Journal of Experimental Psychology: General
2011, Vol. 140, No. 1, 117–139


0096-3445/11/$12.00 DOI: 10.1037/a0022171

# Subjective Confidence in Perceptual Judgments: A Test of the Self-Consistency Model

Asher Koriat
University of Haifa

Two questions about subjective confidence in perceptual judgments are examined: the bases for these judgments and the reasons for their accuracy. Confidence in perceptual judgments has been claimed to rest on qualitatively different processes than confidence in memory tasks. However, predictions from a self-consistency model (SCM), which had been confirmed for general-information questions (Koriat, 2010) and social attitudes (Koriat & Adiv, 2010), are shown to hold true also for perceptual judgments. In SCM, confidence is modeled by the procedure for assessment of statistical level of confidence: For a 2-alternative, forced-choice item, confidence is based on the consistency with which the choice is favored across a sample of representations of the item, and acts as a monitor of the likelihood that a new sample will yield the same choice. Assuming that these representations are drawn from commonly shared populations of representations associated with each item, predictions regarding the basis of confidence were confirmed by results concerning the functions relating confidence and choice latency to interparticipant consensus and to intraparticipant consistency for majority and minority choices. With regard to the confidence-accuracy (C/A) relationship, the consensuality principle, documented for general-knowledge tasks (Koriat, 2008a), was replicated for perceptual judgments: Confidence correlated with the consensuality of the choice rather than with its correctness, suggesting that the C/A correlation is due to the relationship between confidence and self-consistency and is positive only as long as the correct choices are the consistently made choices. SCM provides a general model for the basis and accuracy of confidence judgments across different domains.

*Keywords:* subjective confidence, metacognition, self-consistency, perceptual comparisons, confidence–accuracy relationship

Research on metacognitive judgments and their accuracy has been conducted over the years within somewhat independent traditions. In the area of learning and memory, metacognition researchers have investigated the determinants and consequents of judgments of learning, feelings of knowing, and subjective confidence (see Dunlosky & Metcalfe, 2009; Koriat, 2007). In the area of judgment and decision making, a great deal of work has concerned the calibration of assessed probabilities, primarily the overconfidence bias that is generally observed when people assess the likelihood that their answers to two-alternative, forced-choice (2AFC) questions are correct (Lichtenstein, Fischhoff, & Phillips, 1982; Yates, 1990). Researchers in perception and psychophysics have long explored different quantitative theories of the processes underlying confidence in psychophysical judgments (Baranski & Petrusic, 1994; Juslin & Olsson, 1997; Vickers, Smith, Burt, & Brown, 1985). Forensic psychologists, in turn, have focused on applied questions, investigating the extent to which the confidence of different eyewitnesses is diagnostic of the accuracy of their testimonies (Bothwell, Deffenbacher, & Brigham, 1987; Read, Lindsay, & Nicholls, 1998; Sporer, Penrod, Read, & Cutler, 1995). Finally, research on metacognition has been flourishing in such areas as developmental psychology (Waters & Schneider, 2010), social cognition (Ross, 1997; Tormala & Rucker, 2007; Yzerbyt, Lories, & Dardenne, 1998), animal cognition (Kornell, 2009; Smith, Beran, Couchman, & Coutinho, 2008), and neuroscience (Lieberman, 2000; Modirrousta & Fellows, 2008).

In recent years, several points about metacognitive judgments have emerged as the core issues across these domains. Five such issues were listed by Koriat (2007). The first concerns the bases of metacognitive judgments: What are the processes underlying the monitoring of one's own knowledge (e.g., Koriat, 2007; Schwartz, 1994)? The second concerns the accuracy of metacognitive judgments—the correspondence between subjective and objective indices of knowing and the factors that affect that correspondence (e.g., Dunning, Heath, & Suls, 2004; Schwartz & Metcalfe, 1994). The third concerns the processes underlying the accuracy and inaccuracy of metacognitive judgments (Koriat, 1995). The fourth and fifth concern the effects of monitoring on control (Nelson & Narens, 1990): Assuming that metacognitive judgments exert measurable effects on cognition and behavior (Nelson, 1996), what are these effects, and how do they influence actual performance (e.g., Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1999; Koriat & Goldsmith, 1996; Son & Schwartz, 2002)?

The research reported in this study concerns specifically subjective confidence in perceptual judgments, but a model of sub-

Asher Koriat, Department of Psychology, University of Haifa.

This research was conducted at the Institute of Information Processing and Decision Making. I am grateful to Shiri Adiv for her help in all stages of the project, to Dana Klein and Rinat Gil for their help in the analyses, and to Hila Sorka, Ela Kniaz, Shani Finkelstein, and Naama Yarkonei for their help in preparing the stimuli and conducting the experiments.

Correspondence concerning this article should be addressed to Asher Koriat, Department of Psychology, University of Haifa, Haifa 31905, Israel. E-mail: akoriat@research.haifa.ac.il

jective confidence is examined that appears to have some generality across domains. The model has some bearing on all of the questions mentioned earlier. Predictions derived from the model have been found to hold for confidence in general knowledge (Koriat, 2010). They were also confirmed for the degree of certainty with which people hold their social attitudes and the speed with which the attitude is formed (Koriat & Adiv, 2010). Both of these have been seen as measures of attitude strength (Bassili, 1996), which is assumed to affect the likelihood that attitudes will be translated to behavior. Results indicating that the model's predictions also hold true for perceptual comparisons could have implications regarding the question of whether confidence in sensory–perceptual tasks is based on the same or different processes than confidence in general knowledge (e.g., Winman & Juslin, 1993).

I begin by reviewing the work in which confidence in perceptual tasks is compared with confidence in general knowledge. I then outline the self-consistency model (SCM) of subjective confidence and describe its predictions regarding the basis of confidence judgments and the processes underlying their accuracy.

## Comparing Confidence for Perceptual Tasks and General-Knowledge Tasks

Several researchers argued that there is a fundamentally different basis for confidence in perception and in general knowledge. One of the key findings cited in support of this argument is that confidence judgments in tasks tapping general knowledge typically yield an overconfidence bias (Lichtenstein et al., 1982; see Brenner, Koehler, Liberman, & Tversky, 1996; Griffin & Brenner, 2004; Hoffrage, 2004). That is, people tend to express confidence in their judgments that exceeds the accuracy of those judgments. In contrast, their expressions of confidence in their judgments involving perceptual tasks generally evidence either good calibration or even a tendency toward underconfidence. The idea that perceptual judgments are more accurate than what people feel dates back to Peirce and Jastrow (1884) and Fullerton and Cattell (1892; see Björkman, Juslin & Winman, 1993). These researchers observed that participants' responses to perceptual judgments tend to be correct even when the participants feel that their responses were based on a complete guess. In later years, Dawes (1980) argued that the overconfidence bias in general-knowledge tasks derives from people's tendency to overestimate the power of their intellectual knowledge, but because the sensory coding system is remarkably accurate, perceptual tasks may not be vulnerable to such overconfidence bias. Although his results were generally in line with this proposal, they were not conclusive. However, a follow-up study by Keren (1988) provided clearer evidence in support of Dawes' proposal. Keren found overconfidence in a general-knowledge task but not in a perceptual task. He proposed that the more perception-like a task is, the less it should be likely to induce overconfidence bias. However, when a perceptual task requires additional higher processing beyond sensory encoding, various cognitive distortions may impair participants' ability to monitor their performance accurately.

Other researchers also observed systematic differences in calibration between perceptual tasks and general-knowledge tasks. Björkman et al. (1993) and Winman and Juslin (1993) reported evidence for a *pervasive underconfidence bias* in perceptual tasks and concluded that a qualitative difference exists between the two tasks in the process underlying confidence judgments. A conceptualization of this difference was proposed by Juslin and Olsson (1997) in terms of the distinction between two origins of uncertainty. According to them, Thurstonian uncertainty is characteristic of sensory discrimination tasks in which the uncertainty derives from random noise in the nervous system that causes moment-to-moment variations in the sensations. In contrast, Brunswikian uncertainty, which is characteristic of general-knowledge tasks, derives from a less-than-perfect validity of the cues that people use to infer their answers. Because all participants have become adapted to the same environment, they may be expected to rely on the same cues, and hence a correlation is expected between participants in the proportion of incorrect answers to different items. Juslin and Olsson (1997) proposed that separate models of confidence are needed for the two types of tasks.

Baranski and Petrusic (1994, 1995, 1999) disputed this view, reporting very similar patterns of results for confidence in general-knowledge and perceptual tasks. They observed a comparable "hard–easy effect" for the two tasks: underconfidence for easy judgments and progressive overconfidence as item difficulty increased. Also, for both tasks, confidence tended to decrease with the latency of making a choice. While Baranski and Petrusic admitted that the two classes of tasks differ quite dramatically in terms of the nature of the information on which the decision process is based (sensory stimulation vs. long-term memory), they favored a unified theoretical account of calibration in which a common basis for confidence in the two types of tasks is assumed (see also Merkle & Van Zandt, 2006; Petrusic & Baranski, 1997).

Gigerenzer, Hoffrage, and Kleinbölting (1991) also argued against granting a special status to perceptual tasks as far as confidence judgments are concerned. According to their theory of probabilistic mental models (PMM), individuals' choice of an answer to a 2AFC item, as well as their confidence in that answer, is based on cues that they use to infer the answer. Such is assumed to be the case for both general-knowledge and perceptual tasks. The overconfidence bias that has been observed for general-knowledge tasks is believed to derive from the failure of researchers to sample items that are representative of the natural environment. When a representative set of items is used, confidence judgments should be well calibrated, and this should be also true in the case of perceptual tasks.

Much of the work contrasting confidence for perceptual and general-knowledge tasks has concerned calibration, that is, the absolute correspondence between confidence and accuracy (Griffin & Brenner, 2004). A second aspect of confidence accuracy, however, is resolution or relative confidence. Resolution refers to the within-person confidence/accuracy (C/A) correlation, which reflects the ability of participants to discriminate between correct and incorrect answers (Baranski & Petrusic, 1995; Nelson, 1984; Yaniv, Yates, & Smith, 1991; Yates, 1990). It should be stressed that resolution can be perfect when calibration is very low and vice versa (Koriat & Goldsmith, 1996).

The motivation for SCM derived from observations concerning the resolution of confidence judgments in general-knowledge tasks. For many years, this resolution has troubled psychologists, who ask: How do people discriminate between correct and wrong answers? In attempting to address this question, Koriat (2008a) reviewed several observations that suggest that confidence judg-

ments are actually correlated with the consensuality of the response rather than with its correctness. SCM was developed primarily to explain this principle by specifying the basis of confidence judgments. However, this model also yielded several novel predictions, including some that concern calibration. In the following section, I review the evidence for the consensuality principle. This principle will be shown to hold true for perceptual judgments as well. I will then turn to examine SCM itself and its predictions regarding perceptual judgments.

## The Resolution of Confidence Judgments: The Consensuality Principle

When participants choose an answer to 2AFC general-information questions and indicate their confidence in their answers, a moderate-to-high C/A correlation is generally observed, suggesting that people can monitor the accuracy of their answers. However, results reported by Koriat (2008a) suggest that this correlation is simply due to the fact that for a typical set of general-knowledge questions, participants are more often correct than wrong so that the correct answer is the one that is consensually chosen. Therefore, for a typical set of items, it is unclear whether in making confidence judgments, individuals discriminate between correct and wrong answers or between consensual and nonconsensual answers. Indeed, in several studies in which researchers dissociated correctness from consensuality by including a sufficiently large number of items for which participants agreed on the wrong answer, confidence was found to correlate with the consensuality of the answer rather than with its correctness. Thus, for consensually correct (CC) items, in which most participants chose the correct answer, the C/A correlation was positive, whereas for consensually wrong (CW) items, the C/A correlation was negative. This pattern was observed for a word-matching task in which participants guessed the meaning of words from noncognate languages (Koriat, 1976), for feelings-of-knowing judgments about an elusive memory target (Koriat, 1995), for confidence in 2AFC general-information questions (Koriat, 2008a), and also for the memory of studied sentences (Brewer & Sampaio, 2006; see also Sampaio & Brewer, 2009). Of course, in none of these studies were participants informed about others' answers. Nevertheless, a participant's confidence in his or her choice increased systematically with the proportion of other participants who made that choice.

These results clearly speak against the direct-access view of metacognition (see Schwartz, 1994), according to which participants have privileged access to the accuracy of their answers. Rather, although participants are generally successful in discriminating between the correct answer and the wrong answer, their success seems to derive from their reliance on some cues that are generally diagnostic of the accuracy of the answer. These cues would seem to underlie the consensuality of the response—the extent to which it is found compelling by the majority of participants. Thus, the C/A relationship is due to a confidence/consensuality (C/C) relationship. The results demonstrate the intimate link between metaknowledge accuracy and knowledge accuracy: Metaknowledge is accurate because knowledge itself is accurate (Koriat, 1993).

In this study, then, one of my aims was to examine whether the consensuality principle holds true also for confidence in perceptual tasks. Participants judged which of two lines was longer (Experiment 1) or which of two shapes covered more area (Experiments 2) and indicated their confidence in their answer. On the basis of the results, the items were classified as CC or CW items. I examined whether the correct answer was associated with stronger confidence than the wrong answer for the CC items and whether the wrong answer was associated with stronger confidence for the CW items.

My more central aim, however, was to test predictions from SCM as a general account of the basis of subjective confidence that could also explain the consensuality results. The assumption was that understanding the cognitive basis of confidence judgments could provide a key for understanding the accuracies and inaccuracies of these judgments.

## The Self-Consistency Model of Subjective Confidence

What is the basis of confidence judgments? Most current discussions of metacognitive judgments reflect a cue-utilization approach (vs. a direct-access view), according to which metacognitive judgments are inferential in nature, based on a variety of cues. A distinction is customarily drawn between two classes of cues that give rise, respectively, to information-based or to experience-based metacognitive judgments (see Koriat, Nussinson, Bless, & Shaked, 2008). Consider 2AFC general-information questions. Information-based confidence judgments are assumed to rest on an analytic inference in which individuals retrieve and consult various considerations to reach an answer and also to yield an educated assessment of the likelihood that the answer is correct (Griffin & Tversky, 1992; Juslin, Winman, & Olsson, 2003; Koriat, Lichtenstein, & Fischhoff, 1980; McKenzie, 1997, 1998; Nelson & Narens, 1990). In contrast, experience-based confidence judgments are said to rest on the feedback from the very experience of making a choice. They may be based on such mnemonic cues as the degree of deliberation or conflict experienced in making a choice and on the time and effort it took to reach that choice (Kelley & Lindsay, 1993; Koriat, Ma'ayan, & Nussinson, 2006; Robinson, Johnson, & Herndon, 1997). These contentless mnemonic cues are assumed to give rise directly to a subjective feeling of certainty or uncertainty (see Koriat, 2000, 2007). Indeed, when participants were asked to list four reasons in support of their answer, their confidence in the answer was lower than when they were asked to list only one supporting reason (Koriat et al., 2008; see also Haddock, Rothman, Reber, & Schwarz, 1999). Presumably, retrieving more reasons is experienced subjectively as more difficult than retrieving fewer reasons, so that the effects of mnemonic cues (ease of retrieval) can override the effects of the content of declarative arguments.

