participants. Second, all participants expressed strong convictions in the correctness of their response to the extent that when I told some participants that they were wrong, the typical reaction was 'that's impossible!'. This is similar to the phenomenal feeling that philosophers associate with a priori or analytic truths: Such truths feel *necessarily* correct.

In the Maluma–Takete example, there is no right or wrong answer. However, similar observations were made with similar tasks in which there was a correct answer. One such task required the matching of antonymic words from non-cognate languages (e.g. *tuun–luk*) with their English equivalents (*deep–shallow*). This task had been used by researchers to examine the idea that a universal sound-meaning symbolism that has been incorporated in the formation of all languages, and people have an intuitive feel for it. I was interested to know whether correct matches tend to be endorsed with stronger confidence than wrong matches. In one study, (Koriat 1975) participants' matches were found to be significantly better than chance, averaging 58.1%. In addition, the percentage of correct matches increased steeply with confidence judgements, suggesting that participants were successful in monitoring the correctness of their matches. The latter result presented a puzzle. Neither the information-based approach nor the experience-based approach offers a hint regarding the cues that participants might use to monitor their knowledge. The finding is reminiscent of the direct-access view that rationalists posit with regard to a priori propositions that are accessed through intuition.

An important feature of the word-matching task is that no simple algorithm exists for determining whether the answer is correct or wrong. However, such is also the case in many memory tasks in which participants are successful in monitoring the correctness of their answers. Thus, perhaps, there is some general principle that underlies the accuracy of monitoring in a variety of tasks, including memory tasks and the word-matching task.

In attempting to uncover such a principle, I reasoned that perhaps the observation that participants' matches were largely accurate ('knowledge') creates a confounding for the assessment of the confidence–accuracy correlation ('metaknowledge'). Because the correct match is the one that is consensually endorsed, perhaps confidence judgements are correlated with the consensuality of the match rather than with its correctness. To examine this possibility I tried to dissociate

between correctness and consensuality by including many items for which participants are likely to agree on the *wrong* match (Koriat 1976). The results clearly indicated that confidence ratings correlated with the consensuality of the match rather than with its correctness: For consensually-correct (CC) items, for which most participants chose the correct answer, correct answers were endorsed with stronger confidence, whereas for consensually-wrong (CW) items it was the *wrong* answers that were associated with stronger confidence. The *consensuality principle*—that confidence is correlated with the consensuality of the answer rather than with its correctness—has been replicated since for several other tasks as will be detailed later.

The conclusion from these results is that when a representative sample of items is used, participants are successful in monitoring the correctness of their responses, but they do that *indirectly* by relying on some cues that are correlated with accuracy. These cues would seem to underlie the consensuality of the response—the extent to which it tends to be endorsed by the majority of people. Thus, what Tulving and Madigan (1970) regarded as a truly unique characteristic of human memory turns out to be an artefactual consequence of the fact that in virtually all studies that examined the confidence-accuracy correlation in memory tasks, the consensually endorsed answer is the correct answer. That is, the percentage of correct answers in 2AFC questions is practically always above 50%. Thus, metaknowledge accuracy and knowledge accuracy are intimately linked: metaknowledge is accurate as long as knowledge itself is accurate.

The consensuality principle was also confirmed for response latency. Previous studies had established that response speed is diagnostic of accuracy, being faster for correct than for wrong answers (Kelley and Lindsay 1993; Robinson et al. 1997; Koriat et al. 2006). However, we showed that this is true only for CC items, whereas for CW items the opposite relationship is found (Koriat 2008, 2012).

The consensuality principle is a descriptive principle that does not offer a process account of the basis of confidence judgements and their accuracy. However, it may provide a lead to the question of how we know that we know. It suggests that what makes a person confident in a particular answer is what makes most people favour that answer in the first place. This idea motivated the development of the self-consistency model of subjective confidence (Koriat 2011, 2012; Koriat and Adiv 2011). Before describing the model, I would like to spell out its underlying metatheoretical assumptions.

A preamble to the model: philosophical perspectives

In this section, I would like to place the present proposal with respect to two major issues in epistemology. The first concerns the distinction between the rationalist and empiricist positions regarding the origin of knowledge, and the second concerns the distinction between correspondence and coherence theories of truth.

The origin of knowledge: rationalism versus empiricism

A central issue in the philosophy of knowledge is associated with the traditional distinction between rationalism and empiricism (see Edwards 1996; Markie 2008). The rationalist approach focuses on intuitive knowledge—a priori propositions whose truth is self-evident. Rationalists, such as Descartes, Spinoza, and Leibniz, maintained that there are significant aspects of our concepts and knowledge that are gained independent of sense experience. These are knowable either by direct intuition, or by deduction from intuited propositions. The examples mentioned include mathematical propositions, logical arguments, ethical or moral propositions, and even metaphysical beliefs (e.g. that God exists). Some rationalists posit that such truths are innate. Carruthers (1992), for example, argued that knowledge of some of the principles of folk-psychology (e.g. that

pain tends to be caused by injury) is innate. Innateness generally implies universality. As I noted earlier, my early research on the ability to guess the meaning of foreign words (Koriat 1975) was inspired by the notion that intuited, universal truths are phenomenologically experienced as self-evident.

In contrast, empiricists such as Locke and Berkeley argued that the origin of knowledge resides in the external world. According to them, sense experience is the ultimate source of knowledge and therefore the focus should be on a posteriori propositions whose justification relies on empirical observations.

