Journal of Experimental Psychology: General
2018, Vol. 147, No. 5, 613–631

© 2018 American Psychological Association
0096-3445/18/$12.00 http://dx.doi.org/10.1037/xge0000397

# When Reality Is Out of Focus: Can People Tell Whether Their Beliefs and Judgments Are Correct or Wrong?

Asher Koriat
University of Haifa

Can we tell whether our beliefs and judgments are correct or wrong? Results across many domains indicate that people are skilled at discriminating between correct and wrong answers, endorsing the former with greater confidence than the latter. However, it has not been realized that because of people's adaptation to reality, representative samples of items tend to favor the correct answer, yielding object-level accuracy (OLA) that is considerably better than chance. Across 16 experiments that used 2-alternative forced-choice items from several domains, the confidence/accuracy (C/A) relationship was positive for items with OLA >50%, but consistently negative across items with OLA <50%. A systematic sampling of items that covered the full range of OLA (0–100%) yielded a U-function relating confidence to OLA. The results imply that the positive C/A relationship that has been reported in many studies is an artifact of OLA being better than chance rather than representing a general ability to discriminate between correct and wrong responses. However, the results also support the ecological approach, suggesting that confidence is based on a frugal, "bounded" heuristic that has been specifically tailored to the ecological structure of the natural environment. This heuristic is used despite the fact that for items with OLA <50%, it yields confidence judgments that are counterdiagnostic of accuracy. Our ability to tell between correct and wrong judgments is confined to the probability structure of the world we live in. The results were discussed in terms of the contrast between systematic design and representative design.

*Keywords:* ecological approach, heuristics, metacognition, representative and systematic design, subjective confidence

In a scene from Woody Allen's movie *Deconstructing Harry*, a cameraman attempts to take a shot of Robin Williams who plays Harry. Noticing that the image is blurred, he tries to adjust the camera lens but fails to achieve a sharper image, finally realizing that there is a more fundamental problem: It is Harry himself who is "out of focus."

This episode suggests a distinction between two types of problems that researchers may run into when sampling information from the world. The first is when the camera is out of focus, as when the samples drawn are unrepresentative, yielding a distorted portrayal of reality. The second is when reality itself is out of focus. In that case, errors may stem precisely from the attempt to draw representative samples that are faithful to reality, not considering the possibility that reality itself might be biased with respect to the issue investigated.

In this article I focus on people's metacognitive accuracy—the ability to discriminate between true and false beliefs and judgments. I will examine how the two types of problems mentioned above may affect researchers' conclusions about the question how good people are at telling whether their beliefs or judgments are right or wrong.

In what follows I will first review the long-standing historical controversy in cognitive research between researchers who stress the importance of a representative design that mirrors the conditions that exist in the natural ecology versus those who favor the systematic study of phenomena independent of the distribution of events in the outside world. In recent years the plea for a representative design has been voiced particularly by advocates of the ecological approach to cognition, who extended their research to issues about metacognitive accuracy.

Turning then to questions about metacognitive accuracy, a distinction will be drawn between two aspects of metacognitive accuracy, calibration and resolution. Focusing first on calibration, I examine the claim of proponents of the ecological approach that the poor calibration that has been documented in several studies derives merely from researchers' failure to sample items representatively from the natural ecology (i.e., our camera is out of focus). Focusing next on resolution, I will argue that it is precisely representative sampling that may be responsible for researchers' conclusions about people's general ability to tell correct from wrong answers. This is because samples of items drawn from the natural environment to which people have adapted are bound to be biased (i.e., the natural environment is "out of focus"). The experiments to be presented examine these arguments.

This article was published Online First January 25, 2018.

I am grateful to Miriam Gil and Dan Manor for their help in the analyses of the results, and to Etti Levran (Merkine) for her help in copyediting. Shai Raz and Irit Lindenfeld helped in the data collection. This work was supported by Grant 2013039 from the United States–Israel Binational Science Foundation.

Correspondence concerning this article should be addressed to Asher Koriat, Department of Psychology, University of Haifa, Haifa, Israel, 3498838. E-mail: akoriat@univ.haifa.ac.il

## Representative Design Versus Systematic Design in Cognition Research

Historically, there has been a tension between two experimental paradigms for the study of human cognition, systematic design and representative design (Brunswik, 1955a, 1955b; see Hoffrage & Hertwig, 2006). Systematic design, which is characteristic of mainstream cognitive science, involves the laboratory-based investigation of psychological processes under controlled conditions. These conditions allow researchers to isolate variables and to examine systematically the effects of variations in these variables on different aspects of performance. Systematic design helps specify the basic laws relating stimulus variations to performance and behavior.

Egon Brunswik (1944, 1955b, 1956), however, argued that the method of systematic design destroys the causal structure of the natural environment to which psychological processes have been adapted, and leads researchers to use artificial conditions and stimuli that hardly exist in the world. He called for a design of experiments that is representative of the organism's ecology, arguing that the conditions of experiments should represent the real-life conditions over which generalization is to be achieved.

Other researchers also emphasized the study of cognitive processes under naturalistic, real-world conditions. James Gibson (1979) objected to the practice of setting up laboratory situations that are convenient for the experimenter but atypical for the individual. In his ecological approach to visual perception, he emphasized the structure and richness of the sensory experience afforded to perceivers in the natural environment.

In the area of memory, Ulric Neisser (1978, 1985) dismissed the results observed within the laboratory-based research tradition, arguing for the study of memory under naturalistic conditions. His call for the ecological study of memory has sparked a heated debate between proponents of naturalistic memory research and researchers who favor laboratory-based research (to which *American Psychologist* devoted its January 1991 issue). Some researchers claimed that memory in real-life situations differs in significant ways from the kind of memory that has been traditionally investigated (see Koriat & Goldsmith, 1996a).

In developmental psychology, Bronfenbrenner (1977) argued that research on human development should focus on the progressive accommodation of individuals to their changing environments. He called for ecologically valid research that is carried out in a naturalistic setting and involves objects and activities from everyday life.

In recent years, Brunswik's plea for a representative design has enjoyed a renewed interest among researchers in the area of judgment and decision making in the context of the so-called Neo-Brunswikian approach (see Juslin, & Montgomery, 2007). Advocates of the ecological approach (see Björkman, 1994; Dhami, Hertwig, & Hoffrage, 2004; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Hoffrage & Hertwig, 2006; Juslin, 1994) argued that some of the cognitive illusions that had been documented in the literature, such as the overconfidence bias (Gigerenzer et al., 1991; Juslin, 1994; Juslin, Winman, & Olsson, 2000), the hard-easy effect (Juslin, 1993; Juslin et al., 2000), the hindsight bias (Winman, 1997; Winman, Juslin, & Björkman, 1998), and the availability bias (Sedlmeier, Hertwig, & Gigerenzer, 1998) are not real but stem from researchers' failure to sample items representatively from the natural ecology (see Gigerenzer, 2004).

## The Monitoring of One's Own Knowledge

I now examine how the contrast between representative design and systematic design bears on the question of people's metacognitive accuracy, which is the focus of the present study. I will review briefly what is known about people's ability to discriminate between true and false beliefs and judgments. Later I will examine critically the research testifying for people's ability to monitor the accuracy of their knowledge.

Issues concerning metacognitive accuracy have been discussed by philosophers, statisticians, psychologists, and forensic scientists (Carruthers, 2011; Dunlosky & Metcalfe, 2009; Koriat, 2007; Proust, 2013; Schwarz, 2015). In the philosophical Traditional Analysis of Knowledge (TAK), propositional knowledge is defined as Justified True Belief (JTB). Controversies, however, exist about what makes a belief justified so that if true, it will be *known* to be so. Different proposals have been advanced about epistemic justification but little agreement exists to the extent that adherents of the philosophical skepticism view have raised the question whether knowledge, in the first place, is possible.

In turn, within experimental psychology, the study of metacognitive judgments has produced empirical evidence suggesting that people are relatively skilled at monitoring the accuracy of their knowledge and judgments (see Dunlosky & Metcalfe, 2009; Metcalfe & Dunlosky, 2008). The general accuracy of metacognitive judgments has led several authors to suggest that these judgments are based on participants' direct access to the underlying memory traces (see Schwartz, 1994). For example, it was proposed that judgments of learning (JOLs) during study are based on detecting the strength of the memory trace that is formed following learning (e.g., Cohen, Sandler, & Keglevich, 1991). Similarly, feeling-of-knowing (FOK) judgments were assumed to monitor the actual presence of the elusive target in memory (Hart, 1965). In the case of confidence judgments too, a direct access view generally underlies the use of such judgments in the context of strength theories of memory (see Van Zandt, 2000).

However, the view that has gathered a great deal of support in recent years is that metacognitive judgments are inferential in nature, based on a variety of cues and beliefs that have some validity in predicting correct performance (Benjamin & Bjork, 1996; Jacoby, Kelley, & Dywan, 1989; Koriat, 2007). Results suggest that JOLs made during study rest on the ease with which items are encoded or retrieved during learning (Karpicke, 2009; Koriat & Ma'ayan, 2005; Koriat, Ma'ayan, & Nussinson, 2006), and on beliefs about the variables that affect memory performance (Mueller, Tauber, & Dunlosky, 2013). FOK judgments were said to rely on the familiarity of the pointer that serves to probe memory, and on the accessibility of partial information about the elusive memory target (Koriat, 1993; Koriat & Levy-Sadot, 2001; Schwartz & Metcalfe, 1992). In turn, confidence judgments were claimed to rest on mnemonic cues such as the fluency of selecting or retrieving an answer (e.g., Kelley & Lindsay, 1993; Koriat et al., 2006), or on the considerations retrieved from memory (Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980). Reliance on these cues was assumed to explain both the accuracy and inaccuracy of metacognitive judgments.

Discussions of metacognitive accuracy and inaccuracy have distinguished between two aspects of the confidence/accuracy correspondence, calibration and resolution. This distinction applies to different metacognitive judgments, but we shall focus here on confidence judgments. Calibration (or "bias," "absolute accuracy," Dunlosky & Metcalfe, 2009; Yates, 1990) refers to the extent to which people's judgments demonstrate overconfidence (judgments are more optimistic than actual performance) versus underconfidence (judgments are less optimistic than actual performance; Lichtenstein, Fischhoff, & Phillips, 1982). A simple measure of calibration is the difference between mean confidence and mean accuracy across items (when both are assessed on the same scale). Monitoring resolution (also called "relative accuracy," "discrimination accuracy," or "type 2 sensitivity," Fleming, & Lau, 2014; Nelson & Dunlosky, 1991; Yaniv, Yates, & Smith, 1991) refers to the extent to which metacognitive judgments discriminate between correct and wrong answers. In metacognition research, resolution has been measured by the within-person confidence—accuracy (C/A) gamma correlation (Nelson, 1984) but other measures have also been used (Benjamin & Diaz, 2008; Fleming & Lau, 2014; Higham, Perfect, & Bruno, 2009; Yaniv et al., 1991).

Much of the work on calibration was carried out in the area of judgment and decision-making. The implicit assumption underlying that work is that calibration ought to be perfect, and the challenge is to explain why confidence judgments deviate from perfect calibration (see Griffin & Brenner, 2004). In contrast, research in metacognition has focused primarily on resolution (see Dunlosky & Metcalfe, 2009; Koriat, 2007). This research was initially motivated by observations indicating that metacognitive judgments are generally accurate in predicting memory performance (Brown & McNeill, 1966; Hart, 1965), which raised the question how do people know that they know? In the context of the inferential approach to metacognitive judgments, this question generally meant why resolution is better than chance (rather than why it is not perfect; see Koriat, 2016).