Koriat (2010) proposed that participants facing a 2AFC general-information question typically engage in an analytic-like process, retrieving and consulting different considerations. Participants need not articulate or be fully conscious of these considerations. Once they settle on an answer and have to assess their confidence, they do not go over the entire deliberation process again but rely on the gist of the process that they used to determine the choice (Stephen & Pham, 2008). That gist consists of such gross cues as the feeling of conflict or doubt they had experienced in making a choice, the amount of effort invested, and the time it took to reach the choice. These mnemonic cues represent the feedback from the process of making a choice and mirror roughly the balance of

evidence in favor of the alternative answers. In general, as participants move from making a choice to assessing their degree of confidence in that choice, the contribution of information-driven processes decreases and that of mnemonic cues increases.

Although the process underlying the choice of an answer may be quite complex, the retrospective review of that process, which underlies the participants' confidence, can be modeled by a simple majority vote. From the perspective of the process underlying confidence, the choice process is conceptualized as involving a sequential sampling of different representations of the item (Koriat, 2010). In each iteration, participants draw a representation of the item and reach a tentative choice (i.e., a subdecision). The sampling of representations continues until a preset sample size has been reached or until a series of representations yields the same subdecision a number of times in succession (e.g., three times, see Audley, 1960). The ultimate overt choice is assumed to represent the choice most consistently favored across the series of subdecisions. Subjective confidence in that choice is based primarily on the degree of consistency among the subdecisions, that is, on the proportion of representations supporting the chosen answer.

Underlying this model is the assumption that participants behave essentially like intuitive statisticians who attempt to make conclusions about a population on the basis of a sample of observations. By repeating the choice procedure several times, participants obtain an assessment of the amount of unreliability and doubt involved. Unreliability then affects subjective confidence in the same way that sample variance affects statistical level of confidence. Also, like statistical level of confidence, subjective confidence is assumed to represent an assessment of reproducibility—the likelihood that a new sample of representations drawn from the same population will yield the same choice. Thus, according to SCM, reliability is used as a cue for validity: Although confidence judgments are construed subjectively as pertaining to the correctness of the answer, such judgments actually act as monitors of the consistency with which that answer is favored across representations and the likelihood that it will be chosen when the item is presented again.

In addition, although the major cue for confidence is self-consistency, a frugal cue for self-consistency is choice latency, that is, the amount of time it took to reach an answer. Results reported by Koriat (2010) and Koriat and Adiv (2010) indicate that differences in choice latency mimic rather faithfully differences in self-consistency, so that reliance on choice speed as a cue for confidence can yield the same pattern of confidence judgments as that expected for consistency-based confidence.

The model described can be extended to perceptual comparison tasks. In the case of general-information questions (Koriat, 2010) and social attitudes (Koriat & Adiv, 2010), some variability is assumed in the representations that are sampled in making a choice. The term *representation* is used to refer broadly to any interpretation of a question or a statement or to any consideration that may speak for one response option rather than the other. A similar assumption may be made with regard to perceptual comparison tasks, particularly when these tasks require some higher processing beyond sensory encoding (Keren, 1988). When participants are presented with a perceptual-comparison task for which the answer is not immediately obvious (the *uncertainty zone*, see Winman & Juslin, 1993), it is by exploring different aspects or features of the stimuli that they appreciate the degree of doubt or certainty associated with their decision. Confidence in the answer can then be modeled by a process in which participants sample different representations from a pool of representations and base their confidence on the degree to which these representations argue consistently for one of the choices. Several authors have postulated a fluctuation that occurs in the encoding of stimuli (e.g., Bower, 1972; Estes, 1950; Juslin & Olsson, 1997). This fluctuation can be spontaneous but can also be subject to intentional effects. Various factors that affect the fluctuation of perceptual representations have been discussed in the context of multistable perceptual phenomena (Palmer, 1999). Of course, memory tasks and perceptual tasks differ in the nature of the representations that are explored (Baranski & Petrusic, 1995). However, assuming that confidence rests on structural cues, such as the amount of deliberation experienced or the time it took to reach a choice, then the predictions of SCM may hold true for perceptual judgments as well.

## Predictions of the Self-Consistency Model

Several gross predictions that derive from SCM have been tested and have been generally confirmed across several 2AFC tasks (Koriat, 2010; Koriat & Adiv, 2010). These predictions, which were tested for perceptual judgments, will now be outlined.

One prediction concerns reproducibility: Confidence acts as a monitor of the likelihood that the same choice would be made in a subsequent encounter with the same item. Results in support of this idea were obtained for a word-matching task and for general-information questions (Koriat, 2010). They were also confirmed for confidence in social attitudes (Koriat & Adiv, 2010). In Experiments 1 and 2 of the present study, participants were presented five times with the same set of perceptual comparisons. Assuming that confidence in one's answer is based on self-consistency, it would be expected that confidence in a choice in the first block should predict the likelihood of repeating that choice in subsequent blocks.

This prediction assumes that in responding to a perceptual-comparison task, participants sample representations from roughly the same population of representations. Indeed, a cardinal assumption underlying SCM is that in responding to 2AFC items, whether they involve general-information questions or beliefs and attitudes, participants with the same experience and beliefs draw representations largely from the same, commonly shared population of representations associated with each item. This assumption is critical for the explanation of the consensuality results but also provides additional predictions pertaining to the effect of within-participant consistency and cross-participant consensus on confidence judgments. A similar assumption may be made with regard to perceptual comparisons: In attempting to choose between two response options, participants sample representations of the stimuli largely from the same pool of representations.

Assuming that each item is associated with a population of representations, the most important property of that population is the proportion of representations favoring the most dominant majority answer. This proportion is designated $p_{\mathrm{maj}}$ (Koriat, 2010). Let us assume tentatively that in responding to each item, participants draw a sample of seven representations, each of which yields a binary subdecision, and that the overt choice is dictated by the majority vote. A simple index of self-consistency is the magni-

tude of the majority vote. However, Koriat (2010) and Koriat and Adiv (2010) used an index of broader generality, which is related to the standard deviation of the subdecisions: $1 - \sqrt{\hat{p}\hat{q}}$, with the range .5–1.0. Figure 1A indicates the pattern that is predicted from the binomial distribution when samples of seven representations are drawn randomly from populations of representations that differ in $p_{\text{maj}}$. Three features should be noted in this figure. First, mean self-consistency (and hence confidence) for each item should increase with $p_{\text{maj}}$, the proportion of representations favoring the majority option. Second, however, self-consistency is systematically higher for majority than for minority choices. Finally, whereas for majority choices, self-consistency increases steeply with $p_{\text{maj}}$; for minority choices, it decreases but much more shallowly.

Why should self-consistency differ for majority and minority choices? The answer lies in the relationship between the mean and the variance. For example, with $p_{\text{maj}} = .75$, a sample of seven representations has a .445 likelihood of yielding six or seven representations that favor the majority choice. In contrast, the likelihood that it will yield six or seven representations that favor the minority choice is only .001. In general, then, as long as $p_{\text{maj}}$ differs from .50, minority samples should have a lower self-consistency and hence lower confidence than majority samples.

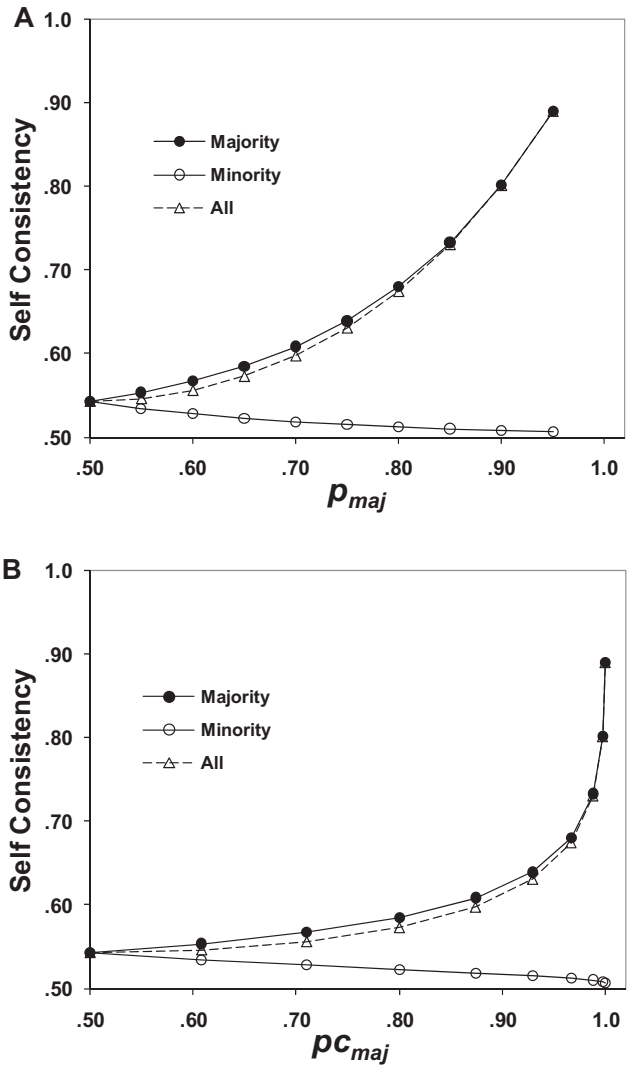


*Figure 1.* Panels A and B present the expected self-consistency scores for majority and minority answers as a function of $p_{maj}$ (proportion of representations favoring the most dominant, majority answer) and $pc_{maj}$ (likelihood of choosing the majority answer), respectively, in Experiment 1.

Self-consistency is expected to increase with increasing $p_{\text{maj}}$, but a similar increase, somewhat more accelerated, is expected for the likelihood of choosing the majority answer, which is designated $pc_{\text{maj}}$. For example, when $p_{\text{maj}} = .75$, samples of seven can be expected to lead to a .93 proportion of participants' choosing the majority alternative. Figure 1B plots the same functions as Figure 1A, but the $p_{\text{maj}}$ values in the *x*-axis were replaced with the corresponding $pc_{\text{maj}}$ values. This figure specifies a set of testable predictions because the $pc_{\text{maj}}$ associated with an item can be estimated. It can be estimated first from response consistency—the proportion of times that the most frequent response is selected by a participant across repeated presentations of the same item. Second, it can be estimated from response consensus—the proportion of participants making the consensually preferred choice. Because these two properties are assumed to be specific to each item, they will be referred to as *item consistency* and *item consensus*, respectively.

Two sets of predictions could be tested in Experiments 1 and 2. The first concerned within-person reliability in responding to the same item. Assuming some fluctuation across blocks in the sampled representations underlying choice, one would expect that confidence in a choice should vary with the choice made. Thus, if the responses to each item were divided for each participant between those that were more frequent and those that were less frequent across blocks, confidence in the frequent response would be expected to be higher on average than confidence in the rare response. Furthermore, item consistency can be defined for each participant and item as the proportion of times that the more frequent response was made across blocks. Confidence in the frequent response was expected to increase steeply with item consistency for the frequent responses, but to decrease, somewhat more shallowly, for the rare responses.

A similar set of predictions could be made with regard to cross-participant consensus. On the basis of the results from Block 1 only, a majority (consensual) choice can be defined for each item as the choice most frequently made across participants. It was predicted that confidence would be higher for the majority choice than for the minority choice. This prediction accords with the consensuality results of Koriat (2008a). It implies that for the same item, those participants who choose the more popular response should express greater confidence in their choice than those who choose the less popular response. Note that this prediction follows from a simple stochastic model in which no interaction between participants was assumed. In addition, confidence in the majority choice was expected to increase with item consensus—the proportion of participants who endorsed the majority choice. In contrast, confidence in the minority choice should tend to decrease with item consensus.

Another set of predictions concerned choice latency. As noted earlier, a relationship between choice latency and confidence has been found in several studies, suggesting that confidence is influenced by the ease with which a decision is reached (Kelley & Lindsay, 1993; Koriat et al., 2006; Nelson & Narens, 1990; Rob-

inson et al., 1997; Zakay & Tuvia, 1998). In order to accommodate the effects of choice latency, I supplemented SCM with the assumption that the sequential sampling of representations underlying choice terminates when a sequence of representations yields the same subdecision three times in a row (see Audley, 1960). Assuming that choice latency reflected the actual number of representations needed to reach a decision, I found that a simulation experiment yielded results that closely mimicked those expected for the theoretical parameter of self-consistency and also those demonstrated for confidence (Koriat, 2010; Koriat & Adiv, 2010). Furthermore, several empirical findings indicated that the results obtained with response latency mimicked those obtained for confidence as far as the effects of item consensus and majority versus minority choices are concerned. These results suggest that choice latency can serve as a frugal cue for self-consistency so that reliance on choice latency as a cue for confidence can yield a pattern of results similar to that expected when confidence is based on self-consistency. Thus, although choice latency is not affected always by the same variables that affect confidence (see Petrusic & Baranski, 2009; Wright & Ayton, 1988), it is claimed to be sensitive to self-consistency. Therefore, the predictions for response speed are similar to those stated for confidence.

Another set of predictions concerned the accuracy of confidence judgments and, specifically, the consensuality principle. According to SCM, the consensuality of a response to a given item is diagnostic of the degree of internal consistency within the sample of representations underlying choice and confidence for that item. To the extent that different participants draw representations from a commonly shared pool of representations, the larger the proportion of participants who opt for a response, the stronger will be the confidence associated with that response. Therefore, the consensuality pattern that has been observed for confidence in general-information questions (Koriat, 2008a) would be expected to be observed also for perceptual judgments. To test this prediction required a sufficiently large set of stimuli for which the consensually endorsed answer was likely to be the wrong answer. Thus, a classification of items as CC and CW was expected to yield the expected interaction between confidence and accuracy: The C/A correlation should be positive for the CC items and negative for the CW items. A similar pattern of results was expected for response latency: For CC items, response latency should be faster for correct answers, whereas for CW items, it should be faster for wrong answers.

Because the task was repeated five times in both Experiments 1 and 2, we were able to examine whether a consistency principle, analogous to the consensuality principle, also was found in a within-individual analysis. Thus, if items were classified for each participant according to whether the frequent choice was the correct choice or the wrong choice, the C/A relationship would be positive for the former items but negative for the latter items. A similar pattern should be observed for the relationship between response speed and accuracy.

The final predictions concern calibration. These predictions will be detailed in the introduction to Experiment 2. Briefly, according to SCM, the overconfidence bias that has been observed for general-knowledge questions derives largely from the discrepancy between reliability and validity: Whereas confidence judgments serve as monitors of reliability or self-consistency, these judgments are evaluated in calibration studies against correctness. However, reliability is virtually always higher than validity. In Experiment 2, I examined whether this is also true of confidence in perceptual judgments. We also examined whether these judgments exhibit a different pattern of calibration than that characteristic of knowledge questions when calibration is evaluated against some criterion of reliability.

## Experiment 1

In Experiment 1, participants were presented with 40 pairs of line drawings. For each pair, they chose the one that was longer and indicated their confidence in the correctness of their choice. The task in its entirety was repeated for a total of five times (blocks).

### Method

**Participants.** Thirty-nine psychology undergraduates (24 women and 15 men) participated in the experiment either for pay or for course credit.

**Stimulus materials.** The experimental materials consisted of 40 different line drawings. These were paired to form 40 pairs so that each drawing appeared twice, but each time it was paired with a different drawing (see Figure 2 for examples). Ten additional line drawings were used to create the practice trials. Each of the stimuli subtended a visual angle of approximately 5.8°.

The pairing of the stimuli was based on the results of an exploratory study in which the likelihood of making a correct answer to each pair was estimated. On the basis of these results, I planned the stimulus pairs used in the experiment to yield a sufficiently large number of pairs on which participants would be likely to agree on the wrong answer.