Most philosophers, however, admit both sources of knowledge. Albert Einstein discussed the 'eternal antithesis between the two inseparable components of our knowledge, the empirical and the rational':

> We reverence ancient Greece as the cradle of western science. Here for the first time the world witnessed the miracle of a logical system which proceeded from step to step with such precision that every single one of its propositions was absolutely indubitable. I refer to Euclid's geometry. This admirable triumph of reasoning gave the human intellect the necessary confidence in itself for its subsequent achievements ….
>
> But before mankind could be ripe for a science which takes in the whole of reality, a second fundamental truth was needed, which only became common property among philosophers with the advent of Kepler and Galileo. Pure logical thinking cannot yield us any knowledge of the empirical world; all knowledge of reality starts from experience and ends in it. Propositions arrived at by purely logical means are completely empty as regards reality. Because Galileo saw this, and particularly because he drummed it into the scientific world, he is the father of modern physics – indeed, of modern science altogether. (Einstein 1934/1954, p. 271.)

At the risk of oversimplification, I would like to stress two aspects that distinguish the two types of knowledge. First, for rationalists, truth lies within: it can be grasped through 'pure reason'. In a sense, its acquisition is based on direct access (or else on a deduction from directly accessed truths). For empiricists, in contrast, knowledge originates from the outside world and hence ultimately relies on empirical observations. Second, there is a consensus that intuition and deduction provide beliefs whose truth is self-evident, and is beyond any doubt ('absolutely indubitable' in Einstein's words). These beliefs are endowed with a sense of necessity. Empiricists, in contrast, admit a degree of uncertainty, arguing, for example, that we can never be sure that our sensory impressions are true.

These comments suggest that perhaps different processes underlie confidence judgements when knowledge originates from within than when it originates from without. Taken together, however, the results of Koriat (1975, 1976, 2008, 2011) suggest otherwise. Furthermore, with regard to intuitive knowledge, the extensive work on intuitive feelings by experimental psychologists (see Lieberman 2000; Hogarth 2001; Myers 2002; Kahneman 2003; Plessner et al. 2007) raises concern about the assumptions among some philosophers that there exists an intimate link between intuition and a priori, innate knowledge, and that intuition provides knowledge whose truth is absolutely certain. Not only has there been evidence that intuitive, gut feelings can have their origin in experience (Westcott 1968; Reber 1989), but also that intuitive feelings that are held with strong subjective certainty are sometimes wrong (Denes-Raj and Epstein 1994; Koriat 1994, 1998; see Nagel 2007). This evidence blurs the distinction between knowledge originating from within and knowledge originating from without, and invites a common framework in which subjective confidence in both types of knowledge can be analysed.

To build such a framework, consider the psychological situation of a participant who is required to assess the confidence in the answer to such questions as 'Which city has more inhabitants, Hanover or Bielefeld?' (Gigerenzer et al. 1991), or 'What is the capital of Australia, Canberra or

Sydney?' (Fischhoff et al. 1977). The pertinent clues for the answer must be retrieved from one's own memory rather than (directly) from the outside world. In this respect, the situation is not different from that underlying the verification of analytic truths. Such is also the case when the propositions concern semantic knowledge, episodic memory, or social and metaphysical beliefs (e.g. 'There is a supreme being controlling the universe'; see later). In attempting to validate one's memories or beliefs (see Ross 1997) or to judge the source of one's memories (see Mitchell and Johnson 2000; Lindsay 2008) one must make do with a variety of pieces of information accessed from within. Indeed, a recent functional magnetic resonance imaging study suggests that the neural activity related to metacognitive judgements is characterized by a shift away from externally directed cognition toward internally directed cognition (Chua et al. 2009). So what is the basis of one's degree of certainty in an answer that is retrieved from memory?

The first postulate underlying SCM is that although the validation of one's own knowledge is based on retrieving information from memory, the underlying process is analogous to that in which information is sampled from the outside world with the goal (1) to test a hypothesis about a population, and (2) to assess the likelihood that the conclusion reached is correct. I argue that such is the case whether participants need to validate propositions whose truth is a priori or propositions whose truth is a posteriori. Thus, the prototype for the underlying process is provided by the statistical procedures that are used by researchers in attempting to draw conclusions about the external world: a proximal sample of observations is used to make inferences about some 'true' parameter of a distal population. The critical difference, of course, is that information is sampled from within rather than from without. The model to be sketched as follows incorporates this assumption.

Correspondence versus coherence theories of truth

I turn now to the second issue, which helps introduce the second postulate. This issue concerns the distinction between two major philosophical theories of truth, correspondence theories and coherence theories (Kirkham 1992). Correspondence theories posit that the truth or falsity of a statement is determined only by how it relates to the world, and whether it accurately describes objects or facts. Coherence theories, in contrast, assume that the truth or falsity of a statement is determined by its relations to other statements rather than its relation to the world (Rescher 1973; Walker 1989). In this view, a person's belief is true if it is coherent with his or her body of beliefs, that is, if it is a constituent of a systematically coherent whole.