It should be noted that calibration and resolution are theoretically independent: Calibration can be perfect when resolution is very low, and vice versa (Fleming, & Lau, 2014; Koriat & Goldsmith, 1996b). For example, assume that for a given two-alternative forced-choice (2AFC) test, a person's probability correct is .60. That person might be extremely overconfident and yet exhibit perfect resolution if he assigns a probability of .90 to all correct answers and a probability of .89 to all incorrect answers. However, as will be shown later, calibration and resolution are not independent when the probabilistic structure of the environment is considered.

## Calibration: Representative Design and the Overconfidence Bias

Let us focus first on the calibration of confidence judgments. Many studies indicated that people are overconfident. Specifically, for 2AFC almanac questions, the subjective probability assigned to the correctness of answers exceeds the proportion of correct answers. The overconfidence bias has been observed across a wide range of conditions (Allwood & Montgomery, 1987; Arkes, Christensen, Lai, & Blumer, 1987; Dunning, Heath, & Suls, 2004; Griffin & Brenner, 2004; Hoffrage, 2004; Koriat et al., 1980; Lichtenstein & Fischhoff, 1977; Soll, 1996), and several explanations of this bias have been proposed (Erev, Wallsten, & Budescu, 1994; Griffin & Tversky, 1992; Koriat et al., 1980; Metcalfe, 1998; Nickerson, 1998). However, proponents of the ecological approach to judgment and decision argued that this bias is largely a pseudophenomenon, resulting from researchers' failure to sample information representatively from the natural environment. Indeed, several experiments have yielded evidence indicating that when items are selected representatively from their reference class, the overconfidence bias is either strongly reduced or entirely eliminated (Gigerenzer et al., 1991; Juslin, 1994). Thus, the study of the overconfidence bias (as well as other claimed biases, see Hoffrage & Hertwig, 2006) has revived the long-standing controversy between proponents of a systematic design and advocates of a representative design that respects the ecological structure of the natural environment.

What are the arguments behind these findings? Advocates of the ecological approach have argued that when compiling items for a test, experimenters tend to select items that tax subjects' knowledge, oversampling difficult or misleading items (Björkman, 1994; Hoffrage & Hertwig, 2006). However, in the real world, people are quite good at judging the reliability of their knowledge, because the cues that they use in making their decisions and confidence are generally valid. Therefore, calibration should be relatively good provided general-knowledge items are representatively sampled from their reference class.

An influential theory of confidence judgments that incorporates these arguments has been proposed by Gigerenzer et al. (1991). In their theory of Probabilistic Mental Models, they assume that when presented with a 2AFC almanac question, participants test several cues in turn until they identify a cue that discriminates between the two answers. When that cue determines the choice, its cue validity is then reported as the confidence in the choice. A critical feature of this theory (see also Juslin, 1994) is the emphasis on learning and adaptation. Consistent with Brunswik's view (1956), it is assumed that in the course of their interaction with the environment, people internalize the associations between cues and events in the world. Reliance on the internalized knowledge contributes to the general accuracy of confidence judgments (see Fiedler, 2007). Gigerenzer et al. (1991) reported evidence indicating a strong overconfidence bias when general-knowledge items were selected informally. However, this bias disappeared when the items were randomly selected from their reference class.

In terms of the "out of focus" example mentioned in the introduction, the ecological approach assumes that a representative design can mend the biased picture produced by the informal sampling of stimuli from the environment.

## Monitoring Resolution: Discriminating Between Correct and Wrong Answers

We turn next to resolution, which captures the concern of philosophers with the question whether we can establish that a particular belief is true and another is false. As noted by Lichtenstein and Fischhoff (1977), "resolution is a more fundamental aspect of probabilistic functioning, for it reflects the ability to sort items into subcategories whose percentage correct is maximally different from the overall percentage correct" (p. 181).

Resolution received particular attention in metacognition and educational research because of findings indicating that people rely heavily on their metacognitive judgments in the strategic regulation of cognitive processes and behavior (Dunlosky & Metcalfe, 2009; Goldsmith & Koriat, 2008; Jackson & Kleitman, 2014; Koriat & Goldsmith, 1996b; Sternberg, 1998; Thiede, Anderson, & Therriault, 2003). Social psychologists also noted that confidence affects the likelihood that people translate their beliefs into behavior (Gill, Swann, & Silvera, 1998).

Several studies indicated that metacognitive judgments exhibit better than chance resolution. Thus, JOLs during study were found to predict the future recall of different items (Nelson & Dunlosky, 1991), and FOK judgments were found to predict the likelihood of recalling or recognizing a momentarily unrecallable memory target (Hart, 1965; Koriat, 1993). Retrospective confidence in one's chosen answer has been also found to predict the correctness of that answer, as we now review.

A wealth of studies have indicated that the within-person C/A correlation is better than chance for many tasks in such domains as general-knowledge (Kleitman & Stankov, 2001; Koriat & Goldsmith, 1996b), perceptual judgments (Keren, 1991), memory performance (Kelley & Sahakyan, 2003; Mickes, Hwe, Wais, & Wixted, 2011; Thompson & Mason, 1996), and achievement and intelligence tests (Jackson & Kleitman, 2014; Schraw & Nietfeld, 1998; Sheffer, 2003; Stankov & Crawford, 1996). Recent evidence indicates that this is also largely true of eyewitness testimony (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted & Wells, 2017). For general-information questions, for example, the within-person Gamma correlation averaged .68 for forced-choice questions (.87 for open-ended questions) in the study of Koriat and Goldsmith (1996b). Similarly, with regard to recognition memory, it was noted that "low-confidence recognition decisions are often associated with close-to-chance accuracy, whereas high-confidence recognition decisions can be associated with close-to-perfect accuracy" (Mickes et al., 2011; p. 239; see Tulving & Thomson, 1971). The C/A correlation is particularly strong for sensory discrimination tasks.

Thus, by and large, people are skilled at telling whether their responses are correct or wrong. In fact, the ability to monitor one's own knowledge was seen by Tulving and Madigan (1970) as "one of the truly unique characteristics of human memory" (p. 477). Of course, this ability is far from being perfect, and differs greatly across tasks. Nevertheless, with a few exceptions (see below), whenever a C/A correlation has been reported in the literature, that correlation is generally positive and usually significant. As noted earlier, for some authors, the predictive validity of metacognitive judgments suggested that people can monitor *directly* the accuracy of their response (see Busey, Tunnicliff, Loftus, & Loftus, 2000; Schwartz, 1994). The observation that the C/A correlation is reliably better than chance in many tasks and domains would seem to contrast with the great bewilderment that exists in the philosophy of knowledge about belief justification.

## Object-Level Accuracy and Meta-Level Accuracy

In this article, I argue that a representative design in the broad sense is liable to yield misleading conclusions about people's general ability to discriminate between true and false beliefs and judgments. This is because the natural, real-life environment is "out of focus" with respect to the issue of monitoring resolution. I will clarify this argument using the distinction between object-level and metalevel performance (Nelson & Narens, 1990). Object-level accuracy (OLA) refers to the correspondence between people's first-order decisions and some objective criterion of correctness ("truth"). It can be indexed by the percentage of correct answers or by the Type-1 $d'$ index in signal detection theory (SDT, Green & Swets, 1966). Metalevel accuracy (MLA) in turn, refers to monitoring resolution—the correspondence between confidence in one's first-order decision and the accuracy of that decision. It can be measured by the C/A gamma correlation (Nelson, 1984) or by Type-2 indexes derived from SDT (Benjamin & Diaz, 2008; Fleming & Lau, 2014; Higham et al., 2009; Maniscalco, & Lau, 2014).

Representative sampling creates a bias in OLA for the very reason that proponents of the ecological approach have preached for the use of a representative design. Their argument is that conditions and stimuli ought to be selected from the natural environment because organisms have been adapted to their environment through evolution and learning (Brunswik, 1955b; Dhami et al., 2004; Hoffrage & Hertwig, 2006). However, precisely because of that, representative samples of stimuli are bound to be selective as far as OLA is concerned, yielding a much better accuracy than would be expected by chance. As will be shown below, the bias inherent in representative sampling is critical for the conclusions reached about MLA.

Let us examine the bias in OLA. Consider first general knowledge questions. Even when researchers do not make special effort to select items representatively, the items selected tend to yield more correct answers than wrong answers as a result of people's adaptation and learning. However, representative sampling tends to aggravate the bias. To illustrate, in Experiment 1 of Gigerenzer et al. (1991), OLA averaged 52.9% for 2AFC items selected informally, but 71.7% when items were selected representatively. The respective means in Experiment 2 were 56.2% and 75.3%. Similarly, in Juslin's study (1994), the respective means were 62.8% and 76.2%. In two other studies (Koriat, 2012c, Study 2), in which I selected 2AFC items randomly from their reference classes ("which of two European countries has a larger population/a larger area?"), OLA averaged 79.5% and 78.2%, respectively. Thus, as might be expected, representative sampling is bound to yield OLA that is markedly better than chance.

The point that I am making is trivial. Readers can prove to themselves that it does not take much effort to compile a set of very "difficult" 2AFC almanac questions for which OLA is around 50%. However, let the reader try to assemble a large enough set of almanac questions for which OLA is reliably *below chance* (as some did, Fischhoff, Slovic, & Lichtenstein, 1977; Koriat, 1995, 2008). Thus, typically, for sets of 2AFC almanac items, the distribution of OLA across items covers only half of the potential range—that between 50% and 100%, which is generally taken to represent "item difficulty." There is very little representation of items for which OLA is at the range 0%-50%.

Consider perception next. In his classic study of size perception, Brunswik (1944) had a graduate student estimate the size of various objects in her natural environment when she was interrupted randomly in the course of her activities. His study was intended to demonstrate the high accuracy of size perception for a representative sample of stimuli drawn from the natural environment. Brunswik argued that this high accuracy is the result of people's learning to use proximal visual cues in accordance with their ecological validity.

In general, in justifying the need for a representative design, Brunswik and proponents of the Neo-Brunswikian approach (Dhami et al., 2004; Gigerenzer et al., 1991; Juslin & Montgomery, 2007) have emphasized the contribution of experience to the adaptation of humans to their ecology. However, evolution has undoubtedly also contributed to OLA being considerably better than chance for basic capabilities. Thus, for simple sensory attributes, comparative psychophysical judgments are generally very accurate in mirroring the physical differences between the stimuli, leading researchers to argue that uncertainty in sensory discrimination tasks derives only from random noise in the nervous system (Juslin & Olsson, 1997).

Similarly, human memory is reliable by and large despite the occasional occurrence of memory distortions and memory intrusions. A word that is retrieved from a studied list is much more likely to be correct than wrong (Koriat, & Goldsmith, 1996b; Koriat, Goldsmith, & Pansky, 2000). This is true even for studies using DRM lists (Roediger & McDermott, 1995): Across more than 100 such studies reviewed (Koriat, Pansky, & Goldsmith, 2011), an item freely recalled had about a .90 probability of being correct. Retrieved partial information about an unrecallable word is also more likely to be correct than wrong (Koriat, 1993; Koriat, Levy-Sadot, Edry, & de Marcas, 2003). Recognition memory for words, pictures, and sentences is also quite remarkable (Shepard, 1967).

In sum, a representative sample of items is bound to be selective with respect to OLA because of the contributions of evolution and learning. The question is whether this biased representation of items does not yield misleading conclusions about people's MLA. That such might be the case is suggested by the proposal that the choice of an answer to a 2AFC item, and the confidence in that choice are based on the same process (Gigerenzer et al., 1991; Koriat, 2012a). Thus, if reality itself is "out of focus" as far as OLA is concerned, samples of items that are representative of reality are liable to yield faulty conclusions with regard to MLA.

How can the bias inherent in a representative design be circumvented? Two methodological options will be explored. The first, is to examine what happens when a relatively large set of "unrepresentative" 2AFC items is used for which OLA <50%. The second is to shift to a systematic design in which items are selected to represent the entire range of OLA from 0% to 100%.