**Apparatus and procedure.** The experiment consisted of five blocks. In each block, the entire set of 40 pairs was presented, preceded by two practice pairs. The experiment was conducted on an IBM-compatible personal computer. Participants were told that for each pair, they should judge which of the two lines was longer. They were instructed to click a box labeled *1* or *2* below the line that they judged as being longer and then to click the box labeled *Confirm* that appeared beneath the previous boxes. After clicking the confirm box, they had to judge on a 0–100 scale how confident they were in their response.[1] They were encouraged to use the full range of the confidence scale.

Each trial began when the participants clicked a box labeled *Show line drawing*. The two stimuli then appeared side by side, labeled *1* and *2*, respectively. Each pair remained on the screen until the participants indicated their response by clicking *1* or *2* with the mouse. The computer program measured response latency, defined as the interval between the presentation of the pair and the confirmation response. After participants clicked the confirm box, a confidence scale (0–100) was added beneath the figures, and participants marked their confidence in their answer by sliding a pointer on a scale using the mouse (a number in the range 0–100 corresponding to the location of the pointer on the screen appeared in a box). After participants clicked a second confirm box, the show line drawing box appeared on the screen,

[1] The confidence judgments in Experiment 1 were measured on a 0–100 scale because of an attempt to compare the choice-independent confidence effect (Koriat, 2008b) in this experiment with that of other tasks for which that scale was appropriate (see General Discussion).

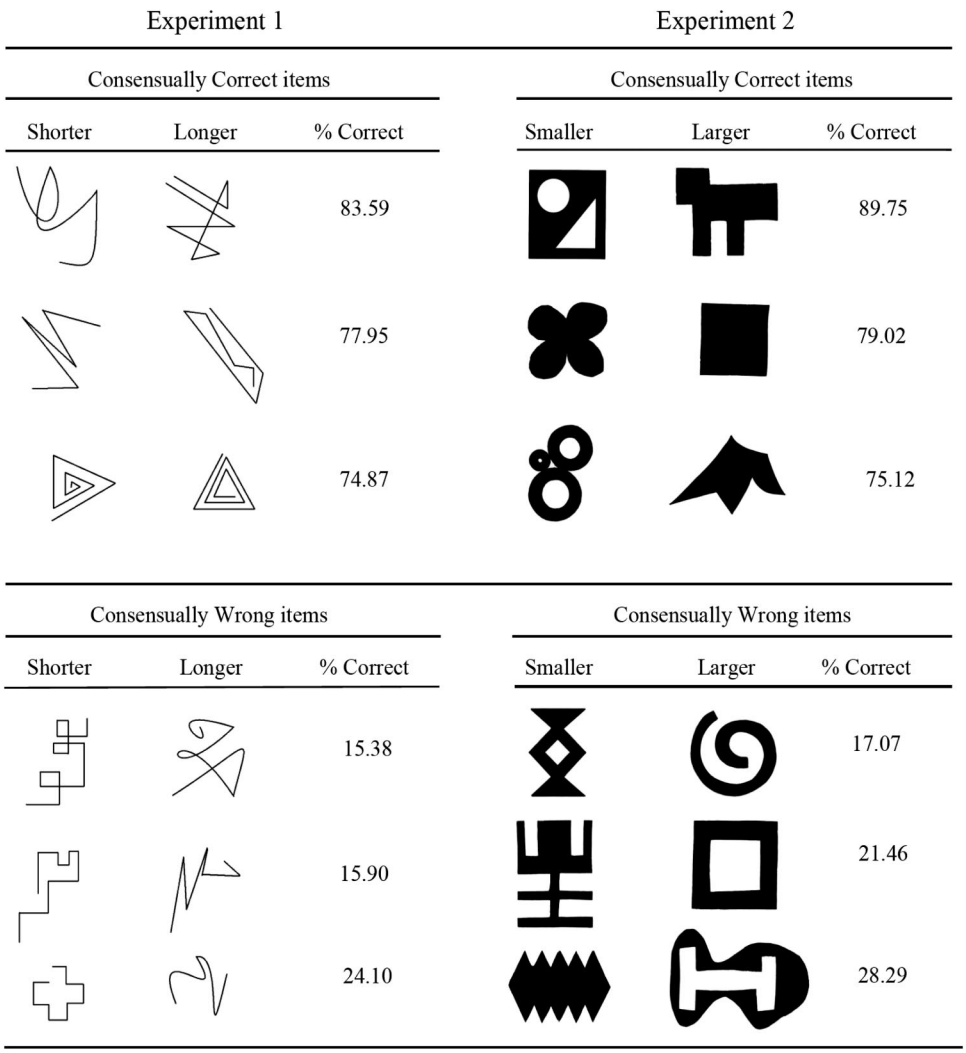


*Figure 2.* Examples of the stimuli used in Experiments 1 and 2, divided into those for which the consensual answer was the correct answer (consensually correct) and those for which the consensual answer was the wrong answer (consensually wrong).

and the next trial began. The order of the 40 experimental pairs was determined randomly for each participant and for each block. There were short breaks between the blocks. The experiment lasted about 45 min.

## Results and Discussion

By and large, participants tended to give the same response to each pair across the five blocks. Thus, the probability of making the Block-1 response again over the next four blocks averaged .76 across participants.

The results were organized around four topics: (a) reproducibility, (b) response consistency, (c) response consensus, and (d) the consensuality principle. Within each topic, the results for confidence judgments are presented first, followed by those for choice latency. In the final section, several analyses that connect some of the previously mentioned topics are presented.

**Reproducibility.** The assumption that confidence acts as a monitor of reliability implies that confidence in a choice predicts the likelihood that an individual will make the same choice in a subsequent presentation of the item. To examine this possibility, I grouped the confidence judgments in Block 1 into six categories, and calculated repetition proportion—the likelihood of making the Block-1 response across the subsequent four blocks—across all participants and items. The results are presented in Figure 3A. The function is monotonic; the Spearman rank-order correlation over the six values was .94, $p < .005$.[2]

Choice speed also predicted reproducibility. In all of the analyses of choice latency reported in this article, latencies that were below or above 2.5 *SD*s from each participant's mean latency for each block were eliminated (3.2% across all blocks). The choice

[2] Other binning procedures led to similar results.

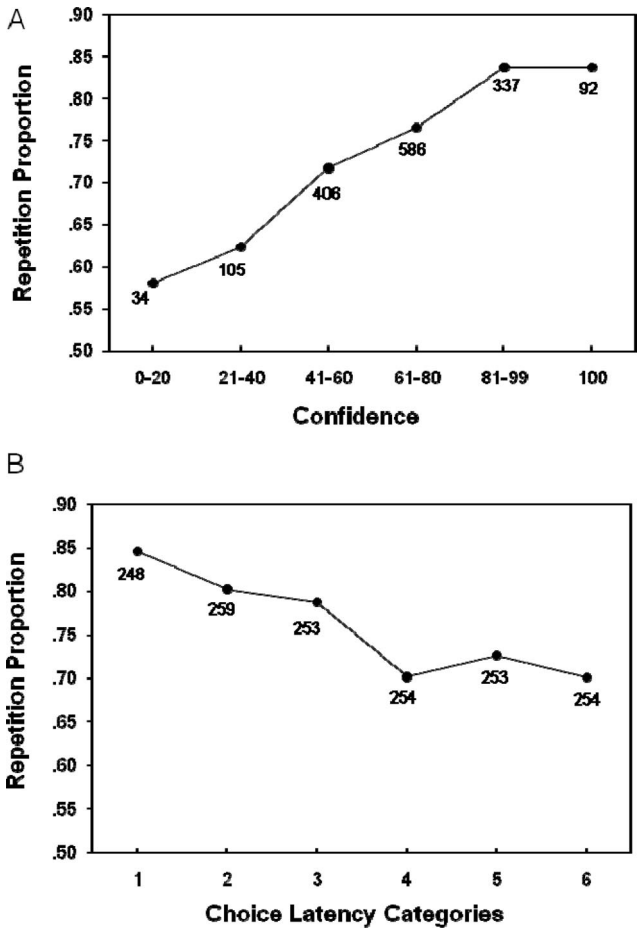


*Figure 3.* Panel A presents the likelihood of repeating Block-1 choices in the subsequent four blocks (repetition proportion) as a function of confidence in these choices in Block 1. Panel B plots repetition proportion as a function of Block-1 choice latency. The number of observations in each category is also included in the figures (Experiment 1).

latencies in Block 1 were grouped into six categories of about the same frequencies across all participants. It can be seen (Figure 3B) that the likelihood of repeating a choice decreased with the time taken to reach that choice in Block 1; the rank-order correlation across the six points was $-.94$, $p < .005$.

**Confidence and choice latency as a function of within-person response consistency.** Assuming that participants sample representations from roughly the same population in each block, I predicted that mean confidence in a choice should increase with item consistency—the number of times that the modal choice is made across the five blocks. However, the more frequent choice should be associated with higher confidence and shorter choice latencies than the less frequent choice.

***The relationship between confidence and within-person response consistency.*** Figure 4A presents mean confidence for the participant's frequent and rare responses as a function of item consistency, that is, the number of times that the frequent response was chosen (item consistency = 3, 4, or 5). Mean confidence increased monotonically with item consistency. However, using only the partial-consistency items (item consistency = 3 or 4), participants were more confident when they chose their more frequent response (62.71) than when they chose their less frequent response (59.03), $t(38) = 5.22$, $p < .0001$. This pattern was exhibited by 31 participants, $p < .0005$, by a binomial test. In addition, confidence in the frequent response increased with item consistency (3 vs. 4), $t(38) = 2.11$, $p < .05$, whereas confidence in the less frequent response decreased with item consistency, $t(38) = 2.03$, $p < .05$. This pattern is precisely what follows from SCM.

***The relationship between choice latency and within-person response consistency.*** Similar analyses were performed on choice latency (see Figure 4B). Choice latencies were shorter for the frequent responses (5.37 s) than for the rare responses (7.28 s), $t(38) = 5.61$, $p < .0001$. In addition, they decreased with item consistency for the frequent choices but increased with item consistency for the rare choices. This pattern roughly mimics the respective pattern for confidence judgments (Figure 4A).

***The postdiction of confidence and latency from response repetition.*** The increase in confidence with item consistency might be due to carry-over effects across repeated presentations. Indeed, confidence sometimes increases with repeated solicitation of a judgment (Hasher, Goldstein, & Toppino, 1977; Holland, Verplanken, & van Knippenberg, 2003; Shaw, 1996). Although this was

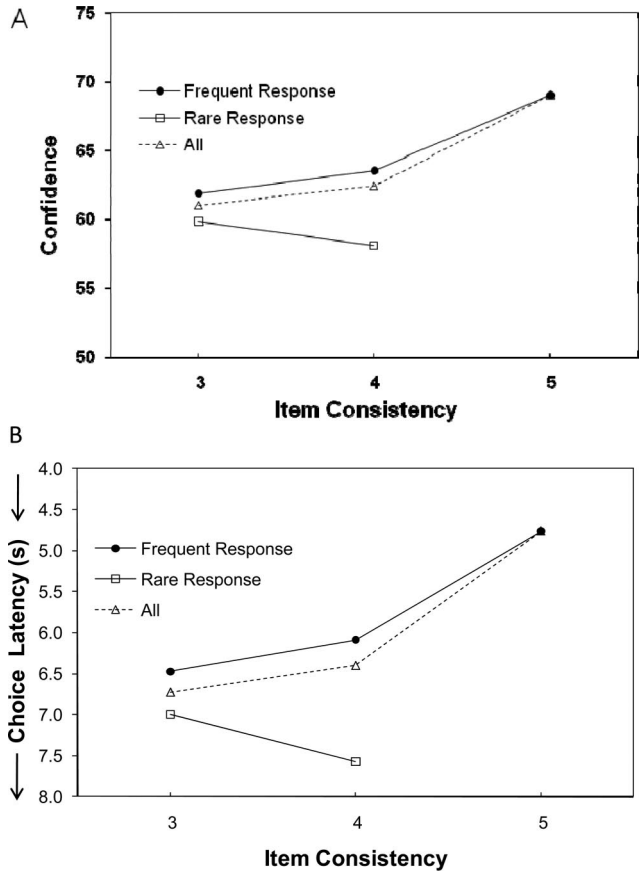


*Figure 4.* Panel A presents mean confidence for each participant's frequent and rare responses as a function of item consistency (the number of times that a response was made across all five blocks). Panel B presents mean choice latency as a function of item consistency (Experiment 1).

not true in the present study (confidence judgments averaged 67.8, 66.31, 65.34, 65.34, and, 66.18, respectively, for Blocks 1–5), it is important to establish that the results on within-person consistency do not derive entirely from carry-over effects across blocks. Such effects are implied by Finn and Metcalfe's (2007) work on the memory for past-test heuristic: Participants may become aware of the repetition of the stimuli and then consult their memory regarding which option they previously selected. If reliance on the memory for past test is associated with higher confidence, then this may explain the observed differences between frequent and rare choices.

In order to show that such is not entirely the case, I attempted to postdict the Block-1 confidence from the frequency with which the Block-1 choice was made across the subsequent blocks. For each participant, each Block-1 choice was classified into one of two categories depending on whether it was repeated two or more times in the subsequent four blocks or one time or not at all. Confidence for the two categories averaged 69.18 and 60.34, respectively, across 39 participants who had both means, $t(38) = 6.57$, $p < .0001$. When the full-consistency items (item consistency = 5) were eliminated for each participant, the respective means were 64.94 and 60.34, $t(38) = 3.39$, $p < .002$. Thus, even for Block-1 responses, confidence discriminates between the more frequent and the less frequent responses: Responses that were made more often across the five blocks yielded higher confidence in Block 1 than responses that were made less often.

**Confidence and choice latency as a function of cross-person response consensus.**

***The relationship between confidence and cross-person consensus.*** According to SCM, different participants also sample representations from roughly the same population of representations characteristic of each item. It follows that the pattern relating confidence to cross-person consensus should be similar to that observed for within-person consistency.

The analyses of the effects of cross-person consensus were carried out across all five blocks. For each of the 40 items, the number of times that each of the two answers was chosen (across the 39 participants × 5 presentations) was determined, and the answer chosen most often was defined as the consensual or majority answer. Item consensus, defined as the percentage of choices of the consensual response to each item, averaged 78.1% across items, (range, 53–100%). For one item, all participants gave the same response throughout the five blocks.

Figure 5A presents mean confidence judgments for majority and minority responses for each of six item-consensus categories (51–59, . . . 90–99, 100). Mean overall confidence judgments increased monotonically with item consensus, as predicted, and when mean confidence and mean item consensus were calculated for each item, the correlation between them over all 40 items was .83 ($p < .0001$). However, when the majority response was chosen, it was endorsed with higher confidence (66.89) than when the minority response was chosen (59.71), $t(38) = 6.43$, $p < .0001$. This difference was consistent across items: For 32 items, confidence was higher for the majority than for the minority response compared with seven items in which the pattern was reversed, $p < .0001$, by a binomial test.

Because for each item the confidence means for majority and minority responses were based on different participants, the results just presented could reflect a between-individual effect: Participants who tended to choose consensual responses also tended to use relatively high confidence ratings. Consistent with previous findings, there were reliable individual differences in mean confidence judgments (see also Kleitman & Stankov, 2001; Stankov & Crawford, 1997): When mean confidence was calculated for each participant and for each block, the correlations across participants between the means for different blocks averaged .92. To control for interparticipant differences, I standardized the confidence judgments of each participant for each block so that the mean and standard deviation of each participant were the same as those of the raw scores across all participants. Average scores were then calculated for each item for majority responses and minority responses. The results were very similar to those exhibited for the raw scores (Figure 5A). Thus, across the 39 items, mean standardized confidence ratings was 66.54 for majority responses, compared with 61.33 for minority responses, $t(38) = 6.06$, $p < .0001$.