The correspondence view of truth reflects the intentions of confidence judgements. Confidence in a proposition reflects the likelihood that that proposition agrees with reality (e.g. that one's name is indeed Daniel, or that Canberra is indeed the capital of Australia). The problem, however, is how can one assess such agreement with reality if one does not have access to reality independent of what one knows or believes about it? Kant stated this problem as follows:

> Truth is said to consist in the agreement of knowledge with the object. According to this mere verbal definition, then, my knowledge, in order to be true, must agree with the object. Now, I can only compare the object with my knowledge by this means, namely, by taking knowledge of it. My knowledge, then, is to be verified by itself, which is far from being sufficient for truth. For as the object is external to me, and the knowledge is in me, I can only judge whether my knowledge of the object agrees with my knowledge of the object. Such a circle in explanation was called by the ancients Diallelos. (Kant 1885, p. 40.)

The resolution of this issue calls for a second postulate: although confidence judgements pertain to correspondence, the mnemonic cue for metacognitive assessments of correspondence is degree of coherence. Confidence in an answer or belief depends on the extent to which that

answer or belief is supported consistently by the various pieces of information that come to mind. Indeed, several discussions have stressed the use of internal consistency as a cue for the validity of one's own beliefs. Ross (1997), for example, noted that people rely on internal coherence in judging the validity of their recollections, and use incoherence and internal contradictions as good reasons to doubt the reality of recollections.

In epistemological discussions, the notion of coherence has been discussed extensively in connection with the justification of beliefs. Unlike *foundationalist* theories, which assume that beliefs are justified on the basis of other beliefs, *Coherentism* theories, claim that a belief is justified by the way it fits together with the rest of the belief system of which it is a part (BonJour 1985). Foundationalists escape the regress problem of an infinite chain of justification (see Moser 1988) by postulating the existence of justified basic beliefs that do not owe their justification to other beliefs (Van Cleve 2005). Coherentists, in contrast, avoid the regress problem without postulating the existence of non-inferential basic beliefs.

The notion of coherence that I assume to underlie confidence judgements is quite loose. First, I assume that what matters is only the internal consistency within the set of thoughts that are activated during the attempt to answer a question or validate a belief. In this respect, coherence or consistency can be said to be output-bound (Koriat and Goldsmith 1996), relative to the set of clues that are activated. Second, what is activated during the choice of an answer is generally an assortment of images, memories, beliefs, associations, and thoughts that cannot always be expressed in a propositional form. Therefore, coherence reflects the extent to which these clues produce a sense of convergence versus a sense of tension or conflict. Indeed, studies of the illusory-truth effect indicate that mere familiarity and fluency can enhance truth judgements. For example, the repetition of a statement increases its perceived truth even when the statements are actually false (Hasher et al. 1977; Bacon 1979; Arkes et al. 1989). Truth judgements are also enhanced by perceptual fluency (e.g. visual contrast; Reber and Schwarz 1999; Hansen et al. 2008; Unkelbach and Stahl 2009) and by manipulations that increase contextual fluency (placing the statement in contexts that provide a continuity of meaning; Parks and Toth 2006).

In sum, because people have no access to the object of their beliefs over and above what they know about it, they rely on a fast assessment of overall coherence (see Bolte and Goschke 2005) as a basis for their judgements about correspondence. In terms of Polanyi's (1958) terminology, the 'object' of metacognitive judgements is correspondence, but the 'tool' is coherence. This state of affairs raises a dilemma for the evaluation of the accuracy of one's confidence judgements: should these judgements be evaluated against correspondence, because this is what participants feel (and state), or should they be evaluated against coherence? As we note later, the discrepancy between the two criteria may explain the overconfidence bias observed in calibration research.

The self-consistency model of subjective confidence rests on the two postulates mentioned earlier. First, it assumes that although information is retrieved from memory, the process is similar to the statistical procedure involved in assessing confidence in a sample-based inference about the outside world. Second, coherence or reliability is used as a cue for validity.

The self-consistency model of subjective confidence

I will now present briefly the SCM of subjective confidence. Underlying SCM is a metaphor of the person as an intuitive statistician (Peterson and Beach 1967; Gigerenzer and Murray 1987; see McKenzie 2005). People's confidence judgements are modelled by the classical procedures of calculating statistical level of confidence when conclusions about a population are to be made based on a sample of observations. When faced with a 2AFC general-information question, or a question about some social or metaphysical belief, it is by replicating the choice process several

times that a person can appreciate the degree of doubt or certainty involved. The assessment of degree of certainty is obtained by sampling different 'representations' or considerations from memory and assessing the extent to which they agree in favouring a particular decision. Subjective confidence essentially represents an assessment of *reproducibility*—the likelihood that a new sample of representations drawn from the same population will yield the same choice. Thus, reliability is used as a cue for validity.

SCM does not pretend to describe the complex processes involved in making a choice, but only to capture the *feedback* from that process. It is assumed that this feedback, which affects confidence, is a crude sense of consistency that can be modelled by a simple count of the proportion of representations favouring each of the two alternatives (see Alba and Marmorstein 1987). A detailed description of the model can be found elsewhere (Koriat 2011, 2012). Here only a brief description will be presented of a specific implementation of the model.

An important assumption of SCM is that in responding to 2AFC items, whether they involve general-information questions or beliefs and attitudes, participants with the same experience draw representations largely from the same, commonly shared population of representations associated with each item. If each representation favours one of the two answers, each item can be characterized by a probability distribution, with p_{maj} denoting the probability that a representation favouring the majority alternative will be sampled.

Given a particular value of n, the number of representations sampled, the parameter p_{maj} for a given item may be estimated from the probability with which the majority alternative is chosen. This probability can be indexed operationally by the proportion of participants who choose the preferred alternative ('item consensus'), or by the proportion of times that the same participant chooses the preferred alternative across repeated presentations ('item consistency'). For example, for an item with a 40–60% between-participant split of choices, item consensus will be 60%.