## Project 1: Monitoring Resolution for Representative and Nonrepresentative Items

The studies conducted at the University of Haifa (Project 1 and Project 2) received the approval of the Ethics Committee of the Psychology Department at University of Haifa. All participants in these studies provided informed consent to a protocol approved by the committee. The materials and raw data for these studies are available for download at https://osf.io/kjzhw.

### Method and Results

In Project 1, I compiled results from 16 experiments each of which included a reasonable number of 2AFC items for which OLA was below 50%. Most of these results have been reported in the past primarily in attempting to clarify the bases of subjective confidence judgments. Here I put together these results to support a general proposition: The findings testifying to people's ability to discriminate between true and false judgments are attributable to the bias inherent in representative samples drawn from the natural ecology. When unrepresentative samples are used, confidence judgments fail to track the accuracy of people's responses, and worse yet, these judgments are *counterdiagnostic* of accuracy. Thus, discrimination ability is confined to the world we live in rather than representing a general aptitude.

The studies to be described are listed in Table 1, which includes several details about each study. The results for the first 11 studies

Table 1
*A List of the Studies Included in Project 1*

| Study | Number of Participants | Number of Items<br>ALL | CC/ Nondeceptive/ Studied | CW/ Deceptive/ Lures |
|---|---|---|---|---|
| 1. Word matching (Koriat, 1976) | 100 | 85 | 38 | 19 |
| 2. General knowledge (Koriat, 2008) | 41 | 105 | 35 | 13 |
| 3. Judgments of line lengths (Koriat, 2011; Experiment 1) | 39 | 40 | 32 | 8 |
| 4. Judgments of areas (Koriat, 2011; Experiment 2) | 41 | 40 | 21 | 15 |
| 5. Geographical relations (Koriat, 2017a; Experiment 3) | 50 | 40 | 21 | 17 |
| 6. Predicting others' preferences (Koriat, 2013) | 41 | 60 | 49 | 10 |
| 7. Predicting others' choices of which of two lines is longer (Koriat, 2017a; Experiment 1, Block 1) | 20 | 40 | 32 | 8 |
| 8. Predicting others' choices of which shape has a larger area (Koriat, 2017a; Experiment 1, Block 2) | 20 | 40 | 25 | 15 |
| 9. Predicting others' responses to social belief items (Koriat & Adiv, 2014) | 41 | 60 | 56 | 4 |
| 10. Predicting others' responses to social attitudes items (Koriat & Adiv, 2014) | 40 | 50 | 43 | 4 |
| 11. Predicting others' word associations (Koriat, 2017b) | 41 | 60 | 24 | 36 |
| 12. Recognition memory for sentences (Sampaio & Brewer, 2009) | 36 | 96 | 72 | 24 |
| 13. Geography questions (Brewer & Sampaio, 2012; Experiment 1) | 48 | 96 | 72 | 24 |
| 14. Geography questions (Brewer & Sampaio, 2012; Experiment 2) | 36 | 102 | 78 | 24 |
| 15. DRM paradigm (Roediger & DeSoto, 2014; Experiment 1) | 48 | 300 | 150 | 50 |
| 16. DRM paradigm (Roediger & DeSoto, 2014; Experiment 2) | 48 | 300 | 150 | 50 |

*Note*. The table indicates the source of each study, the number of participants and items used, and how many of these items were consensually- correct (CC), nondeceptive or studied, and how many were consensually-wrong (CW), deceptive, or nonstudied lures.

come from Koriat's lab. Participants in each study chose the correct answer to a series of 2AFC items and indicated their confidence in their response. The items in each study were then classified on the basis of the empirical results as Consensually-Correct (CC) or Consensually-Wrong (CW). CC items are those for which OLA was better than 50% across participants, and CW items were those with OLA below 50%. Figure 1 plots mean confidence judgments for correct and wrong answers separately for CC and CW items. The results are plotted in the same format for all studies. Thus, items for which OLA was exactly 50% were eliminated from the plots. In some studies, the task was repeated several times, but only the results from the first presentation were

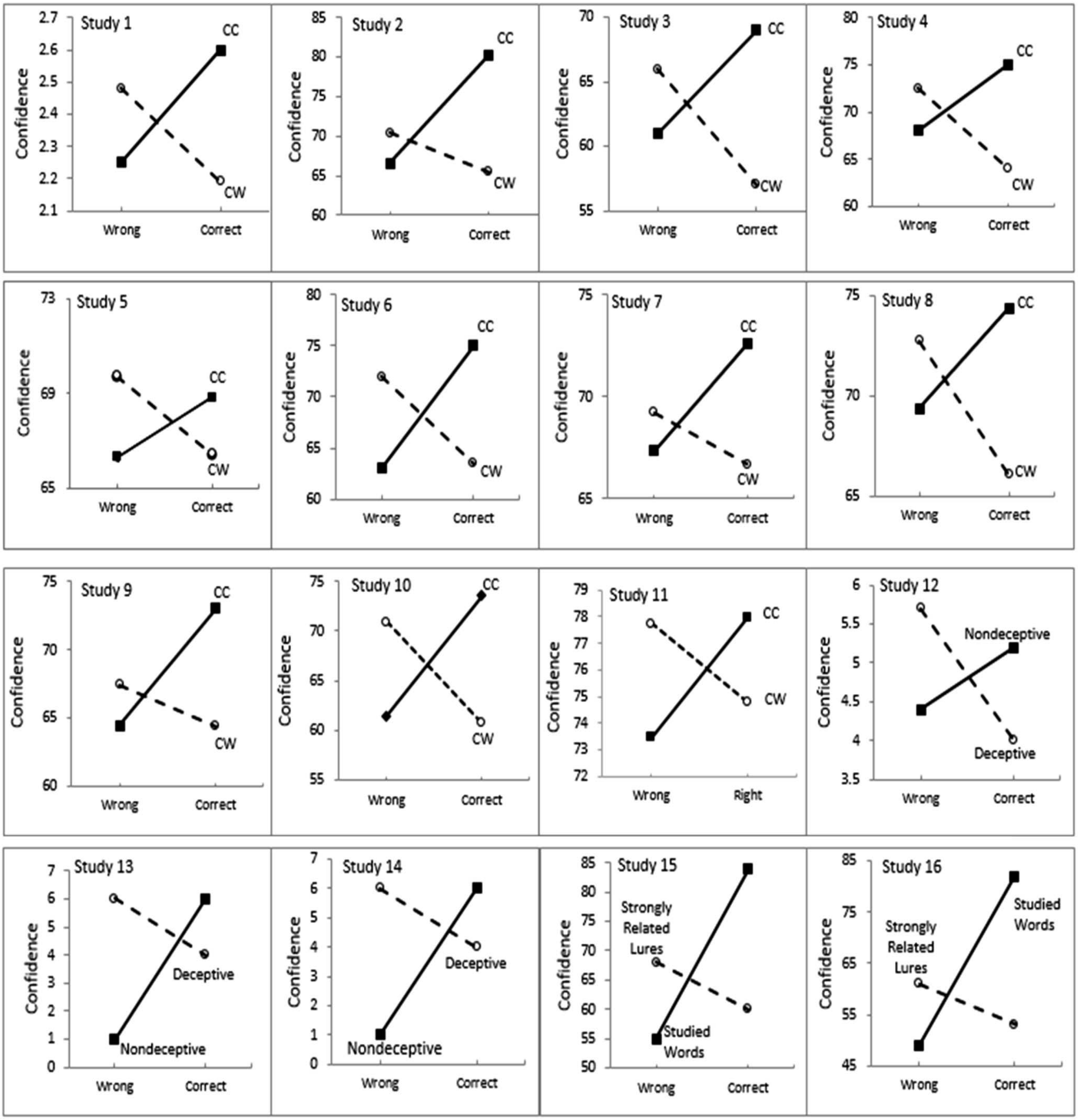

*Figure 1.* Mean confidence for correct and wrong answers, plotted separately for items for which the consensual answer tends to be the correct answer (CC/Nondeceptive/Studied), and for items for which the consensual answer tends to be the wrong answer (CW/Deceptive/Lures; Project 1).

used in Figure 1. In addition, in Study 1 and Study 2, a nonconsensual category was included, but the results for this category were deleted from the plots in Figure 1.

In Studies 1–5, a deliberate attempt was made to include a sufficiently large number of items that would be likely to yield a preponderance of wrong answers. However, the classification of items as CC or CW was determined ad hoc on the basis of the empirical results.

Study 1 was motivated by the results of an earlier study (Koriat, 1975) in which participants guessed the meaning of antonym words from three noncognate languages (e.g., *tuun*—luk) by matching them with their corresponding English translations (*deep*—shallow). The results, somewhat surprisingly, indicated that not only were participants significantly accurate in their guesses but that their accuracy increased significantly with their confidence in their response, indicating that they were able to monitor the accuracy of their guesses. In the subsequent study (Koriat, 1976), the same procedure was used but a subset of items was also included deliberately for which participants' matches were more likely to be wrong. Now it was clear (see Figure 1, Study 1) that accurate monitoring was confined to items for which OLA was better than chance. Thus, for the CC items, confidence increased with accuracy, as in the earlier study (Koriat, 1975), in which the items had been selected representatively (see Slobin, 1968). In contrast, for the CW items, confidence was significantly higher for the *wrong* responses than for the correct responses.

The same interactive pattern was observed in Study 2 (Koriat, 2008) which involved general-information questions. The study deliberately included a set of so-called "deceptive" or "misleading" questions (see Fischhoff et al., 1977; Gigerenzer et al., 1991; Koriat, 1995) for which people tend to choose the wrong answer. Studies 3 and 4 involved perceptual judgments. In Study 3, participants decided which of two irregular lines was longer, whereas in Study 4 they decided which of two shapes had a larger surface area (Koriat, 2011). Study 5 involved geography questions concerning the spatial relationship between two cities, (e.g., *which of the two cities is more to the north*, *Toronto, Canada, or Venice, Italy?*; Koriat, 2017a). The results of all five studies yielded the same crossover interaction: Confidence increased significantly with accuracy for CC items but decreased significantly with accuracy for CW items.

The next six studies (Studies 6–11) involved predictions of others' responses. Participants predicted for each 2AFC item which of the two options would be chosen by the majority of participants. In Study 6 (Koriat, 2013), participants predicted the personal preferences of others and indicated their confidence in their prediction. Their predictions were then compared with the actual normative choices (that were obtained independently), and the items were classified as CC or CW according to the correctness of the prediction made by the majority of participants. Whereas for CC items, confidence in predictions increased significantly with prediction accuracy, for CW items it decreased significantly with prediction accuracy.

For the purpose of the present report, the same type of analysis was applied to five other sets of data involving the prediction of others' responses (see Nickerson, 1998). These data, which were collected for other purposes, included the following: Predicting others' choices of which of two lines is longer (Study 7, Koriat, 2017b), of which shape has a larger area (Study 8, Koriat, 2017a), of others' responses to 2AFC social belief items (Study 9, Koriat & Adiv, 2014), and of others' responses to 2AFC social attitude items (Study 10, Koriat & Adiv, 2014; see Koriat, Adiv, & Schwarz, 2016). In Study 11, participants predicted which of two words people are likely to give as a response to a stimulus word in a word-association task (Koriat, 2017b). Except for Study 11, no attempt was made to include deliberately items for which majority predictions are likely to be wrong; the items were simply divided as CC or CW on the basis of the empirical results. It can be seen (see Figure 1) that in all six studies, participants' confidence increased with prediction accuracy only for CC items, whereas for CW items confidence decreased with prediction accuracy. Note that in Studies 6–10, CC items outnumbered CW items so that participants' predictions were largely correct.