The results presented in Figure 5A indicate that confidence in majority responses increased monotonically with item consensus: The correlation across all 40 items was .85, $p < .0001$ (.86 for the standardized scores, $p < .0001$). In contrast, confidence in minor-

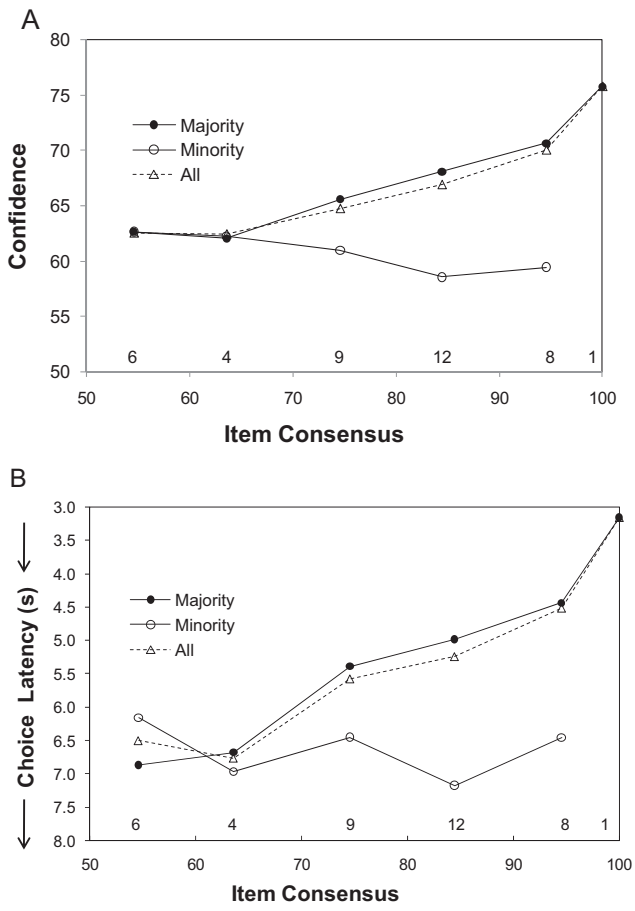


*Figure 5.* Panel A presents mean confidence for majority responses, for minority responses, and for all responses combined as a function of item consensus (the percentage of participants who chose the majority response). Indicated also is the number of items in each confidence category. Panel B presents the same data for choice latency (Experiment 1).

ity responses decreased with item consensus: The correlation between them across the 39 items was −.39, $p < .05$ (−.48, for the standardized scores, $p < .005$). Both of these trends are consistent with SCM.

***The relationship between choice latency and cross-person consensus.*** Similar analyses were conducted for choice latency. The pattern presented in Figure 5B largely mimicked that obtained for confidence. Latency decreased monotonically with item consensus: The correlation between mean latency and item consensus was −.78 across the 40 items, $p < .0001$. Response latencies were longer for minority responses (6.16 s) than for majority responses (5.49 s), $t(38) = 2.52$, $p < .05$. Choice latency for majority responses decreased with item consensus: Across the 39 items, the correlation was −.84 with item consensus, $p < .0001$. The respective correlation for the nonconsensual response was −.09, *ns*. The analyses just presented were also repeated after the choice latency scores were standardized. The results yielded essentially the same pattern as that obtained with the raw scores.

In sum, the pattern obtained for perceptual comparisons was basically the same as that observed for general-information questions (Koriat, 2010) and for social attitudes (Koriat & Adiv, 2010). Participants expressed stronger confidence when they chose the consensual response than when they chose the nonconsensual response. This pattern is in line with SCM if the participants are assumed to draw representations from a shared pool of representations for each item.

**The consensuality principle.** In the analyses presented so far, I have avoided the question of whether the answers chosen were correct or wrong and have focused only on the extent to which the answers were selected consistently within participants or across participants. We now turn to what have been the central questions about subjective confidence: To what extent do confidence judgments monitor the accuracy of the answer, and what are the reasons for the correspondence or miscorrespondence between confidence and accuracy? According to SCM, confidence in an answer should be correlated with the consensuality of the answer rather than its accuracy. This is because consensuality is diagnostic of self-consistency. Hence, the C/A correlation should be positive when the consistently favored answer is the correct answer but negative when the wrong answer is consistently favored.

***Confidence for correct and wrong answers.*** I carried out the following analyses using the data from Block 1 only. The percentage of correct choices ranged from 12.8% to 100% across items. All items eliciting more correct than wrong choices were classified as CC items, and those eliciting more wrong than correct choices were classified as CW. There were 32 CC items, with average percentage correct ranging from 59.0% to 100% ($M = 81.3\%$), and eight CW items, with average percentage correct ranging from 12.8% to 46.2% ($M = 26.0\%$).

To examine the C/A relationship, I averaged the confidence judgments for correct and wrong choices for each participant for the CC and CW items. Because eight participants gave only wrong answers to all CW items, the analyses were based only on the remaining 31 participants. Figure 6A presents the means of these judgments for these participants. The results clearly demonstrated a crossover interaction similar to what was found for general-information questions (Koriat, 2008a). A two-way analysis of variance (ANOVA), Item Class (CC vs. CW) × Correctness, yielded $F(1, 30) = 3.46$, mean square error ($MSE$) = 69.67, $p < .08$, for item class; $F < 1$, for correctness; and $F(1, 30) = 22.42$, $MSE = 77.63$, $p < .0001$, for the interaction. For the CC items, confidence was higher for correct than for wrong answers, $t(30) = 5.13$, $p < .0001$, whereas for the CW items, it was higher for the wrong answers, $t(30) = 2.68$, $p < .05$. The within-person C/A gamma correlation averaged .32 across the CC items, $t(30) = 5.89$, $p < .0001$, but −.25 across the CW items, $t(30) = 2.74$, $p < .01$. Thus, confidence is correlated with the consensuality of the choice rather than with its correctness.

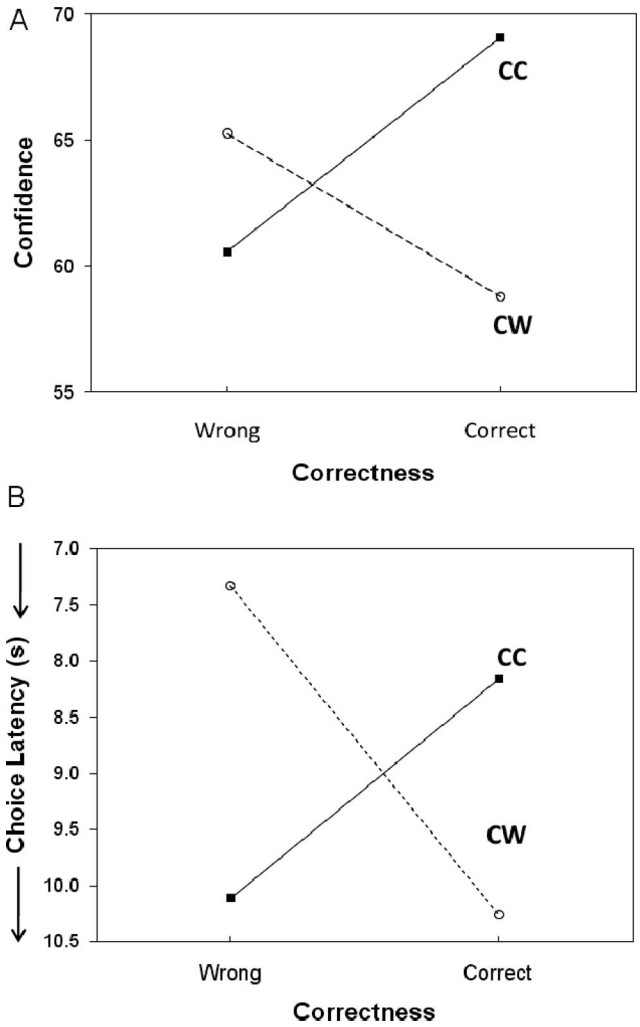


*Figure 6.* Panel A presents mean confidence for correct and wrong answers, plotted separately for the consensually correct (CC) and consensually wrong (CW) items. Panel B presents the same data for choice latency (Experiment 1).

***Choice latency for correct and wrong answers.*** Figure 6B presents the respective results for choice latency based on the same 31 participants. An Item Class (CC vs. CW) × Correctness (correct vs. wrong) ANOVA yielded a significant interaction, $F(1, 30) = 7.10$, $MSE = 25.99$, $p < .05$. The CC items yielded the typical pattern of faster choice latencies for correct than for wrong answers, $t(30) = 3.91$, $p < .001$, whereas the CW items yielded a

trend in the opposite direction, $t(30) = 1.86$, $p < .08$. Overall, the interactive pattern (Figure 6B) is very similar to that observed for confidence judgments (Figure 6A). The gamma correlation between choice latency and accuracy averaged −.28 for the CC items, $t(30) = 5.57$, $p < .0001$, and .25 for the CW items, $t(30) = 2.60$, $p < .05$.

***Choice latency as a potential cue for confidence.*** Whereas the previous section concerned cue validity—the validity of choice latency as a cue for accuracy, this section examines cue utilization—participants' reliance on latency as a basis for confidence. The choice latencies of each participant in Block 1 were split at the median of each class of items. Mean confidence judgments for below-median (short) and above-median (long) choice latencies are presented in Figure 7 as a function of the actual mean choice latency for short and long responses. For both classes of items, confidence decreased with increasing choice latency. A Choice Latency (short vs. long) × Class (CC vs. CW) ANOVA on confidence judgments yielded a significant effect for choice latency, $F(1, 38) = 9.19$, $MSE = 65.67$, $p < .005$. The difference in confidence between short and long choice latencies was significant for both the CC items, $t(38) = 2.50$, $p < .05$, and CW items, $t(38) = 2.51$, $p < .05$. These results suggest that choice latency influenced confidence in the same way for the CC and CW items under the heuristic that faster responses are more likely to be correct (see Koriat & Ackerman, 2010). However, this heuristic was valid only for the CC items, whereas for the CW items, it was counterdiagnostic (see also Koriat, 2008a).

On the whole, the results argue against a trace-access view according to which confidence is based on privileged access to perceptual or memory strength (see Koriat 2007; Van Zandt, 2000). Rather, confidence and response time are typically diagnostic of accuracy only because the responses to 2AFC general-knowledge or perceptual items are by and large correct.

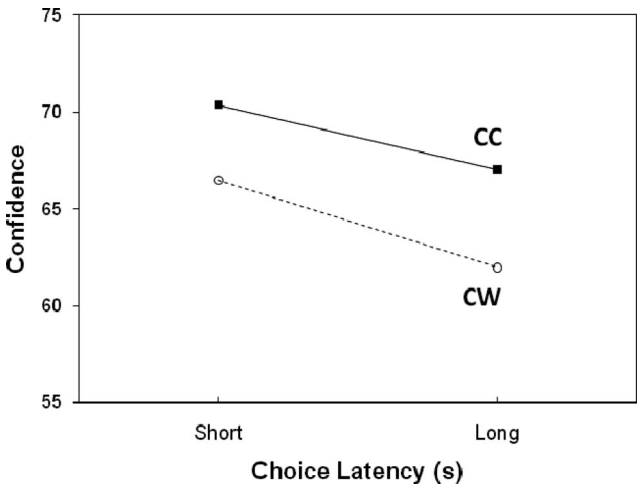


*Figure 7.* Mean confidence for answers with below-median (short) and above-median (long) choice latencies for consensually correct (CC) and consensually wrong (CW) items (Experiment 1).

**Consistency, consensus, and the consensuality principle.** Several issues concerning the relationship among intraindividual consistency, interindividual consensus, and the C/C relationship will be examined in this section.

***The reliability of interitem differences in choice and confidence.*** The assumption that the representations associated with an item are commonly shared implies that properties of items, notably the likelihood of choosing the majority response and confidence in that response, are reliable both within participants and across participants. Indeed, the Cronbach's alpha coefficient as a measure of interparticipant reliability (Crocker & Algina, 1986) was .94 for response choice and .73 for confidence judgments in Block 1, in line with the assumption of some consensus in the representations underlying choice and confidence.

According to SCM, cross-person consensus and within-person consistency should also be correlated. To examine this possibility, I divided participants' choices in Block 1 between those that were repeated two or more times in the subsequent blocks (frequent) and those that were repeated only once or never (rare). For each participant and for each item, I calculated the proportion of other participants (out of 38) who made the same choice in Block 1 as that made by him or her. This proportion was then averaged separately for frequent and rare responses. Across all participants, the proportion of other participants who made the same choice in Block 1 averaged .75 for frequent responses in comparison with .43 for rare responses, $t(38) = 13.4$, $p < .0001$. Thus, the choices that evidenced higher within-participant consistency were the more likely to be chosen by other participants in Block 1.

***The within-person consistency principle.*** The correlation between consistency and consensus suggests that a *consistency principle* analogous to the consensuality principle may hold true within participants. That is, assume that for each person, the items are divided between consistently correct and consistently wrong items according to the answer that is selected more often by the participant across the five blocks. Would the C/A relationship be positive for the consistently correct items and negative for the consistently wrong items? Such, in fact, would be expected to be the case according to SCM.

To examine this question, I divided the items for each participant into those for which the correct answer was chosen three times or more (consistently correct) and those for which the wrong answer was chosen three times or more (consistently wrong). Mean confidence for correct and wrong responses, plotted in Figure 8A, demonstrated a crossover interaction similar to what was observed for cross-person consensus. A two-way ANOVA for these means yielded $F(1, 38) = 57.12$, $MSE = 32.87$, $p < .0001$, for the interaction. For the consistently correct items, confidence was higher for the correct than for the wrong answers, $t(38) = 7.61$, $p < .0001$, whereas the opposite was true for the consistently wrong items, $t(38) = 4.48$, $p < .0001$.

The results just presented included items that contributed only to one of the two classes. When the analysis was repeated after these items had been eliminated for each participant, confidence was overall somewhat lower, but the crossover interaction was still clear (see Figure 8B). A two-way ANOVA, as before, yielded $F(1, 38) = 17.27$, $MSE = 23.83$, $p < .001$, for the interaction.

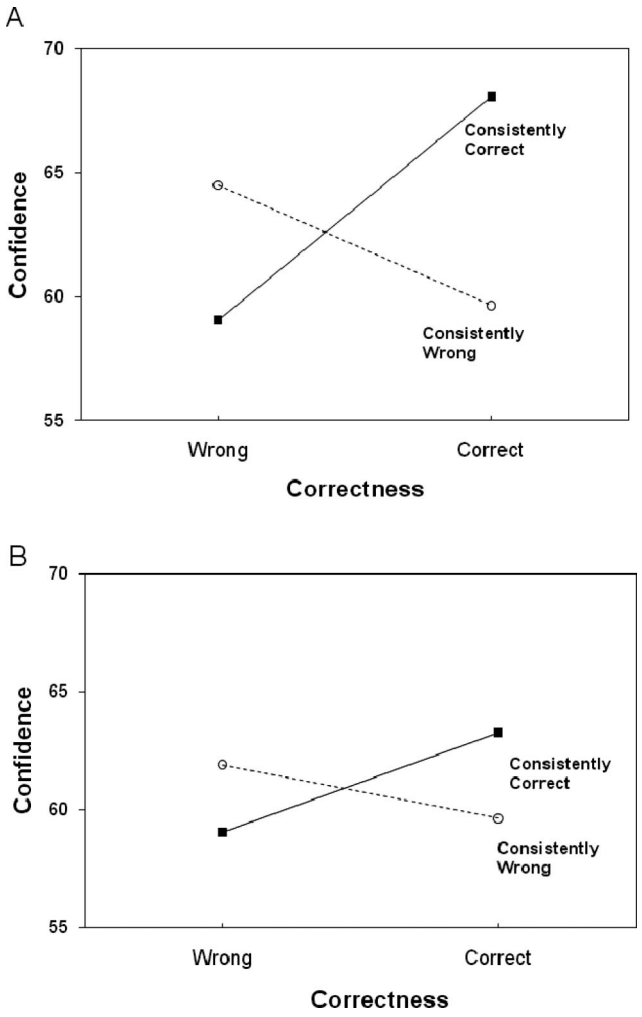


*Figure 8.* Panel A presents mean confidence for correct and wrong answers, plotted separately for the consistently correct and consistently wrong items. Panel B presents the same data after elimination of items that contributed only to one of the two classes (Experiment 1).