One version of the model assumes that participants sample a maximum of seven representations, each of which yields a binary subdecision, and that the overt choice is dictated by the majority vote. However, if three representations in a row yield the same subdecision, the search is stopped and the Run-3 subdecision is reported. An index of self-consistency was used, which is related to the standard deviation of the subdecisions: $1-\sqrt{\hat{p}\hat{q}}$. It is calculated over the actual number of representations sampled.

A simulation experiment that incorporates these simple assumptions yielded the results depicted in Fig. 13.1. These results indicate the functions relating the index of self-consistency to the probability of choosing the majority answer, pc_{maj} for majority and minority choices. Three features should be noted. First, mean self-consistency (and hence, confidence) for each item should increase with pc_{maj}. Second, self-consistency is systematically higher for majority than for minority choices. Finally, whereas for majority choices, self-consistency increases steeply with pc_{maj}, for minority choices, it decreases but much more shallowly.

Why is self-consistency lower for minority than for majority choices? The reason is that when a sample of representations happens to favour a minority choice, the proportion of subdecisions favouring that choice will be smaller on average than when the sample favours the majority choice. For example, for $p_{maj} = 0.70$, and $n = 7$, the likelihood that six or seven representations will favour the majority answer is 0.329, whereas only in 0.004 of the samples will six or seven representations favour the minority answer.

The simulation experiment mentioned earlier indicated that the results for n_{act}, the number of representations actually drawn, mimic very closely those obtained for self-consistency. Assuming that response latency increases with n_{act}, it should be longer for minority than for majority choices.

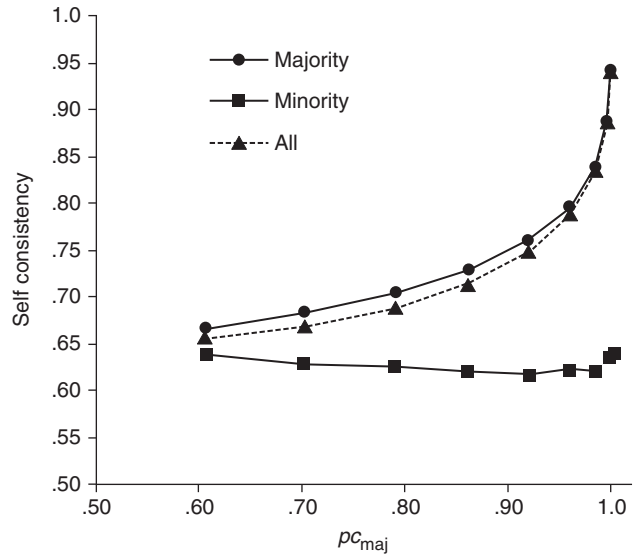

Fig. 13.1 Self-consistency scores as a function of the probability of choosing the majority option (PC_{maj}) based on the results of the simulation experiment. The results are plotted separately for majority and minority choices. Reprinted from figure 2, panel A, in 'The Construction of Attitudinal Judgments: Evidence from Attitude Certainty and Response Latency' by A. Koriat and S. Adiv, *Social Cognition*, 29, 2011, 587, Copyright 2011 by Oxford University Press.

Note that the results in Fig. 13.1 were obtained under the assumption that participants choose the alternative that is favoured by the *majority* of representations in *their* accessed sample of representations. This pattern should be obtained both in a within-individual analysis and in a between-individual analysis.

Empirical evidence

I will present a brief summary of the results of several experiments in which these predictions were tested.

The relationship between confidence and cross-person consensus

As noted, pc_{maj} can be indexed by the proportion of participants who choose the majority answer. To test the predictions of the model, the answer that was chosen by the majority of participants for each item was designated ad hoc as the consensual (majority) answer, and the other as the non-consensual (minority) answer. Mean confidence was then plotted as a function of item consensus—the proportion of participants who chose the majority answer. This was done separately for majority and minority answers. The results yielded a pattern that is qualitatively similar to that depicted in Fig. 13.1. This was true across several tasks: general knowledge, word matching, comparison of the length of two lines, comparison of the area of two figures, social beliefs, and social attitudes (Koriat 2011, 2012; Koriat and Adiv 2011). In all of the tasks, participants made a two-alternative choice and expressed their confidence in the choice. The generality of the findings across domains supports the assumption of SCM that confidence is based on mnemonic cues that are indifferent to the specific content of the representations that are sampled.

The systematic difference between majority and minority choices can explain the consensuality principle: Participants are more confident when their choice is consistent with that of most participants. This should be the case even if all participants are assumed to draw their representations from a commonly shared population (see Fig. 13.1).

The relationship between confidence and within-person consistency

SCM was tested also in a within-individual design. Participants were presented repeatedly with the same set of 2AFC items. The answers to each item were then classified as frequent or rare depending on their relative frequency across repetitions, and confidence was plotted for the frequent and rare answers as a function of item consistency—the relative frequency of the frequent (majority) choice across repetitions.

The predictions of SCM for a within-person design were tested for general knowledge, word matching, perceptual judgements, and social beliefs and attitudes. In all of these domains, the results were in line with predictions. In particular, participants were more confident when they made their more frequent choice than when they made their less frequent choice.