The results for Studies 12–16 come from previous publications in which the division of items into two categories was based on a priori criteria that are related to OLA. For these studies, I simply plotted the results reported in these publications in the same format as that used for Studies 1–11. In Study 12 (Sampaio & Brewer, 2009), participants studied nondeceptive and deceptive sentences, and their recognition memory for these sentences was tested. Deceptive sentences were defined as those that tend to yield schema-based false recognition (e.g., judging that the sentence "The hungry python ate the mouse" was in the list, when the list actually included the sentence "The hungry python caught the mouse"). For nondeceptive sentences confidence was higher for correct responses than for wrong responses, whereas for deceptive sentences (for which mean OLA was below chance) it was higher for the wrong (false alarm) responses. A similar crossover interaction was obtained by Brewer and Sampaio (2006) using deceptive items that contained a possible synonym substitution, thus allowing errors based on gist memory (not shown in Figure 1). For nondeceptive sentences, confidence was significantly higher for correct responses ($M = 5.4$) than for wrong responses ($M = 4.8$) whereas for deceptive sentences it was significantly higher for wrong responses ($M = 5.8$) than for correct responses ($M = 5.8$).

A crossover interaction was also observed in Study 13 that used geography questions (Brewer & Sampaio, 2012; Experiment 1). Deceptive questions in that study were defined as those that produce a high proportion of errors that stem either from hierarchical reasoning (e.g., *Minneapolis, Minnesota, is south of Hamilton, Ontario*) or from alignment errors (e.g., *Lima, Peru is west of Miami, Florida*). The same pattern of results was obtained in Study 14 (Brewer & Sampaio, 2012; Experiment 2) which used similar materials.

In Study 15 (Roediger & DeSoto, 2014; Experiment 1), a DRM paradigm (Roediger & McDermott, 1995) was used. For studied words, confidence was higher for correct than for wrong recognition decisions, as is typically the case. In contrast, for strongly related, nonstudied lures, confidence was higher for wrong decisions than for correct decisions (see also DeSoto & Roediger, 2014; Kurdi, Diaz, Wilmuth, Friedman, & Banaji, 2016[1]). A similar pattern was observed in Study 16 (Roediger & DeSoto, 2014, Experiment 2). Note that for Studies 11–14 mean OLA was below 50% for the deceptive

[1] I analyzed the results of Experiment 1 of Kurdi et al. (2016) for the 10-list condition. For studied items, confidence for correct and wrong recognition judgments averaged 72.08, and 54.15, respectively, $t(59) = 9.07$, $p < .001$. The respective means for strong lures were 58.19 and 64.84, respectively $t(59) = 3.61$, $p < .001$ (eliminating four subjects who did not have means for all conditions). The interaction was significant, $F(1, 59) = 56.94$, $MSE = 159.22$, $p < .0001$.

items, whereas in studies 15–16, it amounted to 56%-57%, although it was much lower than for the studied items. A recent study on face recognition (Sampaio, Reinke, Mathews, Swart, & Wallinger, 2017) also yielded the same type of crossover interaction.[2]

The 16 studies cover a wide range of tasks including general-information, perceptual judgments, episodic memory for words and sentences, and predictions of other's responses in several domains. The generality of the crossover pattern across these tasks is impressive, particularly in the light of the current replication crisis. The results (see Figure 1) converge in demonstrating that MLA depends critically on OLA: For the "representative," CC items, higher confidence was predictive of better accuracy, consistent with many observations suggesting that participants are skilled at discriminating between correct and wrong judgments. In contrast, for the unrepresentative, CW items, confidence actually *decrease*d consistently with accuracy: People were more confident when they were wrong.

The crossover interaction that was exhibited by the results of all 16 studies was described by Koriat (2008, 2011, 2012a) in terms of the consensuality principle: Subjective confidence is actually correlated with the consensuality of the response rather than with its accuracy. The implication is that people are generally successful in monitoring the accuracy of their performance only because in the real world, accuracy and consensuality are confounded: The consensually selected response tends to be the correct response.

## Discussion

What is the theoretical explanation of the crossover C/A interaction documented in Figure 1? I will briefly sketch three theories that have been proposed to account for this interactive pattern, beginning with the self-consistency model (SCM, Koriat, 2012a; Koriat & Adiv, 2016) of the basis of people's convictions in the truth of their beliefs.

In philosophical theories of truth, a distinction is drawn between correspondence theories and coherence theories (Kirkham, 1992; see Hammond, 2000; Koriat, 2012b). For correspondence theories, the truth or falsity of a statement is determined by how that statement corresponds to the world. The problem with these theories, however, is that we have no knowledge about the world over and above what we know about it. As Kant (1885) noted, I can only tell whether my knowledge of the object corresponds to my knowledge of the object. Coherence theories attempted to resolve this problem by proposing that the truth of a belief is determined by its coherence with other beliefs.

SCM assumes that people's subjective confidence is based on coherence (reliability) as a proxy for correspondence (validity). When presented with a 2AFC item, people construct their response on the basis of the cues that they access at the time of making a judgment. Their choice is based on the balance of evidence in favor of the two response options (Vickers, 2001), and their confidence is based on the consistency with which the sampled cues support the chosen option (Kruglanski & Klar, 1987; Slovic, 1966).

Assuming that for each item, people draw their samples largely from the same database, confidence should correlate with the consensuality of the response—the likelihood that that response is selected across participants. This should be true independent of the accuracy of the response. The implication is that the self-consistency heuristic underlying subjective confidence succeeds in monitoring the accuracy of the chosen answer only because in the natural ecology, differences between items lie primarily in the extent to which the cues underlying the response support the correct answer. For CW items, in contrast, it is the wrong answer that is associated with higher self-consistency.

A second theory was proposed by Brewer, Sampaio, and their associates (Brewer & Sampaio, 2006, 2012; Brewer, Sampaio, & Barlow, 2005; Sampaio & Brewer, 2009) to explain their finding that the C/A relationship tends to be positive for nondeceptive items but negative for deceptive items (see Studies 12–14 in Figure 1). Their studies used several tasks involving episodic and semantic memory. Deceptive items were defined on a priori grounds as those that would be expected to yield a high proportion of errors.

The metamemory approach addresses the question how people who no longer have access to the original event that created a memory can produce judgments that are generally successful in predicting the accuracy of the memory for the original event. It was proposed that memory confidence is based on the processes and products of the just-completed memory task, along with the participants' metamemory beliefs about the relation of these processes and products to memory accuracy (see also Koriat, 2015a). Thus, when asked to indicate their confidence in the recall of a studied sentence, participants may rely on the vividness of an image that comes to mind or on the completeness of recall. For example, in a study that examined the cued-recall of studied sentences, Brewer et al. (2005) found that the occurrence of a complete recall (e.g., recalling a whole sentence) was the major factor leading to high confidence for both deceptive and nondeceptive sentences, suggesting that participants have a metamemory belief that full sentence recalls are likely to be accurate. For tasks involving recognition memory (Brewer & Sampaio, 2006, 2012; Sampaio & Brewer, 2009), introspective reports indicated that confidence is strongly related to the feeling of familiarity, the occurrence of an image, and the use of recall as a basis for a recognition decision. It was proposed that because people rely on the same output-based indicators (and associated metamemory beliefs) for both nondeceptive and deceptive items, confidence judgments are valid for nondeceptive items but invalid for deceptive items.

Finally, a third account was proposed by Roediger and DeSoto (2015) to explain the observation that in the DRM paradigm, the C/A relationship is negative for strongly related, nonstudied lures (see Studies 15 and 16 in Figure 1). They proposed that the understanding of inversed C/A relationships requires an understanding of the processes that lead to false memories. Their account of the negative C/A correlation for strongly related lures is based on Tulving's (1974) idea that remembering depends on the overlap between the memory traces that are formed after learning

[2] The study of Sampaio et al. (2017) examined confidence for face recognition. Participants were presented with exemplars of faces constructed digitally as deviations from prototype faces. When presented at test with studied exemplars, prototype faces, and nonstudied exemplars, the prototype faces yielded a high rate of false recognition. Across the studied and unstudied exemplar faces, an item-based analysis yielded significantly higher confidence for correct responses (hits and correct rejections) than for errors (misses and false alarms). In contrast, for the prototype faces, confidence was significantly higher for false alarm responses than for correct rejections.

and the cues provided in the retrieval environment during remembering. In the DRM paradigm, a strongly related lure presented at a recognition test greatly overlaps with features of the stored traces from the study list, and therefore may be falsely judged as old. The more features that overlap, the greater should be the level of false recognition and the higher the judged confidence.

Roediger and DeSoto argued that this account is consistent with an account (Roediger & DeSoto, 2014) in terms of SDT if a "strength of evidence" dimension is conceived not as trace strength but more like the cue-target match discussed in the context of Tulving and Thomson's (1973) encoding specificity principle (see Wixted & Mickes, 2010). Unlike standard signal decision models in which there is a single distribution of items for all nonstudied items, the model proposed assumes that different lures have different strengths of evidence, with strongly related lures having the greatest strength, followed by weakly related lures, and finally by unrelated lures. Assuming that the strength of evidence continuum gives rise to calling an item old and also to the confidence in the response, then a negative C/A relationship should be observed for strongly related lures.

A comparison between the three theories is beyond the scope of this article. What is important is that they all depart from a simple, direct-access approach to confidence, and that they all assume that confidence in a choice should increase with the probability of that choice. For the present study, the important implication, is that the C/A correlation is expected to be positive for typical or representative items, but may be negative across items that are less representative.

## Project 2: Using a Systematic Design to Examine the Confidence-Accuracy Correspondence

I turn next to the second methodological option for overcoming the hazards inherent in a representative design: The use of a systematic design. This design has the advantage that it can help remove the correlation that exists in real-life between accuracy and consensuality (Koriat, 2012a). In addition, it provides information about how the effects on confidence may vary across the various types of samples that people can face during their life. Often, these samples are representative, yielding OLA that is better than chance, but sometimes the items encountered can vary away from representativeness. For example, in Studies 6–10 in Project 1, no deliberate attempt was made to select items with poor OLA, but some of the items turned out to yield OLA below chance level. Thus, in Project 2, 2AFC items were selected that represent different degrees of OLA across the entire 0%-100% range.

### Method

**Participants.** One hundred twenty Hebrew-speaking University of Haifa undergraduates (81 women) participated in the experiment; 82 were paid for their participation, and 38 received course credit.

**Stimulus materials.** To allow a large enough set of CW items, five different tasks were used, and the 2AFC items for each task were selected on the basis of the results of previous experiments in an attempt to produce a balance between CC and CW items. These experiments are listed in Table 2. CW items from each experiment were used, and for each such item, a matched CC item was selected that had approximately the same average item consensus—the percentage of participants who chose the consensual answer. For example, when percent accuracy for a CW item averaged 32%, a matched CC item with close to 68% accuracy was selected. Table 2 lists the experiments from which the items were taken, the number of CC and CW items, and the mean percentage of correct responses for the selected CC and CW items in the original experiments. There were 138 items in total, 69 CC and 69 CW.

**Apparatus and procedure.** Each of the five tasks was administered in a separate block. For each item, participants chose the correct answer and indicated their confidence on a 50%–100% scale. The five tasks were administered in the order in which they are listed in Table 2. The order of the items within each task was random.