The consistency principle was also found for choice latency. For consistently correct items, choice latency averaged 5.18 s and 6.70 s, respectively, for correct and wrong answers. The respective means for consistently wrong items were 7.84 s and 5.40 s, respectively. A two-way ANOVA yielded a significant interaction, $F(1, 37) = 26.23$, $MSE = 5.67$, $p < .0001$.

In sum, as predicted by SCM, in a repeated-testing paradigm, confidence in the choice was correlated with the consistency with which that choice is made across repetitions rather than with the correctness of the choice. The same is true for response speed.

***The joint effects of within-person consistency and cross-person consensus.*** Given that confidence increased with both within-person consistency and between-person consensus, it is of interest to examine the joint contribution of the two variables to confidence judgments. For each participant, the response to an item in Block 1 was classified as (a) consensual or nonconsensual on the basis of the responses of all participants in Block 1 and as (b) frequent or rare, depending on its within-participant frequency across all five blocks. Figure 9 presents mean confidence in the response in Block 1 as a function of the consensuality of the response and its within-person frequency (based on 32 participants who had all four means). It can be seen that the effects of the two factors are additive. A Consensus × Consistency ANOVA yielded $F(1, 31) = 12.75$, $MSE = 110.72$, $p < .005$, for consensus; $F(1, 31) = 13.91$, $MSE = 103.13$, $p < .001$, for consistency; and $F < 1$ for the interaction. Confidence increased with both factors, and the overall effect of the two factors was about the same: Partial $\eta^2$, as an estimate of effect size, was .31 for response consistency and .29 for response consensus. Note that response consistency was found to have a markedly stronger effect on confidence in one's social attitudes than was response consensus (Koriat & Adiv, 2010). Presumably, in the case of perceptual judgments, cross-person consensus and within-person consistency are equally diagnostic of the self-consistency underlying choice and confidence.

In sum, the results of Experiment 1 yielded a large number of observations that accord with SCM. These observations support the predictions of SCM regarding both the basis of confidence judgments and the reasons for their general accuracy. The results are very similar to those observed for general-information questions (Koriat, 2010).

## Experiment 2

In Experiment 2, I had two aims. The first was to generalize the results of Experiment 1 to a task that required comparison of the areas of two geometric shapes. The second was to test predictions of SCM regarding the calibration of confidence judgments. To test these predictions, I had to obtain confidence judgments in the form of assessed probability (50–100%) as is commonly used in calibration studies.

According to SCM, confidence judgments are construed subjectively as pertaining to validity, but they are actually based on cues about reliability. Reliance on reliability is liable to lead to overconfidence because reliability is virtually always higher than validity. For example, when an answer to a general-information question is supported consistently across several considerations, this does not guarantee that the answer is correct. Similarly, the

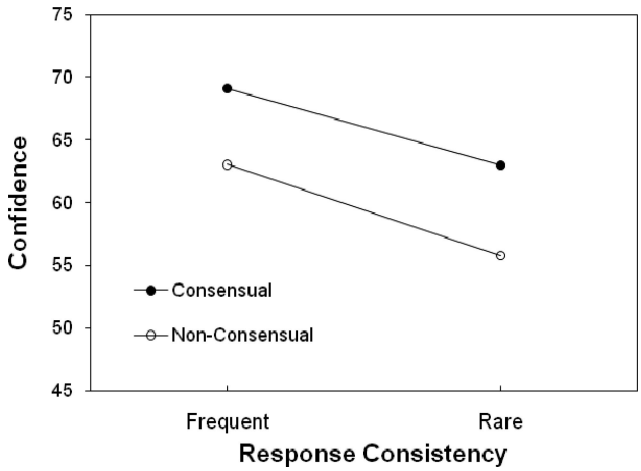


*Figure 9.* Mean confidence for frequent and rare responses plotted separately for consensual and nonconsensual responses (Experiment 1).

fact that in the Müller–Lyer illusion, one line is consistently perceived as being longer than the other line does not prove that it is indeed longer.

I expected the results of Experiment 2 to yield an overconfidence bias because of the deliberate inclusion of items that are likely to yield CW choices. However, this calibration was compared with that observed for a matched set of general-knowledge questions taken from Koriat (2008a). In addition, two questions were addressed. First, is the overconfidence bias for the perceptual task reduced or eliminated when it is evaluated against indexes of reliability or self-consistency rather than correctness? Second, are there systematic differences in calibration between the perceptual and knowledge tasks when calibration is evaluated against reliability or consistency? If such was found to be the case, it could suggest that the two tasks differ in the extent to which confidence is based on self-consistency.

## Method

**Participants.** Forty-one psychology undergraduates (34 women and seven men) participated in the experiment for pay.

**Stimulus materials.** The experimental materials consisted of 40 geometric shapes. These were paired to form 40 pairs so that each geometric shape appeared twice but each time a particular shape appeared, it was paired with a different shape (see Figure 2 for examples). Ten additional shapes were used to create the practice trials. Each of the stimuli subtended a visual angle of approximately 5.8°. As in Experiment 1, the pairing of the stimuli was based on the results of an exploratory study and was aimed to yield a sufficiently large number of CW pairs.

**Apparatus and procedure.** The apparatus was the same as in Experiment 1. The procedure was also the same, with the exception that participants reported their confidence in the form of assessed probability in the range 50–100%.

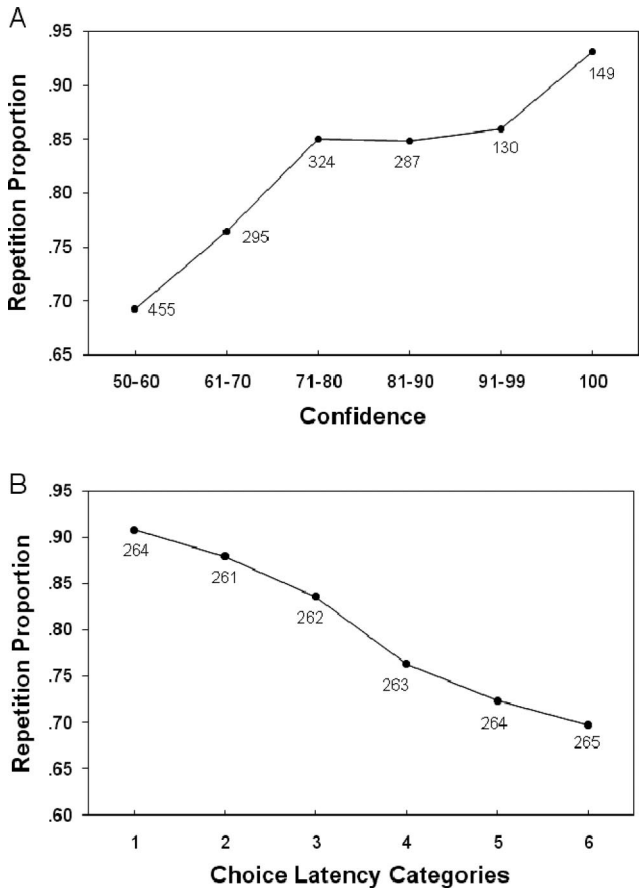


*Figure 10.* Panel A presents the likelihood of repeating the Block-1 choice in the subsequent four blocks (repetition proportion) as a function of confidence in these responses in Block 1. Panel B plots repetition proportion as a function of Block-1 choice latency (Experiment 2).

## Results and Discussion

As in Experiment 1, participants tended to make the same choice across blocks: The likelihood of repeating the Block-1 choice over the next four blocks averaged 79.89% across participants. The results for both confidence and choice latency closely replicated the patterns obtained in Experiment 1. Therefore, I will present them briefly, focusing on the results for confidence.

**Reproducibility.** As in Experiment 1, repetition proportion increased with confidence in Block 1. The Spearman rank-order correlation over the six values in Figure 10A was .94, $p < .005$. A similar result was observed for response speed (Figure 10B) after eliminating outliers (3.4%) as in Experiment 1: The rank-order correlation across the six points was $-1.0$, $p < .0001$.

**The relationship between confidence and within-person response consistency.** A comparison of confidence for the participant's frequent and rare responses yielded the pattern depicted in Figure 11A and based on 39 participants who had all means. Focusing on the partial-consistency items, participants were more confident when they chose their more frequent response (73.65%) than when they chose their less frequent response (66.99%), $t(38) = 9.76$, $p < .0001$. Confidence in the frequent response increased with item consistency (3 vs. 4), $t(38) = 2.86$, $p < .01$, whereas confidence in the less frequent response did not vary with item consistency, $t(38) = 0.17$. Choice latency was shorter for the frequent responses (4.18 s) than for the rare responses (6.44 s), $t(38) = 5.72$, $p < .0001$. It decreased with item consistency for the frequent choices but increased with item consistency for the rare choices (see Figure 11B).

As in Experiment 1, there was little evidence for increased confidence across presentations. Confidence averaged 74.05, 72.08, 73.0, 72.96 and 73.31%, respectively, for Blocks 1–5. Confidence in Block 1 averaged 75.07% for choices that were repeated two times or more and 68.14% for those that were repeated 1 time or never, $t(40) = 6.73$, $p < .0001$. Thus, even for Block 1, participants' confidence discriminated between the more frequent and the less frequent responses.

**The relationship between confidence and cross-person response consensus.** For two items, all participants gave the same response. For the remaining 38 items, the answer that was chosen by the majority of participants across the five blocks was defined as the consensual or majority answer. Figure 12A presents mean confidence for each of the six item-consensus categories. Confidence was higher for the majority answer (73.62%) than for the minority answer (68.18%), $t(37) = 6.66$, $p < .0001$, and for majority responses, it increased monotonically with item consen-

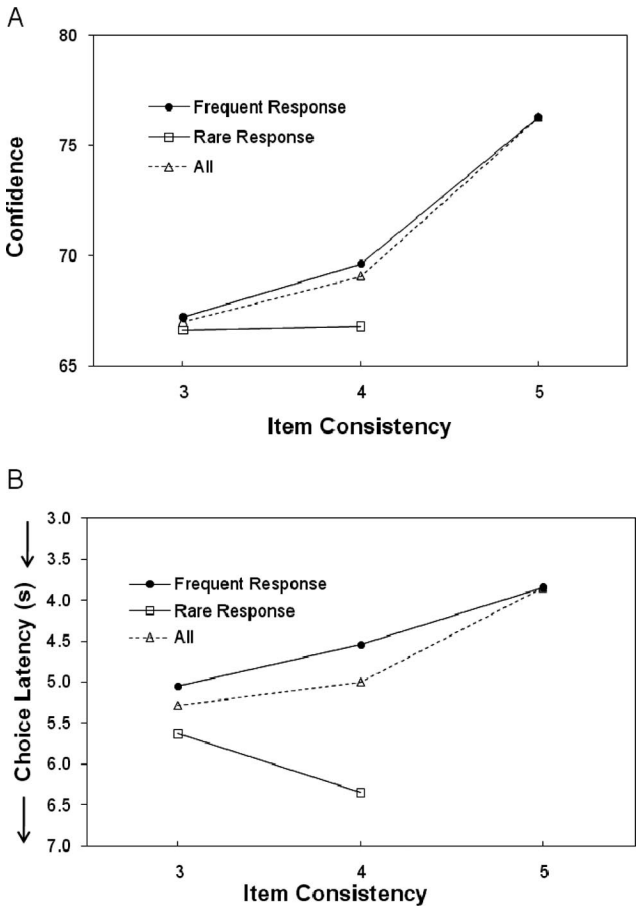


*Figure 11.* Panel A presents mean confidence for each participant's frequent and rare responses as a function of item consistency (the number of times that a response was made across all five blocks). Panel B presents mean choice latency as a function of item consistency (Experiment 2).

sus: The correlation was .71 across the 38 items, $p < .0001$. The respective correlation for minority responses was .23, *ns*. As in Experiment 1, individual differences in confidence were quite reliable: The cross-participant correlations between the participants' mean confidence for different blocks averaged .90. When the results were standardized to remove the contribution of reliable individual differences, the pattern of the means was essentially the same as that depicted in Figure 12A.

Choice latencies were shorter for majority responses (4.13 s) than for minority responses (5.36 s), $t(37) = 5.78$, $p < .0001$ (see Figure 12B). Choice latency for majority responses decreased with item consensus: The correlation across the 38 items was $-.71$, $p < .0001$. The respective correlation for the minority response was positive (.17) but not significant.

**The consensuality principle.** In Block 1, two additional items elicited the same (correct) response across all participants, and they were eliminated from the analysis of confidence resolution. For the 36 items, 21 were classified as CC: Their mean percentage correct ranged from 58.5% to 97.6% ($M = 80.0\%$). The remaining 15 items were classified as CW: Their mean percentage correct ranged from 4.9% to 46.3% ($M = 24.7\%$; see Figure 2 for examples of the stimulus pairs in each category).

Because two participants gave only correct answers to all CC items and two other participants gave only wrong answers to all CW items, the analyses were based only on the remaining 37 participants. The results (see Figure 13A) demonstrate a crossover interaction. A two-way ANOVA yielded $F(1, 36) = 11.66$, $MSE = 25.20$, $p < .005$, for item class; $F < 1$, for correctness; and $F(1, 36) = 58.92$, $MSE = 24.74$, $p < .0001$, for the interaction. For the CC items, confidence was higher for correct answers than for wrong answers, $t(36) = 4.70$, $p < .0001$, whereas for the CW items, confidence was higher for the wrong answers, $t(36) = 5.03$, $p < .0001$. For the 37 participants, the C/A gamma correlation was positive (.30) across the CC items, $t(36) = 5.97$, $p < .0001$, but negative ($-.34$) across the CW items, $t(36) = 5.39$, $p < .0001$. Thus, confidence is correlated with the consensuality of the response rather than with its correctness, similar to what was found in Experiment 1.