Another result that was observed is that confidence in the first presentation predicted the likelihood of making the same choice in subsequent presentations of the item. This is consistent with the assumption that subjective confidence in a choice monitors reproducibility—the likelihood of making the same choice in a subsequent presentation of the item.

Response latency

All of the results summarized so far were replicated when response speed rather than confidence was used as the dependent variable. Thus, response latency was overall shorter for consensual choices than for non-consensual choices and for frequent choices than for rare choices. Overall, the results suggest that response speed is a frugal cue for self-consistency and can be used as a basis for confidence. The results also indicated that the speed of a choice predicts the reproducibility of the choice.

The correlation between confidence and accuracy

The results for the confidence–accuracy correlation also yielded clear support for the consensuality principle. This was true for general-information questions (Koriat 2008), FOK judgements (Koriat 1995), and perceptual judgements (Koriat 2011). It was also observed for sentence memory (Brewer and Sampaio 2006). Both confidence and response speed were correlated with the consensuality of the choice rather than with its correctness: The confidence–accuracy correlation was positive when the consensual choice was the correct choice but negative when it was the wrong choice.

These results disclose the link between knowledge and metaknowledge (Koriat 1993): people know that they know because (or when) they know. Indeed, for the CW items people are 'doubly cursed' (Dunning et al. 2003): they do not know, and do not know that they do not know.

The results for perceptual judgements (Koriat 2011) also supported the consistency principle, which is analogous to the consensuality principle: The confidence-accuracy correlation was positive for items in which the participant's frequent choice was the correct choice but negative for items in which the frequent choice was the wrong choice. These results were also mimicked by the results for response speed.

The calibration of confidence judgements

SCM also provides an account of the overconfidence bias that has been observed in calibration studies (Lichtenstein et al. 1982; Griffin and Brenner 2004). According to SCM, the overconfidence

bias derives, in part, from participants' reliance on reliability as a cue for validity. Reliability (or consistency) is practically always higher than validity. Confidence judgements are assumed to monitor self-consistency but their accuracy is evaluated in calibration studies against correctness. Indeed, the overconfidence bias was reduced or eliminated when confidence was evaluated against indexes of self-consistency rather than against correctness (Koriat 2011).

Interparticipant consensus in choice and confidence

Consistent with SCM, all of the tasks mentioned exhibited a marked degree of cross-person consensus, suggesting that participants share the same core of item-specific representations from which they draw their sample of representations on each occasion. This was true even for social beliefs and social attitudes. Furthermore, cross-person consensus and within-person consistency were correlated so that the choices that evidenced higher within-person consistency were the more likely to be made by other participants.

Discussion

The question of how we can be certain about our beliefs has intrigued philosophers for centuries, and has been addressed in a broad range of domains. Subjective confidence has also attracted much interest in view of the many observations testifying for serious deficiencies in the ability to monitor one's own knowledge and performance (Burton 2008).

In this chapter, I described briefly a model of subjective confidence, focusing on the metatheoretical assumptions underlying the model. In what follows, I discuss these assumptions. SCM assumes that confidence judgements are inferential in nature, relying primarily on cues that derive from task performance. This view departs from the direct access view, which assumes that metacognitive judgements are based on privileged access to memory traces. It also departs from the view that these judgements are mediated by an analytic process in which declarative propositions retrieved from long-term memory are consulted to reach an educated metacognitive assessment. Rather, confidence in a decision is parasitic on the process of making a decision, and is based on mnemonic cues that derive online from that process (Koriat et al. 2008).

As noted in the introduction, one of the central issues in philosophy concerns the origin of knowledge. For rationalist philosophers, the origin of knowledge lies within the person whereas for empiricist philosophers it lies without. However, it was argued that in a typical situation in which participants are required to validate a proposition, they must draw on information that resides within, whether that proposition concerns semantic and episodic memory or so-called a priori truth. Therefore, it was proposed that the self-consistency model might apply not only to memory questions that depend on real-world knowledge but also to statements concerning personal and metaphysical beliefs.

Although the clues for confidence must come from within, it was argued that the process has much in common with that in which information is retrieved from without. Specifically, in testing a hypothesis about a population based on a sample of observations, researchers generally put greater trust in the hypothesis as a function of the level of significance with which the null hypothesis is rejected. That is, they behave as if the correctness of the hypothesis, as well as the likely reproducibility of the observed result, is a monotonically increasing function of level of confidence (see Schervish 1996; Dienes 2011). Statistical level of confidence increases with decreased variance—the extent to which the sampled observations consistently support the hypothesis. Let us examine this idea closely as it bears on the distinction between coherence and correspondence.

Assume that we wish to test the hypothesis that among married couples, husbands are happier than their wives. We draw randomly one couple from a population and find that indeed the

husband is happier than his wife. Apart from the fact that most people will not put too much faith in a conclusion that is based on a sample of $n = 1$, the problem is that such a sample does not allow assessment of the credibility of the conclusion.

The situation changes radically when a larger sample is drawn, e.g. $n = 100$. In this case, statistical level of confidence is based on the *internal consistency* within the sample. If we find that in 80 of the 100 couples husbands are happier, our faith in the hypothesis that in the 'real world' (i.e. in the population as a whole) husbands tend to be happier stems from the consistency *within the sample*. Thus, in a sense, coherence is used as a cue for correspondence.

SCM assumes that in a similar manner, it is by replicating the choice process several times that people appreciate the amount of doubt involved. Subjective confidence in the *validity* of a proposition is then based on the *reliability* with which the proposition is supported across the sample of representations.