The experiment was conducted individually on a personal computer. Each trial began with a probe, which consisted of the question in the case of Tasks 1, 4, and 5, or the statement *to present the stimuli press here* (in the case of Tasks 2 and 3). After clicking *confirm*, the two alternative answers or stimuli were added, and participants indicated their choice by clicking one of them, and then a *confirm* box (participants could change their response but not after clicking *confirm*). A confidence scale (50–100) was then

Table 2
*The Tasks Used in Project 2: Order of Presentation, Task and Source of the Items, Number of Consensually-Correct (CC) and Consensually-Wrong (CW) Items Selected for the Study and Their Mean Percentage Correct, and Number of CC and CW Items Finally Selected and Their Mean Percentage Correct*

| | | | Initial selection | | | Final selection | | |
|---|---|---|---|---|---|---|---|---|
| | | | Number of CC/CW items | Mean percentage correct | | Number of CC/CW items | Mean percentage correct | |
| Order | Task | Experiment | | CC | CW | | CC | CW |
| 1 | General knowledge | Koriat (2008) | 28/28 | 65.3 | 34.2 | 24/24 | 63.4 | 35.9 |
| 2 | Judgment of line lengths | Koriat (2011, Experiment 1) | 8/8 | 74.7 | 26.0 | 7/7 | 76.4 | 24.1 |
| 3 | Judgments of the area of geometric shapes | Koriat (2011, Experiment 2) | 15/15 | 76.6 | 25.0 | 15/15 | 72.6 | 26.72 |
| 4 | Prediction of the preferences of others | Koriat (2013) | 9/9 | 70.2 | 30.3 | 88/ | 71.7 | 32.3 |
| 5 | Geography questions | Koriat (2017a) | 9/9 | 66.5 | 34.3 | 8/8 | 69.0 | 31.8 |
| ALL | | | 138 | 69.6 | 30.8 | 62/62 | 68.6 | 31.0 |

added. Participants indicated their confidence (the chances that their response was correct) by sliding a pointer on the scale using the mouse (a number in the range 50–100 corresponding to the location of the pointer on the screen appeared in a box), and then pressed *confirm*. Participants were instructed to try to make use of the full range of the confidence scale.

**Final item selection.** OLA was calculated for each item. Items were classified anew as CC and CW on the basis of the results. For Tasks 2–5, this classification conformed generally to the original classification, whereas for Task 1, this was true only for 75% of the items. Items were eliminated in an attempt to achieve an equal number of CC and CW items in each task, roughly matched in terms of item consensus (see Table 2). For the remaining items, 62 CC items and 62 CW items, OLA averaged 68.6% and 31.0%, respectively. All the analyses were based on these items.

## Results

**Calibration.** We focus first on calibration. Figure 2 plots the percentage of correct answers as a function of confidence, with confidence grouped into five categories (50–60, 61–70, 71–80, 81–90, and 91–100). Calibration is plotted separately for 4 groups of items that differ in OLA. The figure indicates several trends that are worth noting. First, a strong overconfidence bias is exhibited across all items, with confidence averaging 69.91 when percent correct is around chance (49.81). However, whereas for the CW items confidence and accuracy averaged 69.77 and 31.02, respectively, $t(61) = 18.44$, $p < .0001$, $d = 2.34$, the respective figures for the more representative CC items were 70.04 and 68.59, $t(61) = 1.08$, $p < .29$, $d = 0.14$, consistent with the claim that representative sampling can eliminate the overconfidence bias (Gigerenzer et al., 1991; Juslin, 1994).

Second, the results conform to the hard-easy effect: Overconfidence is reduced as the "difficulty" of the items decreases (Lichtenstein & Fischhoff, 1977). However, this is true even for the CC items. Among these items, those with higher OLAs tended to yield an underconfidence bias (e.g., Griffin & Tversky, 1992).

Finally, although calibration was reasonable for the representative items, the results on the whole testify to the claim that people fail to appreciate their degree of ignorance (Kruger & Dunning, 1999; Lichtenstein & Fischhoff, 1977). In particular, the results indicate that in the case of CW items, people have little awareness that they are erring (see Koriat, 2017a; Brewer & Sampaio, 2012).

**Confidence as a function of OLA.** We turn next to analyses that are pertinent to resolution. Figure 3A presents mean confidence as a function of mean OLA for items grouped into 10 categories of OLA (0–10%, 11–20% . . . 90–100). The function is clearly curvilinear, indicating that overall confidence does not increase monotonically with OLA, but increases with the deviation of OLA from 50%. The effects of OLA yielded a significant quadratic trend, $F(1, 119) = 224.94$, $p < .0001$, which accounted for 69.7% of the variance. The analysis was performed on participants' mean confidence judgments for the items in each of the 10 OLA categories. The linear trend accounted for 7.7% of the variance.

Each point in Figure 3A was based on different participants. Because individuals differ reliably in confidence judgments (see Kleitman & Stankov, 2001; Stankov & Crawford, 1996, 1997), we neutralized these differences by standardizing the confidence judgments of each participant so that the mean and standard deviation of each participant were set as those of the raw scores across all participants. Average scores were then calculated for each category. The results for the standardized scores were practically identical to those in Figure 1.

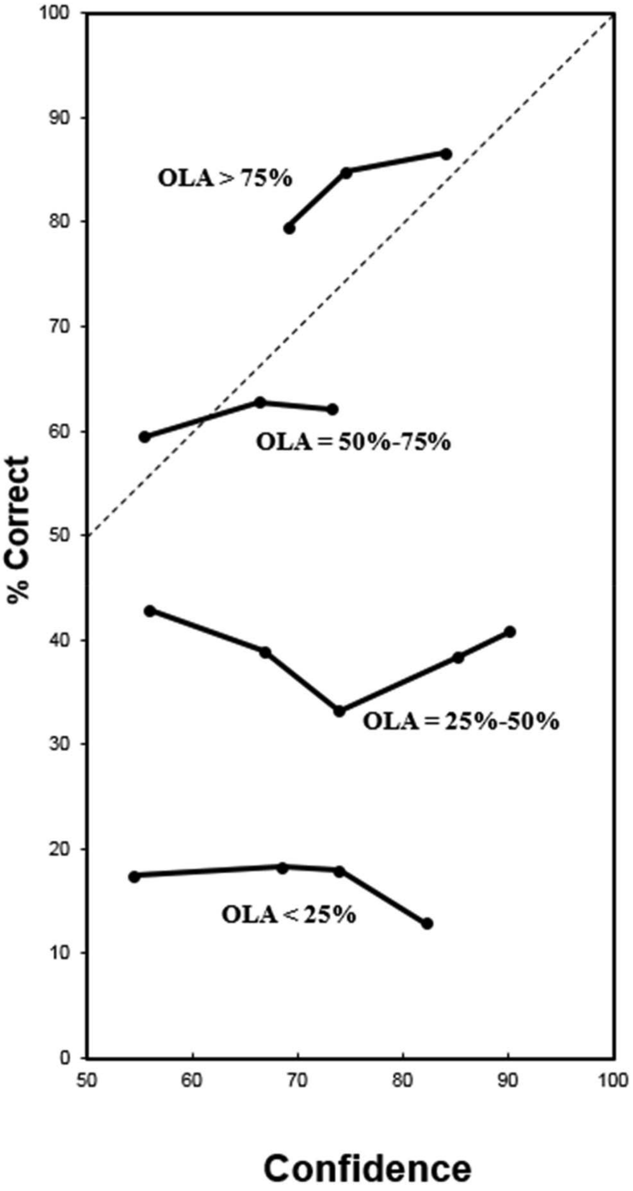

*Figure 2.* Calibration curves for 4 groups of items that differ in Object-Level Accuracy (OLA). The diagonal line indicates perfect calibration.

Figure 3B plots the same results as in Figure 3A but separately for correct and wrong responses. The two functions are largely symmetrical about the 50% chance level. Confidence in the dominant, consensual answer also increased with a deviation of OLA from 50%. The effects of OLA on confidence in the consensual response yielded a significant quadratic trend, $F(1, 119) = 185.74$, $p < .0001$, which accounted for 62.3% of the variance. The analysis was performed on participants' mean confidence judg-

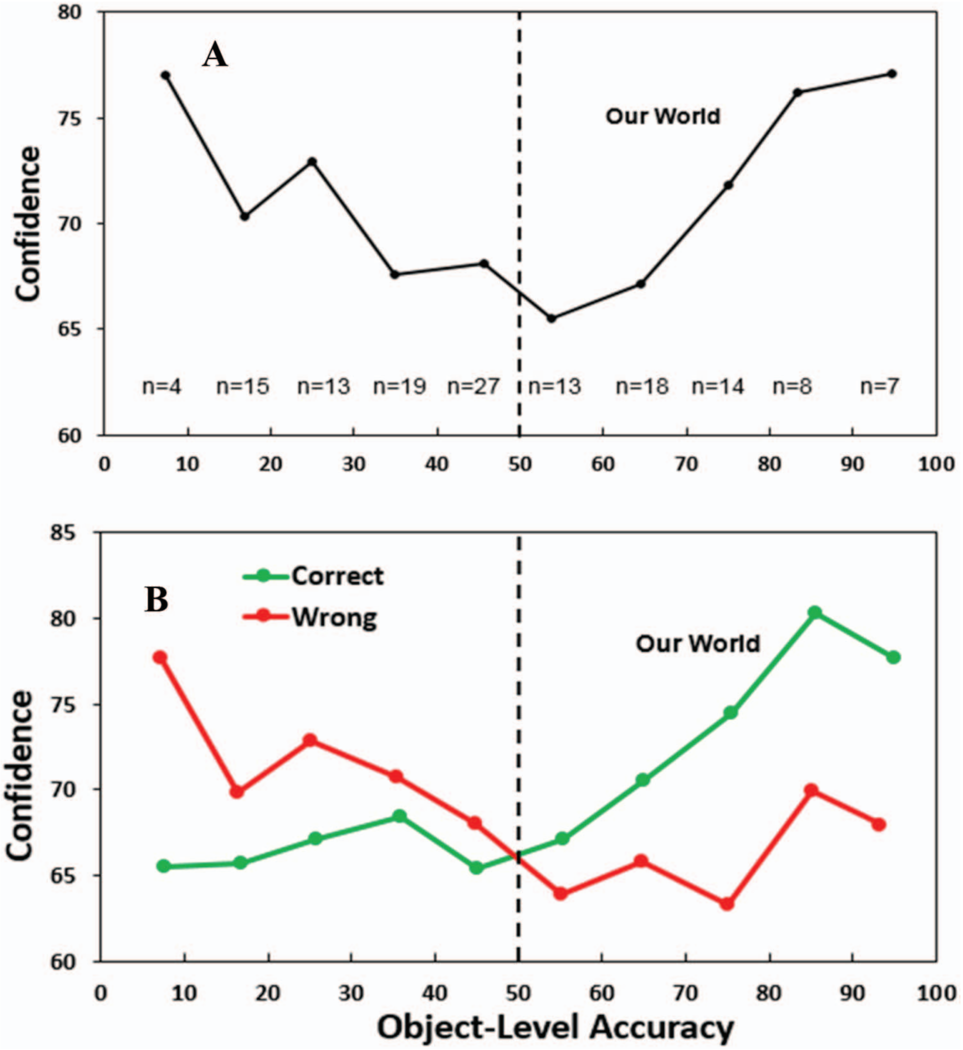

*Figure 3.* Panel A: Confidence as a function of object level accuracy. Indicated also is the number of items in each OLA category. Panel B: Confidence as a function of object level accuracy plotted separately for correct and wrong responses (Project 2).

ments for each of the 10 OLA categories using only the responses to the normative, consensual response. The linear trend accounted for 9.6% of the variance.

**Signal detection analyses.** There has been a great effort in recent years to examine metacognitive resolution within the SDT framework (Benjamin & Diaz, 2008; Fleming & Lau, 2014; Higham et al., 2009; Maniscalco & Lau, 2014). In parallel to the standard, Type-1 SDT that evaluates OLA, Type-2 SDT permits evaluation of MLA.