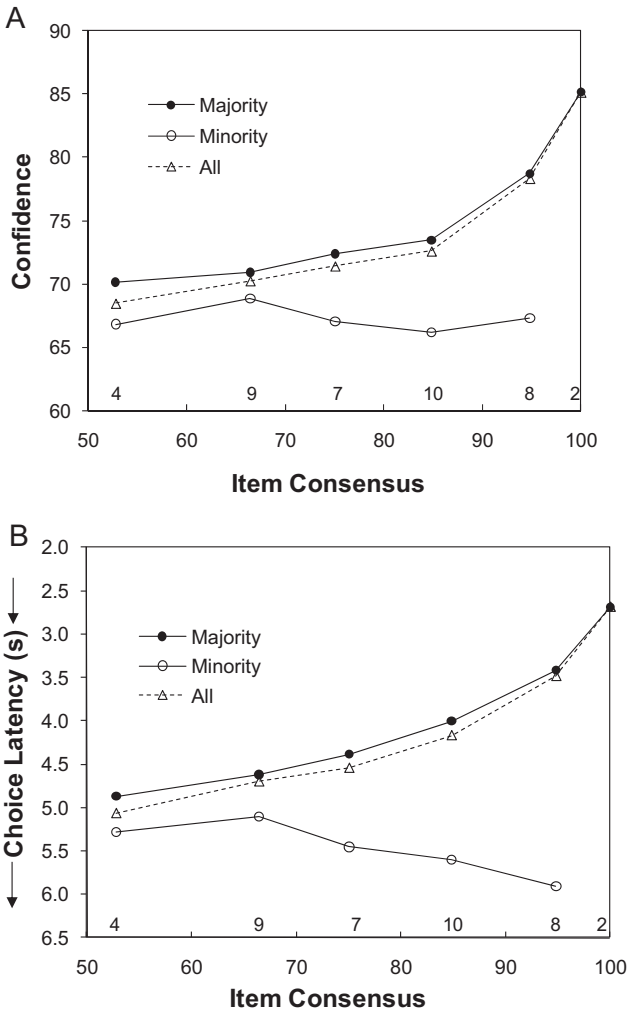


*Figure 12.* Panel A presents mean confidence for majority responses, for minority responses, and for all responses combined as a function of item consensus (the percentage of participants who chose the majority response). Indicated also is the number of items in each confidence category. Panel B presents the same data for choice latency (Experiment 2).

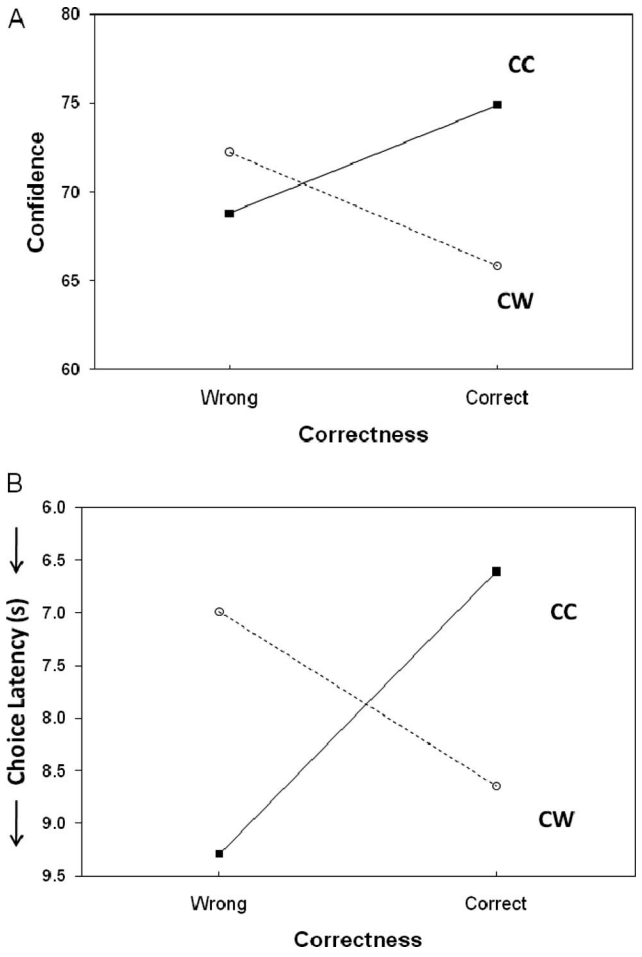


*Figure 13.* Panel A presents mean confidence for correct and wrong answers, plotted separately for the consensually correct (CC) and consensually wrong (CW) items. Panel B presents the same data for choice latency (Experiment 2).

A similar crossover interaction was also evident in the results for choice latency (Figure 13B). An Item Class × Correctness ANOVA yielded $F(1, 36) = 20.39$, $MSE = 8.53$, $p < .0001$, for the interaction. For the CC items, choice latencies were faster for correct than for wrong answers, $t(36) = 4.35$, $p < .0001$, whereas for the CW items the opposite pattern was observed, $t(36) = 3.90$, $p < .0005$. The within-person latency–accuracy gamma correlation, averaged −.39 for the CC items, $t(36) = 7.66$, $p < .0001$, and .23 for the CW items, $t(36) = 3.85$, $p < .001$. With regard to the latency–confidence relationship, confidence decreased with increasing choice latency for both CC and CW items, suggesting that participants relied on the same heuristic indiscriminately.

**Consistency, consensus, and the consensuality principle.** As in Experiment 1, the interparticipant reliability of choice and confidence in Block 1 was high: The Cronbach's alpha coefficient amounted to .97 for response choice and .91 for confidence judgments. There was also a correlation between within-person consistency and between-person consensus: When the responses of each participant were classified as frequent or rare for each item, the proportion of other participants (out of 40) who made the same response in Block 1 averaged .75 for frequent responses in comparison with .49 for rare responses, $t(40) = 10.65$, $p < .0001$.

As in Experiment 1, the C/A relationship was positive for consistently correct items and negative for consistently wrong items. One participant gave only wrong answers to all consistently wrong items. The results for the remaining 40 participants are presented in Figure 14A. For the consistently correct items, confidence was higher for correct answers (74.87%) than for wrong answers (66.25%), $t(39) = 10.25$, $p < .0001$, whereas for the consistently wrong items, confidence was higher for the wrong answers (71.61%) than for correct answers (65.84%), $t(39) = 6.80$, $p < .0001$. I repeated the analyses after eliminating the items for which the participant made the same response throughout the five blocks; the same crossover pattern was observed (Figure 14B): The two-way ANOVA yielded $F(1, 39) = 15.11$, $MSE = 13.07$, $p < .0005$, for the interaction. A crossover interaction was also observed for choice latency between item class and correctness; the two-way ANOVA yielded $F(1, 39) = 12.86$, $MSE = 6.31$, $p < .001$, for the interaction.

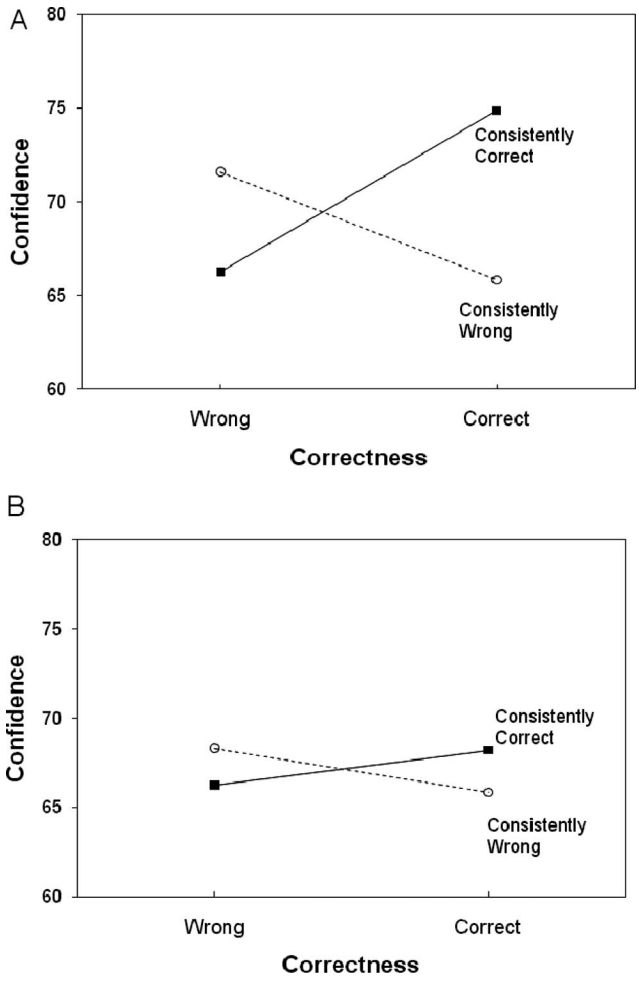


*Figure 14.* Panel A presents mean confidence for correct and wrong answers, plotted separately for the consistently correct and consistently wrong items. Panel B presents the same data after elimination of items that contributed to only one of the two classes (Experiment 2).

**The calibration of confidence judgments.** We turn finally to calibration. Because CW items were deliberately included, I could not conduct an assessment of the over/underconfidence bias. However, it is important to examine what happens to calibration when it is evaluated against some index of self-consistency.

In the following analyses, I compared calibration curves for the perceptual task of Experiment 2 with those obtained for general-information questions in Koriat (2008a). To do so, I used only the results from Block 1 in both experiments. Four perceptual items for which the percentage of correct answers was 100% were eliminated from the analyses. For the remaining items, 36 general-information questions were selected from the 105 questions of Koriat (2008a) to match each of the perceptual items in terms of the percentage of consensual answers. Mean percentage correct for each item averaged 56.98% across items for the perceptual set and 61.59% for the knowledge set. I then assessed calibration for each of the two sets of items using the procedure of calibration studies (see Lichtenstein et al., 1982). The results (accuracy), plotted for seven confidence categories (50, 51–60, . . . 91–99, 100), appear in Figure 15A for the perceptual set and in Figure 15B for the knowledge set. These results disclose a strong overconfidence bias for both sets. For the perceptual set, mean confidence for each participant averaged 72.84% across participants, in comparison with 56.98% for percentage correct. For the knowledge set, the respective means were 73.99% across participants and 61.59% for percentage correct. Thus, the overconfidence bias amounted to 15.86 percentage points for the perceptual set and to 12.40 percentage points for the knowledge set.

The same data were plotted in the same figures except that calibration was evaluated against three different indexes of self-consistency. In the first analysis, confidence judgments were compared with item consensus: For each confidence category, the percentage of consensual (majority) responses across all items in that category was calculated. Mean item consensus (averaged for each participant and then averaged across participants) was 78.05% for the perceptual set and 77.37% for the knowledge set. Thus, for both sets, confidence yielded a small and comparable underconfidence bias, which amounted to 5.21 percentage points for the perceptual set, $t(40) = 3.67$, $p < .001$, and to 3.38 percentage points for the knowledge set, $t(40) = 3.37$, $p < .005$.

In the second analysis, confidence judgments were compared with response consensus: For each target participant and for each item, the percentage of other participants who gave the same response as the target participant to that item was calculated across all items in that confidence category. Mean response consensus across participants was 68.25% for the perceptual set and 67.21% for the knowledge set. Thus, for both sets, confidence yielded a small overconfidence bias, which amounted to 4.59 percentage points for the perceptual set, $t(40) = 3.17$, $p < .005$, and 6.78 percentage points for the knowledge set, $t(40) = 6.30$, $p < .0001$.

A third criterion with which confidence judgments can be compared is repetition, the likelihood of making the same choice in a subsequent presentation of the item. Because the study of Koriat (2008a) included only two presentations of the questions, the evaluation of calibration against repetition was based only on the first two blocks for the perceptual task. The likelihood of making the Block-1 choice in Block 2 (repetition) was calculated for each category for both sets of stimuli. Mean repetition scores averaged 79.89% for the perceptual set and 86.79% for the knowledge set.

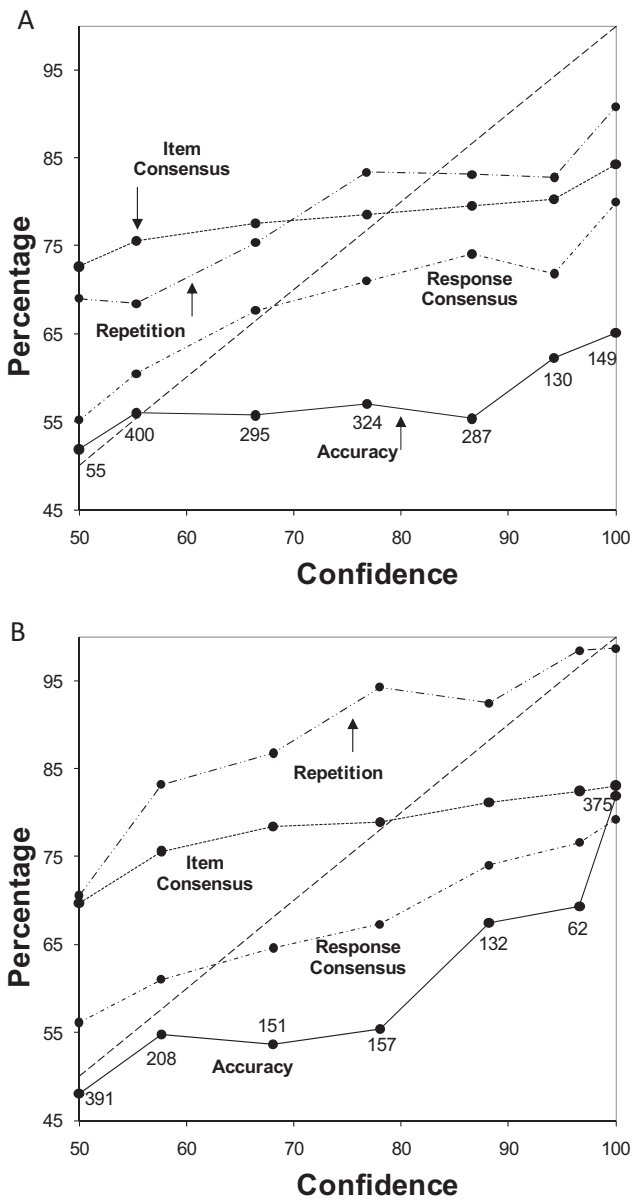


*Figure 15.* Calibration curves derived from four different assessment criteria for the appropriateness of confidence judgment: Accuracy (the likelihood that a participant would choose the correct answer), item consensus (the percentage of consensual responses), response consensus (the percentage of other participants who gave the same response as each target participant), and repetition (the percentage of times that the Block-1 response was repeated in Block 2). The results are plotted for two sets of matched items: perceptual-comparison items (Panel A, items drawn from Experiment 2), and general-information questions (Panel B, questions drawn from Koriat, 2008a).

Thus, the two tasks seemingly can be distinguished by this criterion. For both tasks, confidence yielded an underconfidence bias, but this bias was smaller for the perceptual set (amounting to 7.05 percentage points) than for the knowledge set (12.80 percentage points). A two-way ANOVA, Task (perceptual vs. knowledge) × Measure (confidence vs. repetition) yielded $F(1, 80) = 9.15$,

$MSE = 72.42$, $p < .005$, for task; $F(1, 80) = 78.09$, $MSE = 51.74$, $p < .0001$, for measure; and $F(1, 80) = 6.55$, $MSE = 51.74$, $p < .05$, for the interaction.

In sum, the two matched sets of items yielded a marked overconfidence bias of about the same magnitude. The evaluation of confidence judgments against any of the three indexes of self-consistency yielded a markedly smaller tendency toward overconfidence than when these judgments were evaluated against accuracy. This observation accords with the idea that the overconfidence bias stems, at least in part, from the discrepancy between reliability and validity. Only for response repetition was there a measurable difference between perceptual and general-knowledge tasks. Perhaps, participants can better recall their previous answer to a general-information question than to a perceptual-comparison item and tend simply to repeat it. On the whole, then, the calibration results presented in Figure 15 do not give any clear indication of a qualitative difference in confidence between the perceptual task and the knowledge task in the calibration of confidence judgments.