This view differs from that of the PMM theory, which assumes that confidence is based on the stored validity of a single cue that discriminate between the two alternative answers. Clearly, PMM is an inferential, cue-based model as far as the choice of an answer is concerned. However, when it comes to confidence, the model is more like a trace-access model because confidence is read out directly from the stored validity of the cue.

Unlike PMM theory, SCM assumes that confidence depends on the internal consistency within a *collection* of representations. This assumption avoids the regress problem without postulating a direct-access basis for confidence. The logic underlying SCM is the same as that underlying the (mis)interpretation of statistical level of confidence as capturing the degree of trust in a hypothesis. The finding that confidence in the first presentation of an item predicts the likelihood of making the same choice in subsequent presentations of the item also parallels the (mis)interpretation of statistical level of confidence as capturing the likely reproducibility of the observed effect.

In line with SCM, confidence judgements were found to track both the stable and variable contributions to choice. The stable contributions stem from the constraints imposed by the population of representations available in memory. In general, the polarization of the population of representations associated with an item constraints the extent of fluctuation in judgements that may be expected across occasions and across people. The variable contributions are disclosed by the systematic differences between majority and minority choices, which are assumed to convey information about the specific sample of representations underlying a particular choice (Koriat and Adiv 2011; see also Wright 2010).

The finding that the same pattern of results was obtained across different domains reinforces the assumption that confidence is based on structural, contentless cues. This finding may also be taken to imply that from a psychological point of view the processes underlying confidence in a priori truths are not qualitatively different from those underlying confidence in a posteriori truths. Admittedly, in the case of a priori truths (e.g. that the internal angles of a triangle add up to 180 degrees or that two plus two equals four), there is generally little variance between the outcomes of different representations of the question. However, perhaps that is precisely the cue for the strong conviction associated with such statements: What characterizes a priori beliefs is that however one thinks of them one arrives at the same conclusion. Nevertheless, the question should be entertained whether there are particular beliefs for which we should postulate some sort of direct access.

This question actually applies to episodic knowledge as well, when such knowledge is held with strong confidence (e.g. one's name). Metcalfe (2000), for example, postulated a 'special noetic state' in which metacognitive judgements are based on direct access rather than on inference from cues. Unkelbach and Stahl (2009) also proposed that when judging the truth of a statement,

'participants may simply know the factual truth or falsity of a statement and judge it accordingly' (p. 24). Gigerenzer et al.'s PMM model (1991) also incorporates a strategy (*local mental model*) in which a choice of an answer is based on a direct solution by memory. Only when this strategy fails, do participants construct a PMM that uses probabilistic information from a natural environment. Thus, an important question that we leave open is whether there are beliefs for which subjective confidence depends on a process that is qualitatively different from that postulated by SCM.

Acknowledgement

I am grateful to Shiri Adiv for her help and advice in the preparation of the manuscript and to Rinat Gil and Dana Klein for their help in the statistical analyses.

References

Alba, J. W. and Marmorstein, H. (1987). The effects of frequency knowledge on consumer decision making. *Journal of Consumer Research*, 14, 14–25.

Arkes, H. R., Hackett, C., and Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, 2, 81–94.

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning & Memory*, 5, 241–52.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, 31, 297–305.

Benjamin, A. S. and Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.) *Implicit memory and metacognition*, pp. 309–38. Mahwah, NJ: Erlbaum.

Benjamin, A. S., Bjork, R. A., and 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.

Bolte, A. and Goschke, T. (2005). On the speed of intuition: intuitive judgments of semantic coherence under different response deadlines. *Memory & Cognition*, 33, 1248–55.

BonJour, L. (1985). *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.

Brewer, W. F. and Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory*, 14, 540–52.

Brown, R. and McNeill, D. (1966). The 'tip of the tongue' phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5, 325–37.

Brunswik, E. (1956). *Perception and representative design in psychological experiments*. Berkeley, CA: University of California Press.

Burke, D. M., MacKay, D. G., Worthley, J. S., and Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, 30, 542–79.

Burnet, J. (1930). *Early Greek philosophy*. London: Adam and Charles Black.

Burton, R. A. (2008). *On being certain: Believing you are right even when you're not*. New York: St. Martin's Press.

Carruthers, P. (1992). *Human knowledge and human nature*. Oxford: Oxford University Press.

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, 22, 273–80.

Chua, E. F., Schacter, D. L., and Sperling, R. A. (2009). Neural correlates of metamemory: A comparison of feeling-of-knowing and retrospective confidence judgments. *Journal of Cognitive Neuroscience*, 21, 1751–65.

Cohen, R. L., Sandler, S. P., and Keglevich, L. (1991). The failure of memory monitoring in a free recall task. *Canadian Journal of Psychology*, 45, 523–38.

Costermans, J., Lories, G., and Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 142–50.

Denes-Raj, V. and Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, 66, 819–29.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–90.

Dunlosky, J. and Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage.

Dunning, D. (2007). Prediction: The inside view. In A. W. Kruglanski and E. T. Higgins (Eds.) *Social Psychology: Handbook of basic principles*, pp. 69–90. New York: The Guilford Press.

Dunning, D., Johnson, K., Ehrlinger, J., and Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12, 83–7.

Dunning, D., Heath, C., and Suls, J. M. (2004). Flawed self-assessment. *Psychological Science*, 5, 69–106.