Figure 4 presents the Type-2 receiver operating characteristic (ROC) curves for the CC items, for the CW items, and for all items combined. For the CC items, the ROC curve lies above the diagonal; the area under the ROC (AUROC2; Fleming & Lau, 2014) averaged 0.600 across participants, $t(119) = 14.56$, $p < .0001$, for the difference from random guessing (0.5). In contrast, for the CW items, it lies *below* the diagonal, with AUROC2 averaging 0.432, $t(119) = 9.37$, $p < .0001$, for the difference from 0.5. Across all items, the type-2 ROC curve lies roughly along the diagonal, suggesting little discrimination sensitivity. AUROC2 averaged 0.516, somewhat higher than 0.5, $t(119) = 3.26$, $p < .005$.

Table 3 presents the results of statistical analyses that compared CC and CW items in AUROC2 (after Fleming & Lau, 2014), in Meta $d'$ (after Maniscalco & Lau, 2014), and in Kruskal-Goodman gamma correlation. The table indicates the significance of the difference of each measure from 0 (for Meta $d'$ and gamma) or from .500 (for AUROC2).

For comparison purposes, I applied the same analysis to a task (Koriat, 2012c, Study 2) for which the items were selected representatively (see Figure 4). AUROC2 averaged 0.674, significantly higher than for the CC items, $t(178) = 6.25$, $p < .0001$, $d = 0.97$. This result brings to the fore the constraints imposed on the selection of the CC items by the requirement to ensure matching with the CW items: The difficulty finding CW items with extreme

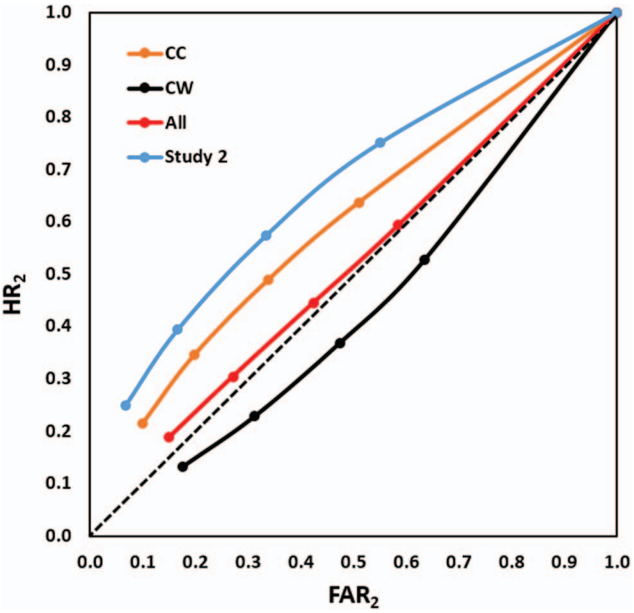

*Figure 4.* Type-2 receiver operating characteristic (ROC) curves for the Consensually- Correct (CC), Consensually-Wrong (CW), and all items combined (All). A ROC curve is also presented for Study 2 (Koriat, 2012c) for which the items were selected representatively (Project 2).

item consensus values. OLA for Study 2 of Koriat (2012c) averaged 78.9%, compared with 69.6% for the CC items in Project 2.

Altogether, the results of Project 2 accord with the consensuality principle (Koriat, 2008, 2012a). First, they yielded the same crossover interaction as in Figure 1: Across subjects, confidence for the CC items was higher for the correct answers than for the wrong answers, $t(119) = 14.88$, $p < .0001$, whereas for the CW items it was higher for the wrong answers than for the correct answers, $t(119) = 7.77$, $p < .0001$ (see Figure 5). This interactive pattern was also observed in an item-based analysis: Across the CC items, confidence was higher for participants who chose the correct answer ($M = 71.70$) than for those who chose the wrong answer ($M = 64.85$), $t(61) = 7.77$, $p < .0001$. For the CW items, in contrast, confidence was higher for participants who chose the wrong answer ($M = 70.79$) than for those who chose the correct answer ($M = 66.64$), $t(61) = 5.90$, $p < .0001$.

A second observation that is consistent with the consensuality principle is shown in Figure 6, which plots the ROC curve that is obtained when confidence judgments are assumed to actually track the consensuality of the response rather than its accuracy. In this analysis, the consensual response for each item was treated as if it were the correct response. The ROC curve across all items now lies above the diagonal, with AUROC2 averaging 0.591, $t(119) = 18.01$, $p < .0001$, for the difference from 0.5. Meta $d'$ averaged

Table 3
*Mean Auroc2, Meta d′ and Gamma Correlation for Consensually-Correct (CC), Consensually-Wrong (CW) Items, and All Items Combined, as Well t Test Comparing These Indexes for CC and CW Items*

| Measure | CC | CW | CC vs. CW | ALL |
|---|---|---|---|---|
| AUROC2 | $+.600^{***}$ | $+.432^{***}$ | $t(119) = 17.26$, $p < .0001$ | $+.516^{**}$ |
| Meta $d'$ | $+.962^{***}$ | $-.556^{***}$ | $t(119) = 13.69$, $p < .0001$ | $+.111^{*}$ |
| Gamma | $+.29^{***}$ | $-.21^{***}$ | $t(119) = 18.28$, $p < .0001$ | $+.05^{**}$ |

*Note*. Auroc2 was calculated after Fleming and Lau (2014), and meta $d'$ was calculated after Maniscalco and Lau (2014). Another measure of metacognitive discrimination is ANDI (1991). It was not used here because it does not take into account the possibility of a meaningful negative relationship between confidence and accuracy.
Significance of difference from .500 for Auroc2 and from 0 for Meta $d'$ and gamma: $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .0001$.

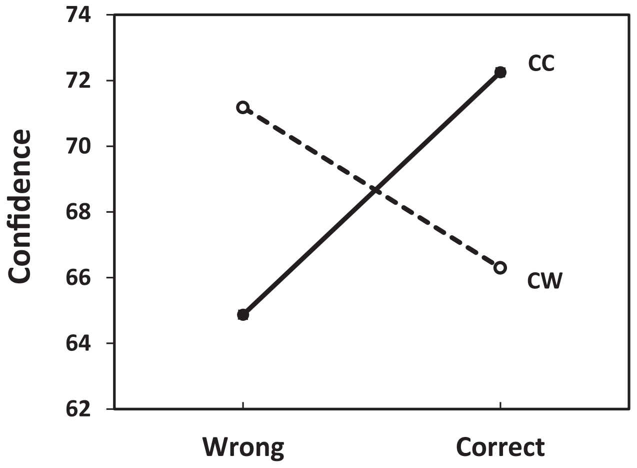

*Figure 5.* Mean confidence for correct and wrong answers, for Consensually-Correct (CC) items and Consensually-Wrong (CW) items in Project 2.

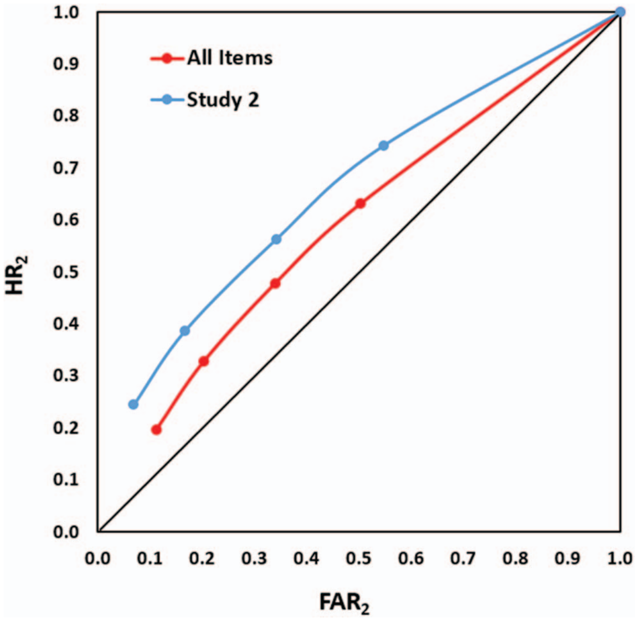

*Figure 6.* Type-2 receiver operating characteristic (ROC) curves for all items in Project 2 using consensuality rather than accuracy as the criterion. Presented also is the corresponding ROC curve for Study 2 (Koriat, 2012c) for which the items were selected representatively.

0.698, $t(119) = 17.69$, $p < .0001$, for the difference from chance level. The within-person gamma correlation between confidence and consensuality averaged +.25 across all items, significantly better than chance, $t(119) = 18.10$, $p < .0001$. It was positive for 114 of the 120 participants, $p < .0001$ by a binomial test.

Figure 6 also includes the respective ROC curve for Study 2 of Koriat (2012c). AUROC2 averaged 0.666, $t(59) = 17.44$, $p < .0001$, for the difference from 0.5. Meta $d'$ averaged 1.267, $t(59) = 16.45$, $p < .0001$, for the difference from chance level. The within-person confidence-consensuality gamma correlation averaged +.43, significantly better than chance, $t(59) = 20.42$, $p < .0001$.

## General Discussion

The question of metacognitive accuracy has been discussed in many contexts: Can people monitor the accuracy of their beliefs and judgments? The results of empirical studies indicate that people's confidence judgments are generally diagnostic of the accuracy of their reports across many domains. One domain for which a large body of research has yielded mostly mixed results is eyewitness testimony (but see Wixted et al., 2015; Wixted & Wells, 2017). However, I will exclude this domain from the present discussion because of the problem of defining what constitute "representative" samples of situations in this domain.

In what follows I will first summarize the results, and examine how they can be accounted for by SCM. Next I will delineate some of the methodological and conceptual implications of the results. Finally, I will discuss the implications of the results for the ecological perspective.

### Summary of the Findings and Their Relationship to SCM

The results on the whole clearly supported the idea that MLA is intimately tied to OLA. The results of Projects 1 (see Figure 1) confirmed the general observation that confidence judgments track the accuracy of one's judgments and beliefs across items with OLA exceeding chance performance. However, for items with OLA below chance level, confidence judgments were in fact counterdiagnostic of accuracy. This was consistently true for a variety of tasks, including word matching, almanac questions, perceptual comparisons, judgments of geographical relations, recognition memory for words and sentences, and the prediction of the majority responses of people's perceptual judgments, word association responses, preferences, beliefs, and attitudes.

The results of Project 2, which were obtained using a systematic design, also yielded the same pattern (Figure 3 and 5). The SDT analyses indicated that for items with OLA >50%, the Type-2 ROC curve lies above the diagonal, whereas for items with OLA <50% it lies below the diagonal (see Figure 4). When the results were analyzed across all items, there was only a very slight increase in accuracy with increased confidence (Figure 4 and Table 3), unlike what has been observed in many studies that examined the C/A relationship. However, the analyses across all items indicated that confidence judgments did track the consensuality of the response—the likelihood that it would be chosen by the majority of participants (see Figure 6), consistent with SCM.

As far as calibration is concerned, the results (see Figure 2) were in line with the claim of the ecological approach that representative sampling can eliminate the overconfidence bias (Gigerenzer et al., 1991; Juslin, 1994): Whereas items with OLA <50% yielded a strong overconfidence bias, those with OLA >50% yielded virtually perfect calibration. These results suggest that in the context of a systematic design, calibration and resolution tend to be correlated when calculated across different sets of items that differ in OLA. This correlation stems from MLA being moderated by OLA. Thus, the items with OLA <50% yielded a strong overconfidence bias as well as a very poor (actually negative) C/A correlation in comparison with the more representative items—those yielding OLA >50%.

The calibration results also confirm the proposition that for items that are likely to elicit erroneous answers, people are largely unaware that they are erring (Kruger & Dunning, 1999; Lichtenstein & Fischhoff, 1977). Indeed participants failed to tell whether an item is "deceptive," likely to draw mostly wrong answers across people, or whether it is nondeceptive. This is possibly because choice and confidence are based on the same process for both types of items (see Koriat, 2017a; Brewer & Sampaio, 2012).