To conclude, the results of Experiment 2 replicated the main findings of Experiment 1. These results were consistent by and large with SCM. They also replicated the consensuality pattern that had been found in Experiment 1. In addition, the results on calibration supported the idea that the overconfidence bias that is sometimes observed derives in part from the discrepancy between reliability and validity and can be reduced or eliminated when confidence judgments are evaluated against some criterion of reliability or self-consistency.

## Experiment 3

Experimental evidence suggests that confidence judgments guide behavioral decisions (Fischhoff, Slovic, & Lichtenstein, 1977; Gill, Swann, & Silvera, 1998; Goldsmith & Koriat, 2008). In Experiment 3, I examined whether the consensuality results have behavioral consequences. Participants were presented with items taken from Experiment 2. They chose the correct answer and were then asked to wager on the correctness of their answer (see Persaud, McLeod, & Cowey, 2007). We examined whether participants would fail to maximize their cash earnings in the case of CW items by betting heavily on the wrong choices.

### Method

**Participants.** Twenty-six psychology undergraduates (21 women and 5 men) participated in the experiment for course credit.

**Stimulus materials.** All 16 CW pairs from Experiment 2 were used (26.8% correct responses across the five blocks in Experiment 2), and 16 CC items were selected to match roughly these items in terms of the percentage of consensual choices (75.3% correct responses).

**Apparatus and procedure.** The apparatus was the same as in the previous experiment. Participants were told that they had a chance to earn money if they would agree to take part in a gambling game. They would have to judge which of two geometric figures had a larger area and then would have to decide for each pair, how much money—between 0 and 10 Israeli shekels (approximately \$0.25–\$2.50 in U.S. dollars)—they would be willing to wager on the correctness of their answer. If the answer was correct, the amount wagered would be added to their earnings, but if it was incorrect, that amount would be deducted. They were informed that their earnings would be based only on their performance and wagers for 10 of the 32 items that would be selected randomly. However, they were assured that they would not pay any losses if there will be any.

The procedure was similar to that of Experiment 2 but instead of making confidence judgments, participants typed in the amount wagered, and then clicked a confirm box, after which the next trial began. At the end of the experiment, participants entered a 3-digit number; on the basis of the number entered, a random set of 10 pairs was selected. The number of correct and wrong answers among the 10 pairs was displayed as well as the amount of cash earned. There was only one block. The order of the experimental pairs was determined randomly for each participant and for each presentation.

### Results and Discussion

I first will examine the choices made. For the CC items, participants chose the correct answer more often (75.72%) than the wrong answer, whereas for the CW items they chose the wrong answer more often (61.30%) than the correct answer, $t(30) = 5.85$, $p < .0001$. For the CC items, choice latency (after outliers were eliminated, as in the previous experiments) was shorter for the correct answers (11.31 s) than for the wrong answers (14.56 s), $t(25) = 2.57$, $p < .05$, whereas for the CW items, it was shorter for the wrong answers (12.90 s) than for the correct answers (15.34 s), $t(25) = 4.46$, $p < .05$.

For the wagering responses, Figure 16 presents the average amount wagered on correct and wrong choices for CC and CW items. A two-way ANOVA yielded $F(1, 25) = 3.37$, $MSE = 0.55$, $p < .08$, for item class; $F < 1$ for correctness; and $F(1, 25) = 14.05$, $MSE = 0.91$, $p < .001$, for the interaction. For the CC items, participants placed larger wagers on correct answers than on

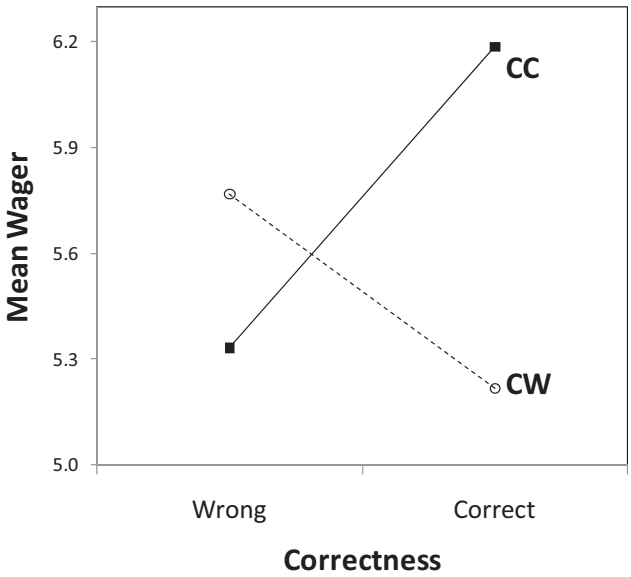


*Figure 16.* Mean wagers placed on correct and wrong choices for consensually correct items (CC) and for consensually wrong items (CW).

wrong answers, $t(25) = 3.04$, $p < .01$, whereas for CW items, they placed larger wagers on the wrong answers, $t(25) = 2.61$, $p < .05$. It is interesting to note that this pattern was observed in an item-based analysis as well. For the 16 CC items, 10 items exhibited a trend of larger wagers placed on the correct answer than on the wrong answer (one tie), whereas for the 16 CW items, 11 items indicated larger wagers on the wrong answer than on the correct answer, $\chi^2(1) = 3.14$, $p < .10$. This trend suggests that the interaction occurs even in a between-person comparison.

The percentage of correct responses in Experiment 3 correlated .89, $p < .0001$, across the 32 items with the respective percentage in Block 1 of Experiments 2. Mean confidence in Experiment 2 (Block 1) predicted the amount wagered in Experiment 3: The correlation across the 32 items was .60, $p < .001$, consistent with the idea that participants rely on their confidence in their strategic regulation of their behavior (Koriat & Goldsmith, 1996).

In sum, the consensuality results appear to have behavioral consequences: Participants' reliance on confidence as a basis for wagering is beneficial only for CC items but is detrimental for CW items.

## General Discussion

This study addressed the two cardinal questions about subjective confidence: the basis of confidence judgments and the reasons for their accuracy. SCM was originally developed to offer an answer to these questions with regard to confidence in general knowledge. The present study provided consistent support for the extension of SCM to confidence in perceptual-comparison tasks. Not only were the results very similar for the two perceptual tasks used, but they were also quite similar to those obtained for general-information questions (Koriat, 2010). In the following sections, I first summarize the findings and then discuss some of their general implications.

### Summary of the Findings

**Confidence monitors reproducibility.** Confidence in the Block-1 choice predicted the likelihood of making that choice in the subsequent presentations of the item, consistent with the proposition that confidence acts as a monitor to reproducibility. The likelihood of repeating the Block-1 choice also decreased monotonically with response latency in Block 1 (Figures 3 and 10).

**Confidence and response consistency.** The predictions of SCM concerning response consistency were clearly confirmed (Figures 4 and 11): (a) Mean confidence increased with increasing within-person consistency; (b) the more frequent responses were associated with higher confidence than were the less frequent responses; (c) confidence in the frequent responses increased with item consistency; and (d) confidence in the rare responses either decreased or did not vary with item consistency. Precisely the same overall pattern was observed for response speed. The difference between frequent and rare choices was observed even in Block 1, when the frequent–rare status of a choice was postdicted from the choices made in subsequent blocks.

**Confidence and response consensus.** The pattern obtained for response consistency was replicated by the results for response consensus when the choices were classified on the basis of cross-person agreement. In particular, majority choices were associated with higher confidence and faster response times than minority choices (Figures 5 and 12).

**Resolution.** The results clearly supported the consensuality principle: Confidence and response speed were correlated with the consensuality of the choice rather than with its correctness. The C/A relationship was positive when the consensual choice was the correct choice but negative when it was the wrong choice. Experiment 3 indicated that the consensuality results can have behavioral consequences. For the CC items, participants placed larger wagers on correct than on wrong answers, whereas the opposite was found for the CW items.

**The consistency principle.** The results also supported the consistency principle, which is analogous to the consensuality principle: The C/A relationship was positive for items in which a participant's modal choice was the correct choice but negative for items in which the modal choice was the wrong choice. These results were also mimicked by the results for response speed.

**The relationship between consistency and consensus.** 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 chosen by other participants in Block 1. This pattern of results supports the idea of a common item–specific population of representations from which different participants sample representations in each encounter with the item.

**The calibration of confidence judgments.** The strong overconfidence bias that was observed in Experiment 2 was reduced or eliminated when confidence was evaluated against indexes of self-consistency. This was true for both the perceptual items and their matched general-information questions, supporting the proposition that the overconfidence bias derives in part from participants' reliance on reliability as a cue for validity.

I turn next to examination of three issues: the bases of confidence judgments, the C/A relationship, and the calibration of confidence judgments. I will conclude by discussing the question whether SCM for confidence in 2AFC tasks can generalize across perceptual and general-knowledge items.

### The Bases of Confidence Judgments

The unique features of SCM can be brought to the fore by reference to the distinction between the Brunswikian and Thurstonian models of uncertainty. According to Juslin and Olsson (1997; see also Dougherty, 2001; Vickers & Pietsch, 2001), sensory tasks are dominated by Thurstonian uncertainty. For these tasks, uncertainty stems from random neural noise that may occasionally result in incorrect responses. Because variability is completely random, little within-person and cross-person consistency is expected in the tendency to make incorrect responses to a stimulus, and the percentage correct should generally exceed 50%. These tasks are expected to yield an underconfidence bias. General-information questions, in contrast, are said to be dominated by Brunswikian uncertainty. Here errors stem from the imperfect validity of the cues that people use to infer the correct answer. Because participants generally rely on exactly the same cue-based inference, errors should be correlated across persons, and the percentage correct may vary across the full range from 0% to 100%. General-knowledge tasks may yield an overconfidence bias.

SCM can be said to combine features from both the Brunswikian and Thurstonian approaches to confidence. With the Brunswikian approach, it shares the assumption that in responding to an item, participants sample representations from a population of representations that is largely shared. This was assumed to be true not only for general-knowledge tasks but also for perceptual-comparison tasks such as those used in the present study (and also for other tasks such as word matching and social beliefs and attitudes). The assumption of a common item–specific population of representations is supported by the within-person consistency and the cross-person consensus in the choice made for different items. Furthermore, the errors were correlated across participants to the extent that for some of the items, mean percentage correct was less than 50%. There was little difference in these respects between the results for the perceptual tasks used in this study and the general-information questions examined by Koriat (2010), and both tasks were liable to yield overconfidence.

At the same time, SCM also incorporates the Thurstonian notion of a random sampling of representations not only for perceptual tasks but also for general-knowledge tasks. Random sampling was expected to yield occasional deviations from the consensual choice. Indeed, the results clearly testified for a certain degree of within-person and between-person variability in the choice made for the same item. What is important is that confidence was shown to track not only the stable contributions to choice, as reflected in the functions relating mean confidence to item consistency and item consensus, but also the variable contributions: Minority choices were associated consistently with lower confidence than majority choices. This was true for both general knowledge and perceptual comparisons (and also for social attitudes, Koriat & Adiv, 2010).

With regard to general knowledge, the assumption in the PMM theory is that the response to a 2AFC item is determined by a single cue that discriminates between the two answers, and confidence simply reflects the perceived validity of that cue. Thus, the notion of a random sampling of a collection of representations does not apply. In contrast, many models of confidence in sensory, psychophysical tasks incorporate the notion of random fluctuations that are due to internal noise (e.g., Audley, 1960; Merkle & Van Zandt, 2006; Vickers, 1979; for a review, see Baranski & Petrusic, 1998). Although some researchers acknowledged that for certain perceptual tasks, the major source of variability is external to the observer (see Olsson, 1999; Vickers & Pietsch, 2001), none has postulated or tested possible relations between confidence, on the one hand, and within-person and between-person consistency, on the other hand.

SCM in its present form includes very rudimentary assumptions, only those that were necessary to bring to the fore the gross regularities that were observed across several tasks. However, the model is rather crude and might benefit from the incorporation of some of the additional processing assumptions (e.g., response bias, thresholds, interval of uncertainty, sampling window, deadline, multiple traces, and so forth) included in other models (see e.g., Dougherty, 2001; Vickers & Pietsch, 2001). It should be stressed, however, that some of the assumptions are domain specific (e.g., pertaining either to memory or to perception e.g., Dougherty, 2001; Juslin & Olsson, 1997), whereas SCM focuses on mnemonic determinants of confidence that are relatively content free and thus can account for the generality of the model across several different domains.

With regard to choice latency, the results indicated that differences in response latency closely mimic those expected and obtained for self-consistency. It was proposed that both self-consistency and choice latency represent mnemonic cues derived from the experience gained in the decision process, primarily the amount of deliberation and conflict involved. The effects of self-consistency and choice latency on confidence are in line with the proposition that metacognitive judgments are based primarily on the feedback from task performance: It is by attempting to answer a question or to solve a problem that one knows whether the answer or the solution is correct (Hertwig, Herzog, Schooler, & Reimer, 2008; Jacoby, Kelley, & Dywan, 1989; Koriat et al., 2006).

It should be stressed that response latency is not affected always by the same variables that affect confidence (see Wright & Ayton, 1988). In fact, Koriat et al. (2006) reported evidence indicating that confidence judgments decrease with choice time when choice time is data driven but increase with choice time when choice time is goal driven. Possibly, mnemonic cues are relied upon according to their cue validity. Indeed, the results of Koriat and Ma'ayan (2005) suggest that metacognitive judgments are based on a flexible and adaptive use of different mnemonic cues according to their relative validity in predicting performance. In the present context, it would seem that across a homogenous set of items, the amount of time spent choosing an answer is diagnostic of the degree of self-consistency and can serve generally as a valuable, frugal cue for accuracy.

## The Resolution of Confidence Judgments

Turning next to the accuracy of confidence judgments, the results replicated closely the consensuality pattern that had been found for general-knowledge questions (Koriat, 2008a), suggesting that even for perceptual tasks, participants do not have direct access to the accuracy of their answers. In addition, the present study added an important observation to those reported by Koriat (2008a): A pattern that is analogous to the consensuality results was observed in a within-person analysis. In both Experiments 1 and 2, confidence in a choice was correlated with the frequency of that choice across the five blocks. regardless of its accuracy.

It should be noted that a pattern of results that accords with the consensuality principle appears in the data presented by Baranski and Petrusic (1995, Figure 2) for a perceptual task that required location comparisons. For that task, certain stimulus configurations had been found to yield illusory errors. The calibration curves for these *misleading/illusory items* (for which the percentage correct was 44% or less) indicated that rate of correct responses decreased with mean confidence. The authors concluded, "This reflects an ability to differentiate correct from incorrect judgments but, as is evident in Figure 2, the skill is actually counter to reality; i.e., error probability increases as confidence in a correct response increases!" (p. 404).

One implication of these results is that for a randomly selected set of perceptual stimuli, participants are able, by and large, to discriminate between correct and wrong answers. A second implication, however, is that this ability does not stem from participants' privileged access to the correctness of their answers. Rather,

participants rely on a metacognitive heuristic that is generally valid because self-consistency is diagnostic of accuracy. Thus, my results as well as those of Baranski and Petrusic (1995) illustrate the idea that the accuracy of metaknowledge is a byproduct of the accuracy of knowledge itself (Koriat, 1993, 2010).

The methodological implication of these results is that ascertaining a representative design in which items are sampled randomly is critical for drawing descriptive conclusions about people's metacognitive skills. However, using a biased selection of items is necessary for the clarification of the processes underlying subjective confidence and its accuracy (Koriat, Pansky, & Goldsmith, 2010). Indeed, this study, as well as that of Koriat (2008a), reinforces the plea for the use of a representative sampling of items in assessments of the accuracy of confidence judgments (Gigerenzer et al., 1991; Juslin, 1994; see Dhami, Hertwig, & Hoffrage, 2004). This plea, however, has been stressed in connection with the assessment of calibration. The results of this study indicate that resolution too varies greatly depending on the characteristics of the items included in the studied sample.