Edwards, P. (Ed.) (1996). *The encyclopedia of philosophy*, Vols. *1 and 2—complete and unabridged*. New York: Macmillan Reference.

Einstein, A. (1934/1954). *Ideas and opinions* (S. Bargmann trans.). New York: Crown Publishers.

Epstein, S. and Pacini, R. (1999). Some basic issues regarding dual-process theories from the perspective of cognitive-experiential self-theory. In S. Chaiken and Y. Trope (Eds.) *Dual process theories in social psychology*, pp. 462–82. New York: Guilford Press.

Fiedler, K. (2007). Information ecology and the explanation of social cognition and behavior. In A. W. Kruglanski and E. T. Higgins (Eds.) *Social psychology: Handbook of basic principles*, pp. 176–200. New York: The Guilford Press.

Finn., B. (2008). Framing effects on metacognitive monitoring and control. *Memory Cognition*, 36, 813–21.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 552–64.

Gigerenzer, G., and Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506–28.

Goldsmith, M. and Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. Benjamin and B. Ross (Eds.) *Psychology of learning and motivation, Vol. 48: Memory use as skilled cognition*, pp. 1–60. San Diego, CA: Elsevier.

Goldstein, D. G. and Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109, 75–90.

Griffin, D. and Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler and N. Harvey (Eds.) *Blackwell handbook of judgment and decision making*, pp. 177–99. Malden, MA: Blackwell Publishing.

Griffin, D. and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411–35.

Hansen, J., Dechêne, A., and Wänke, M. (2008). Discrepant fluency increases subjective truth. *Journal of Experimental Social Psychology*, 44, 687–91.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56, 208–16.

Hasher, L., Goldstein, D., and Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16, 107–12.

Hertwig, R., Herzog, S. M., Schooler, L. J., and Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1191–206.

Hogarth, R. (2001). *Educating intuition*. Chicago, IL: University of Chicago Press.

Jacoby, L. L. and Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. H. Bower (Ed.) *The psychology of learning and motivation: Advances in research and theory*, pp. 1–47. New York: Academic Press.

Jacoby, L. L., Kelley, C. M., and Dywan, J. (1989). Memory attributions. In H. L. Roediger and F. I. M. Craik (Eds.) *Varieties of memory and consciousness: Essays in honour of Endel Tulving*, pp. 391–422. Hillsdale, NJ: Erlbaum.

Juslin, P. and Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, 104, 344–66.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58, 697–720.

Kahneman, D. and Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak and R. G. Morrison (Eds.) *The Cambridge handbook of thinking and reasoning*, pp. 267–93. Cambridge: Cambridge University Press.

Kant, I. (1885). *Kant's introduction to logic, and his Essay on the mistaken subtility of the Four Figures* (T. K. Abbott, trans.) London: Longmans, Green and Co.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, 138, 469–86.

Kelley, C. M. and Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, 32, 1–24.

Kirkham, R. L. (1992). *Theories of truth: A critical introduction*. Cambridge, MA: MIT Press.

Köhler, W. (1947). *Gestalt psychology*. New York: The New American Library of World Literature, Inc.

Koriat, A. (1975). Phonetic symbolism and the feeling of knowing. *Memory & Cognition*, 3, 545–8.

Koriat, A. (1976). Another look at the relationship between phonetic symbolism and the feeling of knowing. *Memory & Cognition*, 4, 244–8.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609–39.

Koriat, A. (1994). Memory's knowledge of its own knowledge: The accessibility account of the feeling of knowing. In J. Metcalfe and A. P. Shimamura (Eds.) *Metacognition: Knowing about knowing*, pp. 115–35. Cambridge, MA: MIT Press.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology*: General, 124, 311–33.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126, 349–70.

Koriat, A. (1998). Illusions of knowing: The link between knowledge and metaknowledge. In V. Y. Yzerbyt, G. Lories, and B. Dardenne (Eds.) *Metacognition: Cognitive and social dimensions*, pp. 16–34. London: Sage.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149–71.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, and E. Thompson (Eds.) *The Cambridge handbook of consciousness*, pp. 289–325. New York: Cambridge University Press.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 34, 945–59.

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, 140, 117–39.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, 119(1), 80–113.

Koriat, A. and Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science,* 13, 441–53.

Koriat, A. and Adiv, S. (2011). The construction of attitudinal judgments: Evidence from attitude certainty and response latency. *Social Cognition*, 29, 577–611.

Koriat, A. and Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, 103, 490–517.

Koriat, A. and Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, 52, 478–92.

Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology:* Human Learning and Memory, 6, 107–18.

Koriat, A., Bjork, R. A., Sheffer, L., and Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, 133, 643–56.

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

Koriat, A., Nussinson, R., Bless, H., and Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky and R. A. Bjork (Eds.) *Handbook of memory and metamemory*, pp. 117–35. New York: Psychology Press.

Koriat, A., Ackerman, R., Lockl, K., and Schneider, W. (2009). The memorizing-effort heuristic in judgments of learning: A developmental perspective. *Journal of Experimental Child Psychology*, 102, 265–79.

Kornell, N. (2011). Failing to predict future changes in memory: A stability bias yields long-term overconfidence. In A. S. Benjamin (Ed.) *Successful remembering and successful forgetting: a Festschrift in honor of Robert A. Bjork*, pp. 365–86. London: Psychology Press.

Kornell, N. and Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, 138, 449–68.

Kornell, N., Rhodes, M. G., Castel, A. D., and Tauber, S. K. (2011). The ease of processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, 22, 787–94.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases*, pp. 306–34. New York: Cambridge University Press.

Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin*, 126, 109–37.

Lindsay, D. S. (2008). Source Monitoring. In H. L. Roediger, III (Ed.) *Cognitive psychology of memory. Vol. 2 of Learning and Memory: A Comprehensive Reference* (J. Byrne, Ed.), pp. 325–48. Oxford: Elsevier

Markie, P. (2008). Rationalism vs. empiricism. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. Stanford, CA: The Center for the Study of Language and Information (CSLI), Stanford University.

McKenzie, C. R. M. (1997). Underweighting alternatives and overconfidence. *Organizational Behavior and Human Decision Processes*, 71, 141–60.

McKenzie, C. R. M. (2005). Judgment and decision making. In K. Lamberts and R. L. Goldstone (Eds.) *Handbook of cognition*, pp. 321–38. London: Sage.

Metcalfe, J. (2000). Feelings and judgments of knowing: Is there a special noetic state? *Consciousness and Cognition*, 9, 178–86.

Mitchell, K. J. and Johnson, M. K. (2000). Source monitoring: Attributing mental experiences. In E. Tulving and F. I. M. Craik (Eds.) *The Oxford handbook of memory*, pp. 179–95. New York: Oxford University Press.

Moser, P. K. (1988). Internalism and coherentism: A dilemma. *Analyses*, 48, 161–3.

Murphy, S. T., Monahan, J. L., and Zajonc, R. B. (1995). Additivity of nonconscious affect: Combined effects of priming and exposure. *Journal of Personality and Social Psychology*, 69, 589–602.

Myers, D. G. (2002). *Intuition: Its powers and perils*. New Haven, CT: Yale University Press.

Nagel, J. (2007). Epistemic intuitions. *Philosophy Compass* 2(6), 792–819.

Nelson, T. O. and Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The 'delayed-JOL effect.' *Psychological Science*, 2, 267–70.

Nelson, T. O., Gerler, D., and Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting prceptual identification and relearning. *Journal of Experimental Psychology: General*, 113, 282–300.

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

Parks, C. M. and Toth, J. P. (2006). Fluency, familiarity, aging, and the illusion of truth. *Aging, Neuropsychology, and Cognition*, 13, 225–253.

Peterson, C. R. and Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68, 29–46.

Plessner, H., Betsch, C., and Betsch, T. (2007) (Eds.) *Intuition in judgment and decision making*. Mahwah, NJ: Erlbaum.

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. Chicago, IL: University of Chicago Press.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 159, 271–95.

Proust, J. (2008). Epistemic agency and metacognition: An externalist view. *Proceedings of the Aristotelian Society*, 108, 241–68.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118, 219–35.

Reber, R., and Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, 8, 338–42.

Reder, L. M. (1988). Strategic control of retrieval strategies. In G. H. Bower (Ed.) *The psychology of learning and motivation: Advances in research and theory*, *Vol. 22*, pp. 227–59. New York: Academic Press.

Rescher, N. (1973). *Coherence theory of truth*. London: Oxford University Press.

Robinson, M. D., Johnson, J. T., and Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology*, 82, 416–25.

Ross, M. (1997). Validating memories. In N. L. Stein, P. A. Ornstein, B. Tversky, and C. Brainerd (Eds.) *Memory for everyday and emotional events*, pp. 49–81. Mahwah, NJ: Erlbaum.

Schervish, M. J. (1996). P Values: What they are and what they are not. *The American Statistician*, 50, 203–6.

Schwartz, B. L. (1998). Illusory tip-of-the-tongue states. *Memory*, 6, 623–42.

Schwartz, B. L. and Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1074–83.

Schwartz, B. L. and Metcalfe, J. (2011). Tip-of-the-tongue (TOT) states: Retrieval, behavior, and experience. *Memory & Cognition*, 39, 737–49.

Shafir, E., Simonson, I., and Tversky, A. (1993). Reason-based choice. *Cognition*, 49, 11–36.

Tulving, E., and Madigan, S. A. (1970). Memory and verbal learning. *Annual Review of Psychology*, 21, 437–84.

Tversky, A. and Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547–67.

Unkelbach, C. and Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness & Cognition*, 18, 22–38.

Van Cleve, J. (2005). Why coherence is not enough: A defense of moderate foundationalism. In M. Steup and E. Sosa (Eds.) *Contemporary debates in epistemology*, pp. 168–80. Oxford: Blackwell.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 582–600.

1 Walker, R. C. S. (1989). *The coherence theory of truth: realism, anti-realism, idealism*. London: Routledge.
2 Westcott, M. R. (1968). *Toward a contemporary psychology of intuition*. New York: Holt, Rinehart and
3 Winston.
4 Wright, J. C. (2010). On intuitional stability: The clear, the strong, and the paradigmatic. *Cognition*,
5 115, 491–503.
6 Yaniv, I. and Meyer, D. E. (1987). Activation and metacognition of inaccessible stored information:
7 Potential bases for incubation effects in problem solving. *Journal of Experimental Psychology:*
8 *Learning, Memory, and Cognition*, 13, 187–205.

Foundations of Metacognition

Foundations of Metacognition

Edited by

Michael Beran

Johannes L. Brandl

Josef Perner

Joëlle Proust

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2012
Impression: 1

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

ISBN 978–0–19–964673–9

Printed in Great Britain by
CPI Group (UK) Ltd, Croydon, CR0 4YY