The results on the whole are consistent with SCM. It was proposed that in the case of 2AFC items, people's choice is based on the retrieval of a small number of cues from memory, and confidence in that choice rests on the consistency with which the choice has been supported across the sampled cues. Because people with similar experience tend to draw their samples from a population of item-specific cues that is largely shared, the average self-consistency associated with each item is reflected in the interperson consensus in the choice made. Therefore, average confidence in a given choice should increase with the consensuality of that choice regardless of the accuracy of the choice. Indeed, this was found to be the case in the present study. The results suggest that the C/A correlation that has been observed in many studies stems from the fact that consensuality and accuracy are correlated across items drawn representatively from the natural environment.

Note that in this study, as well as in previous studies, the results were found to support the consensuality pattern both in a subject-based analysis as well as in an item-based analysis. For example, for items with OLA <50%, some participants did choose the correct answer to some of the items. However, they endorsed that answer with lower confidence than participant who chose the wrong answer to the same items. This observation is consistent with the sampling assumption underlying SCM, which implies some variation in the choices made across people and occasions. However, confidence is expected to track both the stable and variable contributions to the choices reached (see Koriat & Sorka, 2017).

In sum, as Deffenbacher (1980) noted, the faith in the adequacy of certainty as a predictor of accuracy "would appear to be rooted in the firm common sense intuition that accuracy and confidence are strongly and positively related" (p. 244). Our results, however, suggest that this intuition derives from the fact that in the world we live in, we witness only part of the function relating confidence to accuracy (see Figure 3). Although people may not have direct access to the accuracy of their knowledge, they rely on a heuristic that is quite effective in the world they live in.

Note, however, that several researchers who endorsed an inferential view of metacognitive judgments, also postulated the possibility that in some cases people can access directly an answer without having to engage in a probabilistic inference (e.g., Gigerenzer et al., 1991; Metcalfe, 2000; Unkelbach & Stahl, 2009). The possibility of "just knowing" seems particularly plausible in connection with episodic or semantic information that is held with strong confidence (see Koriat, 2012b). This possibility creates a problem for the compilation of CW items, because for some of these items, confidence would be expected to be particularly high for participants who happen to rely on privileged, correct knowledge (but see Prelec, Seung, & McCoy, 2017).[3] If this is true, it may explain why the ROC curve across all items (see Figure 4) was slightly above the diagonal.

## Methodological and Conceptual Implications

What are the methodological implications of the results reported in this study? A question that naturally arises is: Why should we bother about the C/A relationship obtained in a systematic design if that relationship is not true of the natural environment?

Three reasons for using a systematic design can be mentioned. First, although in the course of their life, people are more likely to face "representative" samples of items and situations, they are also likely to face samples that deviate in some way from those commonly encountered. It is therefore important to explore how people cope with different sampling situations that they might meet. For example, in Koriat's study (2011) participants who wagered money on their answer to representative, CC items, placed larger wagers on the correct answers, thus maximizing their earnings. For CW items, in contrast, they lost money by betting heavily on the wrong choices. Also group discussion was found to improve decision accuracy in the case of CC items, but was actually detrimental to accuracy for CW items (Koriat, 2015b).

A second reason involves the distinction between two research agendas that sometimes conflict (see Koriat et al., 2011). The first agenda is to obtain a faithful description of the state of affairs in the real world, for example, to determine whether the overconfidence phenomenon is real. This agenda calls for a representative design that ensures generalization to real-world conditions. The second agenda, however, is that of providing a theoretical explanation of the phenomenon under investigation. This agenda sometimes calls precisely for the use of unrepresentative items and conditions that can help untangle variables that go hand in hand in real life (Koriat, 2012a). In fact, many studies in metacognition have yielded important theoretical insights by deliberately using conditions that are ecologically unrepresentative, even contrived (Benjamin, Bjork, & Schwartz, 1998; Brewer & Sampaio, 2012; Koriat, 1995). Thus, whereas the right side of Figure 3 is relevant to the first agenda, describing what occurs in the real world, the entire figure, is the one that provides a lead to the theoretical explanation of the basis of subjective confidence and its accuracy.

A third reason, finally, is that the appropriate reference class for drawing a representative sample can differ depending on the research question asked. For example, consider a philosopher who wishes to know whether people are endowed with a general ability to tell truth from falsity independent of the structure of a specific ecology. For her, our results might be taken to imply that samples of items drawn from the natural ecology are "biased," yielding misleading conclusions about people's general discrimination ability. In fact, philosophers who are concerned with "universal" truths rather than with regularities that are specific to the accidental properties of a particular ecology, may take our results to suggest that their perplexity about truth and belief justification is not unwarranted.

In general, this study delivers a warning: Beware of a representative design! (see Fiedler, 2000, for a similar warning). Reality is "out of focus" with respect to the issue of monitoring resolution, and perhaps with respect to other issues as well. This should not be surprising given the adaptation of people to their natural ecology through evolution and learning. It is quite intriguing that although many studies have reported a positive C/A relationships for many tasks, never have I seen an acknowledgment of the possibility that these relationships might be completely attributable to OLA being better than chance for the tasks used. One exception, perhaps, is Deffenbacher's (1980) optimality hypothesis for eyewitness testimony, which states that the likelihood of obtaining positive correlations between eyewitness confidence and accuracy should increase with the optimality of the information-processing conditions during encoding, retention and testing. Deffenbacher's hypothesis, however, concerns between-individual rather than within-individual C/A correlations.

In fact, in many attempts to model choice and decision behavior, researchers have relied heavily on confidence judgments. However, they failed to consider the possibility that the results might be specific to the tasks used (e.g., psychophysical judgments, recognition memory), for which OLA is considerably better than chance.

## Implications of the Results for the Ecological Perspective

Whereas the foregoing discussion underscored the perils lurking in a representative design, the results can actually be seen to provide strong support for the basic tenets of the ecological approach, which has emphasized the adaptation of organisms to the probabilistic structure of the natural environment (see Fiedler & Juslin, 2006).

Fiedler (2007), who stressed the need to consider the texture and contents of the stimulus environment that impinges on the individual's mind, described several pervasive biases in the information ecology. Here we focused on one pervasive constraint: People live in a world in which their type-1 judgments are considerably better than chance. Therefore, they are exposed to biased samples of information for which items and situations differ mostly in the extent to which they yield better than chance judgments. Note that in discussing the adaptation of organisms to their ecology, Brunswik (1956) emphasized the contribution of learning much more than that of evolution. However, evolution has undoubtedly also contributed to OLA being considerably better than chance for many basic processes such as sensory discrimination and memory performance.

The results of the present study are consistent with the ecological perspective in suggesting that in the course of their adaptation to the natural environment, organisms not only learn the ecological validity of different cues, but the very heuristics that they use are

[3] One of the CW questions in Koriat (2008) was whether the capital of Australia is Sydney or Canberra. One of the participants approached me after the experiment and told me that he knew the correct answer because he had lived most of his life in Canberra before moving to Israel. This might be an example of reliance on "privileged knowledge."

specifically tailored to the probabilistic structure of the environment. What is impressive is that people rely routinely on the self-consistency heuristic despite the fact that this heuristic is *counterdiagnostic* of accuracy for items that are unrepresentative of the natural ecology.

This pattern of results provides strong support for Simon's (1956, 1982) notion of bounded rationality and for the theoretical framework of Gigerenzer and his associates on fast and frugal heuristics (Gigerenzer & Goldstein, 1996; Gigerenzer, Hertwig, & Pachur, 2011; Gigerenzer, Todd, & the ABC Research Group, 1999; Hertwig, Herzog, Schooler, & Reimer, 2008). According to Simon, given the constraints of limited knowledge, time, and computational capabilities, information-processing systems satisfice rather than optimize. People exploit regularities in the world that allow them to rely on simplifying mechanisms. The notion of bounded rationality implies domain specificity (see Fiedler, 2007): People do not strive for general algorithms that provide optimal solutions under all conditions, but make do with satisficing heuristics that yield reasonable solutions that fit the architecture of a particular environment.

The self-consistency heuristic can be labeled a "bounded" heuristic, one whose effectiveness is confined to the probabilistic structure of a particular ecology. Although this heuristic is liable to yield illusions of knowing (Koriat, 1998) and metacognitive myopia (Fiedler, 2000, 2012) for a few unrepresentative items, it has the advantage of being fast and frugal, and of producing metacognitive judgments that are accurate for most items in the natural environment. Perhaps the best evidence for the overall usefulness of this heuristic is the failure of researchers to recognize that the positive C/A relationship that has been observed across many studies is actually confined to items for which OLA is better than chance.

Another bounded heuristic is the accessibility heuristic assumed to underlie FOK judgments. Koriat (1993) argued that when the retrieval of a memory target fails, FOK is based on the mere accessibility of partial information about the target regardless of its accuracy (see also Brewer & Sampaio, 2012). Indeed, both correct partial information and wrong partial information were found to contribute equally to the FOK. However, FOK judgments were nevertheless accurate in predicting the future recognition of the elusive memory target. This is because the partial cues retrieved about the elusive target were much more likely to be correct than wrong. Thus, FOK judgments are accurate because memory itself is largely accurate. In support of this idea, the FOK-recognition relationship was found to be positive only across typical ("representative") memory questions that tend to elicit primarily correct partial information, whereas for questions that are assumed to elicit a preponderance of incorrect partial information, correct recognition decreased with increased FOK judgments. This pattern mimics the pattern that was found for the C/A correlation in the present study. Perhaps other heuristics, such as the processing fluency heuristic are similarly bounded (see Schwarz, 2004; Unkelbach, 2006).

In sum, the results suggest that people do not have privileged knowledge about the accuracy of their knowledge. Nevertheless, they succeed in discriminating between correct and wrong beliefs and judgments across real-life items in many domain. Their success derives from the application of a frugal heuristic that exploits the statistical structure of the environment. Although the application of that heuristic across the board yields inflated metacognitive judgments for a few so-called "misleading" or "deceptive" items, this is a small price to pay given the simplicity of the heuristic and its general adaptive value.

## References

Allwood, C. M., & Montgomery, H. (1987). Response selection strategies and realism of confidence judgments. *Organizational Behavior and Human Decision Processes, 39,* 365–383. http://dx.doi.org/10.1016/0749-5978(87)90029-X

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39,* 133–144. http://dx.doi.org/10.1016/0749-5978(87)90049-5

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Mahwah, NJ: Erlbaum.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68. http://dx.doi.org/10.1037/0096-3445.127.1.55

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

Björkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes, 58,* 386–405. http://dx.doi.org/10.1006/obhd.1994.1043

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

Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language, 67,* 59–77. http://dx.doi.org/10.1016/j.jml.2012.04.002

Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language, 52,* 618–627. http://dx.doi.org/10.1016/j.jml.2005.01.017

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32,* 513–531. http://dx.doi.org/10.1037/0003-066X.32.7.513

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning & Verbal Behavior, 5,* 325–337. http://dx.doi.org/10.1016/S0022-5371(66)80040-3

Brunswik, E. (1944). Distal focussing of perception: Size-constancy in a representative sample of situations. *Psychological Monographs, 56,* 1–28. http://dx.doi.org/10.1037/h0093505

Brunswik, E. (1955a). In defense of probabilistic functionalism: A reply. *Psychological Review, 62,* 236–242. http://dx.doi.org/10.1037/h0040198

Brunswik, E. (1955b). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217. http://dx.doi.org/10.1037/h0047470

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

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

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199596195.001.0001

Cohen, R. L., Sandler, S. P., & Keglevich, L. (1991). The failure of memory monitoring in a free recall task. *Canadian Journal of Psychology/Revue canadienne de psychologie, 45,* 523–538. http://dx.doi.org/10.1037/h0084303

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior, 4,* 243–260. http://dx.doi.org/10.1007/BF01040617

DeSoto, K. A., & Roediger, H. L., III. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science, 25,* 781–788. http://dx.doi.org/10.1177/0956797613516149

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

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

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5,* 69–106. http://dx.doi.org/10.1111/j.1529-1006.2004.00018.x

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519–527. http://dx.doi.org/10.1037/0033-295X.101.3.519

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107,* 659–676. http://dx.doi.org/10.1037/0033-295X.107.4.659

Fiedler, K. (2007). Information ecology and the explanation of social cognition and behavior. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 176–200). New York, NY: Guilford Press.