## The Calibration of Confidence Judgments

Unlike the findings suggesting an underconfidence bias for perceptual tasks (Björkman et al., 1993; Juslin & Olsson, 1997; Winman & Juslin, 1993), the results of Experiment 2 yielded an overconfidence bias of about the same magnitude for matched general-knowledge and perceptual items. The overconfidence bias observed in Experiment 2 was clearly related to the same type of item-selection feature that has been claimed to contribute to the overconfidence bias in general-information questions—the deliberate inclusion of items that tend to result in a preponderance of erroneous judgments. In this respect, the results agree with the general premise of PMM theory (Gigerenzer et al., 1991), that the calibration of confidence judgments depends on the validity of the cues that underlie choice and confidence.

As noted earlier, however, the task used in Experiment 2 probably engaged cognitive processes beyond sensory encoding, and as a result the items tended to elicit consistent choices (and errors) both within persons and across persons (for other perceptual tasks exhibiting these features, see Vickers & Pietsch, 2001). In addition, for some of the items, mean percentage correct was less than 50%, as sometimes occurs for general-information questions (Koriat, 2008a). For a task involving simple sensory discriminations, consensuality and correctness are strongly correlated. Thus, it is quite possible that perceptual tasks that involve very simple sensory discriminations (Juslin & Olsson, 1997) are based on a process that differs from that assumed in SCM. These tasks might also yield a reliable underconfidence bias.

Consistent with SCM, the evaluation of confidence judgments against indices of self-consistency reduced or eliminated the overconfidence bias that was observed when confidence was evaluated against accuracy. This was true for both the perceptual items and their matched general-information questions. This observation supports the proposition that the overconfidence bias derives in part from the reliance of participants on reliability as a cue for validity.

## A General Process Underlying Subjective Confidence?

To what extent is SCM a general model of confidence judgments? In addition to the results presented in this study regarding confidence in perceptual judgments, several results, mostly unpublished, testify for the generality of the model across several 2AFC tasks.[3] Support for a large number of predictions derived from SCM regarding the basis of confidence judgments has been obtained so far for general-information questions (data based on Koriat, 2008a, as well as other data; see Koriat, 2010), a word-matching task (Koriat, 2010), social attitudes (Koriat & Adiv, 2010), and social beliefs (unpublished data). Support was also obtained for tasks requiring participants to guess other participants' responses to questions concerning social beliefs and social attitudes (unpublished data). In addition, for all the tasks in which there exists a criterion of accuracy, the results confirmed the consensuality principle, suggesting that the C/A correlation actually reflects a C/C correlation and that both correlations stem from the dependence of confidence on self-consistency.

The similarity of the results across domains accords with the assumption of SCM that confidence judgments are based primarily on mnemonic cues that reside in the feedback from task performance rather than on the content of declarative considerations. Clearly, the considerations that people make in attempting to choose between two different answers should differ in regards to perceptual judgments, memory questions, and social beliefs and attitudes. However, mnemonic cues such as self-consistency and choice latency (and also accessibility, i.e., the number of clues that come to mind, in the case of general-knowledge questions, see Koriat, 2008b) are structural in nature, indifferent to the content of the considerations that guide the choice itself. Hence the architecture of the processes underlying confidence seems to be the same for 2AFC questions regardless of their content.

Possibly, however, SCM for 2AFC tasks is confined to situations in which the stimuli evoke some conflict or deliberation between different representations or considerations. Thus, it might not hold true for perceptual-comparison task for which the answer is immediately obvious, that is, tasks that are not within the "uncertainty zone" (see Winman & Juslin, 1993). Similarly, it might not apply to simple sensory discriminations such as those involving comparison of the length of (straight) lines or the weight of objects (Juslin & Olsson, 1997). In terms of the findings of Keren (1988), these are tasks that do not require additional higher processing beyond sensory encoding.

Although SCM is very rudimentary and incorporates strong assumptions, it yielded a large number of gross predictions that were generally confirmed. In future work, investigators must attempt to refine the model in order to allow more detailed, quantitative predictions. It also would be interesting see to what extent the SCM can be extended to tasks other than 2AFC tasks, for example, those involving more than two alternative options.

[3] The results of the unpublished studies can be obtained from the author.

## References

Audley, R. J. (1960). A stochastic model for individual choice behavior. *Psychological Review, 67,* 1–15. doi:10.1037/h0046438

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55,* 412–428.

Baranski, J. V., & Petrusic, W. M. (1995). On the calibration of knowledge

and perception. *Canadian Journal of Experimental Psychology, 49,* 397–407. doi:10.1037/1196-1961.49.3.397

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 929–945. doi:10.1037/0096-1523.24.3.929

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception & Psychophysics, 61,* 1369–1383.

Barnes, A. E., Nelson, T. O., Dunlosky, J., Mazzoni, G., & Narens, L. (1999). An integrative system of metamemory components involved in retrieval. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance. Interaction of theory and application* (pp. 287–313). Cambridge, MA: MIT Press.

Bassili, J. N. (1996). Meta-judgmental versus operative indexes of psychological attributes: The case of measures of attitude strength. *Journal of Personality and Social Psychology, 71,* 637–653. doi:10.1037/0022-3514.71.4.637

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics, 54,* 75–81.

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology, 72,* 691–695. doi:10.1037/0021-9010.72.4.691

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 85–123), Washington, DC: Winston.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes, 65,* 212–219. doi:10.1006/obhd.1996.0021

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory, 14,* 540–552. doi:10.1080/09658210600590302

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York, NY: Holt, Rinehart, & Winston.

Dawes, R. M. (1980). Confidence in intellectual judgments vs. confidence in perceptual judgments. In E. D. Lanterman & H. Feger (Eds.), *Similarity and choice: Papers in honour of Clyde Coombs* (pp. 327–345). Bern, Switzerland: Huber.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130,* 959–988. doi:10.1037/0033-2909.130.6.959

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General, 130,* 579–599. doi:10.1037/0096-3445.130.4.579

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

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

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57,* 94–107. doi:10.1037/h0058559

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

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564. doi:10.1037/0096-1523.3.4.552

Fullerton, G. S., & Cattell, J. M. (1892). On the perception of small differences. In *Publications of the University of Pennsylvania*: *Philosophical Monograph Series,* No.2. Philadelphia, PA: University of Pennsylvania Press.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528. doi:10.1037/0033-295X.98.4.506

Gill, M. J., Swann, W. B., Jr., & Silvera, D. H. (1998). On the genesis of confidence. *Journal of Personality and Social Psychology, 75,* 1101–1114. doi:10.1037/0022-3514.75.5.1101

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

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler, & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 177–198). Malden, MA: Blackwell. doi:10.1002/9780470752937.ch9

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435. doi:10.1016/0010-0285(92)90013-R

Haddock, G., Rothman, A. J., Reber, R., & Schwarz, N. (1999). Forming judgments of attitude certainty, importance, and intensity: The role of subjective experiences. *Personality and Social Psychology Bulletin, 25,* 771–782. doi:10.1177/0146167299025007001

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16,* 107–112. doi:10.1016/S0022-5371(77)80012-1

Hertwig, R., Herzog, S. M., Schooler, L. J., & 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,* 1191–1206. doi:10.1037/a0013025

Hoffrage, U. (2004). Overconfidence. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgment, and memory* (pp. 235–254). Hove, England: Psychology Press.

Holland, R. W., Verplanken, B., & van Knippenberg, A. (2003). From repetition to conviction: Attitude accessibility as a determinant of attitude certainty. *Journal of Experimental Social Psychology, 39,* 594–601. doi:10.1016/S0022-1031(03)00038-6

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

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 226–246. doi:10.1006/obhd.1994.1013

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104,* 344–366. doi:10.1037/0033-295X.104.2.344

Juslin, P., Winman, A., & Olsson, H. (2003). Calibration, additivity, and source independence of probability judgments in general knowledge and sensory discrimination tasks. *Organizational Behavior and Human Decision Processes, 92,* 34–51. doi:10.1016/S0749-5978(03)00063-3

Kelley, C. M., & 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. doi:10.1006/jmla.1993.1001

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica, 67,* 95–119. doi:10.1016/0001-6918(88)90007-8

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15,* 321–341. doi:10.1002/acp.705

Koriat, A. (1976). Another look at the relationship between phonetic

symbolism and the feeling of knowing. *Memory & Cognition, 4,* 244–248.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639. doi:10.1037/0033-295X.100.4.609

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124,* 311–333. doi:10.1037/0096-3445.124.3.311

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition, 9,* 149–171. doi:10.1006/ccog.2000.0433

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

Koriat, A. (2008a). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 945–959. doi:10.1037/0278-7393.34.4.945

Koriat, A. (2008b). When confidence in a choice is independent of which choice is made. *Psychonomic Bulletin & Review, 15,* 997–1001. doi:10.3758/PBR.15.5.997

Koriat, A. (2010). *The self-consistency model of subjective confidence.* Manuscript submitted for publication.

Koriat, A., & Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science, 13,* 441–453. doi:10.1111/j.1467-7687.2009.00907.x

Koriat, A., & Adiv, S. (2010). *The construction of attitudinal judgments: Evidence from attitude certainty and response latency.* Manuscript submitted for publication.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490–517. doi:10.1037/0033-295X.103.3.490

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118. doi:10.1037/0278-7393.6.2.107

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52,* 478–492. doi:10.1016/j.jml.2005.01.001

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

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory*: *Essays in honor of Thomas O. Nelson* (pp. 117–135). New York, NY: Psychology Press.

Koriat, A., Pansky, A., & Goldsmith, M. (2010). An output-bound perspective on false memories: The case of the Deese–Roediger–McDermott (DRM) paradigm. In A. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in Honor of Robert A. Bjork* (pp. 301–332). London, England: Psychology Press.

Kornell, N. (2009). Metacognition in humans and animals. *Current Directions in Psychological Science, 18,* 11–15. doi:10.1111/j.1467-8721.2009.01597.x

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York, NY: Cambridge University Press.

Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin, 126,* 109–137. doi:10.1037/0033-2909.126.1.109

McKenzie, C. R. M. (1997). Underweighting alternatives and overconfidence. *Organizational Behavior and Human Decision Processes, 71,* 141–160. doi:10.1006/obhd.1997.2716

McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 771–792. doi:10.1037/0278-7393.24.3.771

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology: General, 135,* 391–408. doi:10.1037/0096-3445.135.3.391

Modirrousta, M., & Fellows, L. K. (2008). Medial prefrontal cortex plays a critical and selective role in "feeling of knowing" meta-memory judgments. *Neuropsychologia, 46,* 2958–2965. doi:10.1016/j.neuropsychologia.2008.06.011

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

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51,* 102–116. doi:10.1037/0003-066X.51.2.102

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 125–173). San Diego, CA: Academic Press.

Olsson, H. (1999). A sampling model of confidence in sensory discrimination. *Acta Universitatis Upsaliensis: Comprehensive summaries of Uppsala Dissertations from the Faculty of Social Sciences, 87.*

Palmer, S. (1999). *Vision science: From photons to phenomenology.* Cambridge, MA: MIT Press.

Peirce, C. S., & Jastrow, J. (1884). On small differences of sensation. *Memoirs of the National Academy of Science, 3,* 73–83.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience, 10,* 257–261. doi:10.1038/nn1840

Petrusic, W. M., & Baranski, J. V. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *American Journal of Psychology, 110,* 543–572. doi:10.2307/1423410

Petrusic, W. M., & Baranski, J. V. (2009). Probability assessment with response times and confidence. *Acta Psychologica, 130,* 103–114. doi:10.1016/j.actpsy.2008.10.008

Read, J. D., Lindsay, D. S., & Nicholls, T. (1998). The relation between confidence and accuracy in eyewitness identification studies: Is the conclusion changing? In C. P. Thompson, D. J. Herrmann, J. D. Read, D. Bruce, D. G. Payne, & M. P. Toglia (Eds.), *Eyewitness memory: Theoretical and applied perspectives* (pp. 107–130). Mahwah, NJ: Erlbaum.

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82,* 416–425. doi:10.1037/0021-9010.82.3.416

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

Sampaio, C., & Brewer, W. F. (2009). The role of unconscious memory errors in judgments of confidence for sentence recognition. *Memory & Cognition, 37,* 158–163. doi:10.3758/MC.37.2.158

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feeling of knowing. *Psychonomic Bulletin & Review, 1,* 357–375.

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93–113). Cambridge, MA: MIT Press.

Shaw, J. S., III. (1996). Increases in eyewitness confidence resulting from postevent questioning. *Journal of Experimental Psychology: Applied, 2,* 126–146. doi:10.1037/1076-898X.2.2.126

Smith, J. D., Beran, M. J., Couchman, J. J., & Coutinho, M. V. C. (2008).

The comparative study of metacognition: Sharper paradigms, safer inferences. *Psychonomic Bulletin & Review, 15,* 679–691. doi:10.3758/PBR.15.4.679

Son, L. K., & Schwartz, B. L. (2002). The relation between metacognitive monitoring and control. In T. J. Perfect & B. S. Schwartz (Eds.), *Applied metacognition* (pp. 15–38). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511489976.003

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118,* 315–327. doi:10.1037/0033-2909.118.3.315

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence, 25,* 93–109. doi:10.1016/S0160-2896(97)90047-7

Stephen, A. T., & Pham, M. T. (2008). On feelings as a heuristic for making offers in ultimatum negotiations. *Psychological Science, 19,* 1051–1058. doi:10.1111/j.1467-9280.2008.02198.x

Tormala, Z. L., & Rucker, D. D. (2007). Attitude certainty: A review of past findings and emerging perspectives. *Social and Personality Psychology Compass, 1,* 469–492. doi:10.1111/j.1751-9004.2007.00025.x

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600. doi:10.1037/0278-7393.26.3.582

Vickers, D. (1979). *Decision processes in visual perception.* New York, NY: Academic Press.

Vickers, D., & Pietsch, A. (2001). Decision-making and memory: A critique of Juslin and Olsson's (1997) sampling model of sensory discrimination. *Psychological Review, 108,* 789–804. doi:10.1037/0033-295X.108.4.789

Vickers, D., Smith, P., Burt, J., & Brown, M. (1985). Experimental paradigms emphasizing state or process limitations: II. Effects on confidence. *Acta Psychologica, 59,* 163–193. doi:10.1016/0001-6918(85)90018-6

Waters, H. S., & Schneider, W. (Eds.). (2010). *Metacognition, strategy use, and instruction.* New York, NY: Guilford.

Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgments: Two different accounts. *Scandinavian Journal of Psychology, 34,* 135–148. doi:10.1111/j.1467-9450.1993.tb01109.x

Wright, G., & Ayton, P. (1988). Decision time, subjective probability, and task difficulty. *Memory & Cognition, 16,* 176–185.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110,* 611–617. doi:10.1037/0033-2909.110.3.611

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice–Hall.

Yzerbyt, V. Y., Lories, G., & Dardenne, B. (Eds.). (1998). *Metacognition: Cognitive and social dimensions.* Thousand Oaks, CA: Sage.

Zakay, D., & Tuvia, R. (1998). Choice latency times as determinants of post-decisional confidence. *Acta Psychologica, 98,* 103–115. doi:10.1016/S0001-6918(97)00037-1

Received April 18, 2010
Revision received October 22, 2010
Accepted October 25, 2010 ■