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 57, pp. 1–55). San Diego, CA: Elsevier. http://dx.doi.org/10.1016/B978-0-12-394293-7.00001-7

Fiedler, K., & Juslin, P. (2006). Taking the interface between mind and environment seriously. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 3–29). New York, NY: Cambridge University Press.

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. http://dx.doi.org/10.1037/0096-1523.3.4.552

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8,* 443. http://dx.doi.org/10.3389/fnhum.2014.00443

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Gigerenzer, G. (2004). The irrationality paradox. *Behavioral and Brain Sciences, 27,* 336–338. http://dx.doi.org/10.1017/S0140525X04310083

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669. http://dx.doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199744282.001.0001

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

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.

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. http://dx.doi.org/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.

Green, D. M., & Swets, J. A. (Eds.). (1966) *Detection theory and psychophysics*. New York, NY: Wiley.

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. 177–199). Malden, MA: Blackwell. http://dx.doi.org/10.1002/9780470752937.ch9

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

Hammond, K. R. (2000). Coherence and correspondence theories in judgment and decision making. In T. Connolly & H. R. Arkes (Eds.), *Judgment and decision making: An interdisciplinary reader* (2nd ed., pp. 53–65). New York, NY: Cambridge University Press.

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

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. http://dx.doi.org/10.1037/a0013025

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 57–80. http://dx.doi.org/10.1037/a0013865

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

Hoffrage, U., & Hertwig, R. (2006). Which world should be represented in representative design? In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 381–408). Cambridge, UK: Cambridge University Press.

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: Capturing decision tendencies in a fictitious medical test. *Metacognition and Learning, 9,* 25–49. http://dx.doi.org/10.1007/s11409-013-9110-y

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. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology, 5,* 55–71. http://dx.doi.org/10.1080/09541449308406514

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. http://dx.doi.org/10.1006/obhd.1994.1013

Juslin, P., & Montgomery, H. (Eds.). (2007) *Judgment and decision making: Neo-Brunswikian and process-tracing approaches*. New York, NY: Psychology Press.

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. http://dx.doi.org/10.1037/0033-295X.104.2.344

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107,* 384–396. http://dx.doi.org/10.1037/0033-295X.107.2.384

Kant, I. (1885). *Kant's introduction to logic: And his essay on the mistaken subtilty of the four figures* (T. K. Abbott, Trans.). London, UK: Longmans, Green.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138,* 469–486. http://dx.doi.org/10.1037/a0017341

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. http://dx.doi.org/10.1006/jmla.1993.1001

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language, 48,* 704–721. http://dx.doi.org/10.1016/S0749-596X(02)00504-1

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273. http://dx.doi.org/10.1016/0001-6918(91)90036-Y

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

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

Koriat, A. (1975). Phonetic symbolism and feeling of knowing. *Memory & Cognition, 3,* 545–548. http://dx.doi.org/10.3758/BF03197529

Koriat, A. (1976). Another look at the relationship between phonetic symbolism and the feeling of knowing. *Memory & Cognition, 4,* 244–248. http://dx.doi.org/10.3758/BF03213170

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639. http://dx.doi.org/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. http://dx.doi.org/10.1037/0096-3445.124.3.311

Koriat, A. (1998). Illusions of knowing: The link between knowledge and metaknowledge. In V. Y. Yzerbyt, G. Lories, & B. Dardenne (Eds.), *Metacognition: Cognitive and social dimensions* (pp. 16–34). London, England: Sage. http://dx.doi.org/10.4135/9781446279212.n2

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289–326). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816789.012

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

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General, 140,* 117–139. http://dx.doi.org/10.1037/a0022171

Koriat, A. (2012a). The self-consistency model of subjective confidence. *Psychological Review, 119,* 80–113. http://dx.doi.org/10.1037/a0025648

Koriat, A. (2012b). The subjective confidence in one's knowledge and judgments: Some metatheoretical considerations. In M. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *The foundations of metacognition* (pp. 213–233). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199646739.003.0014

Koriat, A. (2012c). When are two heads better than one and why? *Science, 336,* 360–362. http://dx.doi.org/10.1126/science.1216549

Koriat, A. (2013). Confidence in personal preferences. *Journal of Behavioral Decision Making, 26,* 247–259. http://dx.doi.org/10.1002/bdm.1758

Koriat, A. (2015a). Knowing by doing: When metacognitive monitoring follows metacognitive control. In S. D. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. Roediger, III, (Eds.), *Remembering: Attributions, processes, and control in human memory: Papers in honor of Larry L. Jacoby* (pp. 185–197). New York, NY: Psychology Press.

Koriat, A. (2015b). When two heads are better than one and when they can be worse: The amplification hypothesis. *Journal of Experimental Psychology: General, 144,* 934–950. http://dx.doi.org/10.1037/xge0000092

Koriat, A. (2016). Metacognition: Decision-making processes in self-monitoring and self-regulation. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. 1, pp. 356–379). Malden, MA: Wiley–Blackwell.

Koriat, A. (2017a). Can people identify "deceptive" or "misleading" items that tend to produce mostly wrong answers? *Journal of Behavioral Decision Making, 30,* 1066–1077. http://dx.doi.org/10.1002/bdm.2024

Koriat, A. (2017b). [Confidence in the predictions of others' responses in a word-association task]. Unpublished raw data.

Koriat, A., & Adiv, S. (2014). *Confidence in the predictions of others' beliefs and attitudes*. Manuscript in preparation.

Koriat, A., & Adiv, S. (2016). The self-consistency theory of subjective confidence. In J. Dunlosky & S. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 127–147). New York, NY: Oxford University Press.

Koriat, A., Adiv, S., & Schwarz, N. (2016). Views that are shared with others are expressed with greater confidence and greater fluency independent of any social influence. *Personality and Social Psychology Review, 20,* 176–193. http://dx.doi.org/10.1177/1088868315585269

Koriat, A., & Goldsmith, M. (1996a). Memory metaphors and the real life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences, 19,* 167–188. http://dx.doi.org/10.1017/S0140525X00042114

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

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51,* 481–537. http://dx.doi.org/10.1146/annurev.psych.51.1.481

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 34–53. http://dx.doi.org/10.1037/0278-7393.27.1.34

Koriat, A., Levy-Sadot, R., Edry, E., & de Marcas, S. (2003). What do we know about what we cannot remember? Accessing the semantic attributes of words that cannot be recalled. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1095–1105. http://dx.doi.org/10.1037/0278-7393.29.6.1095

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1037/0096-3445.135.1.36

Koriat, A., Pansky, A., & Goldsmith, M. (2011). 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. 297–328). London, UK: Psychology Press.

Koriat, A., & Sorka, H. (2017). The construction of category membership judgments: Towards a distributed model. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed., pp. 773–794). Amsterdam, the Netherlands: Elsevier. http://dx.doi.org/10.1016/B978-0-08-101107-2.00031-2

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121–1134. http://dx.doi.org/10.1037/0022-3514.77.6.1121

Kruglanski, A. W., & Klar, Y. (1987). A view from a bridge: Synthesizing the consistency and attribution paradigms from a lay epistemic perspective. *European Journal of Social Psychology, 17,* 211–241. http://dx.doi.org/10.1002/ejsp.2420170208

Kurdi, B., Diaz, A. J., Wilmuth, C. A., Friedman, M. C., & Banaji, M. R. (2016, May). *Moderators of the confidence–accuracy relationship in recognition memory*. Poster presented at the Annual Meeting of the Association for Psychological Science, Chicago, IL.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior & Human Performance, 20,* 159–183. http://dx.doi.org/10.1016/0030-5073(77)90001-0

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. http://dx.doi.org/10.1017/CBO9780511809477.023

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of Type 1 and Type 2 data: Meta-d′, response-specific meta-d′, and the unequal variance SDT model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Berlin, Germany: Springer. http://dx.doi.org/10.1007/978-3-642-45190-4_3

Metcalfe, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics? *Personality and Social Psychology Review, 2,* 100–110. http://dx.doi.org/10.1207/s15327957pspr0202_3

Metcalfe, J. (2000). Feelings and judgments of knowing: Is there a special noetic state? *Consciousness and Cognition, 9,* 178–186. http://dx.doi.org/10.1006/ccog.2000.0451

Metcalfe, J., & Dunlosky, J. (2008). Metamemory. In H. Roediger (Ed.), *Cognitive psychology of memory* (Vol. 2, pp. 349–362). Oxford, UK: Elsevier.

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140,* 239–257. http://dx.doi.org/10.1037/a0023007

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

Neisser, U. (1978). Memory: What are the important questions? In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 3–24). London, UK: Academic Press.

Neisser, U. (1985). The role of theory in the ecological study of memory: Comment on Bruce. *Journal of Experimental Psychology: General, 114,* 272–276. http://dx.doi.org/10.1037/0096-3445.114.2.272

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

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2,* 267–271. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00147.x

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* (Vol. 26, pp. 125–173). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/S0079-7421(08)60053-5

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220. http://dx.doi.org/10.1037/1089-2680.2.2.175

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature, 541,* 532–535. http://dx.doi.org/10.1038/nature21054

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199602162.001.0001

Roediger, H. L., & DeSoto, K. A. (2015). Understanding the relation between confidence and accuracy in reports from memory. In S. D. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. Roediger, III, (Eds.), *Remembering: Attributions, processes, and control in human memory: Papers in honor of Larry L. Jacoby* (pp. 347–367). New York, NY: Psychology Press.

Roediger, H. L., III, & DeSoto, K. A. (2014). Confidence and memory: Assessing positive and negative correlations. *Memory, 22,* 76–91. http://dx.doi.org/10.1080/09658211.2013.795974

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814. http://dx.doi.org/10.1037/0278-7393.21.4.803

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

Sampaio, C., Reinke, V., Mathews, J., Swart, A., & Wallinger, S. (2017). High confidence in falsely recognizing prototypical faces. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology,* 1–33. http://dx.doi.org/10.1080/17470218.2017.1329844

Schraw, G., & Nietfeld, J. (1998). A further test of the general monitoring skill hypothesis. *Journal of Educational Psychology, 90,* 236–248. http://dx.doi.org/10.1037/0022-0663.90.2.236

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

Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1074–1083. http://dx.doi.org/10.1037/0278-7393.18.5.1074

Schwarz, N. (2004). Metacognitive experience in consumer judgment and decision making. *Journal of Consumer Psychology, 14,* 332–348. http://dx.doi.org/10.1207/s15327663jcp1404_2

Schwarz, N. (2015). Metacognition. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA handbook of personality and social psychology: Attitudes and social cognition* (pp. 203–229). Washington, DC: APA. http://dx.doi.org/10.1037/14341-006

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 754–770. http://dx.doi.org/10.1037/0278-7393.24.3.754

Sheffer, L. (2003). *The reliability and structure of metacognitive skills and their relationship to cognitive performance* (Unpublished doctorate dissertation). University of Haifa, Haifa, Israel.

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

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63,* 129–138. http://dx.doi.org/10.1037/h0042769

Simon, H. A. (1982). *Models of bounded rationality*. Cambridge, MA: MIT Press.

Slobin, D. I. (1968). Antonymic phonetic symbolism in three natural languages. *Journal of Personality and Social Psychology, 10,* 301–305. http://dx.doi.org/10.1037/h0026575

Slovic, P. (1966). Cue-consistency and cue-utilization in judgment. *The American Journal of Psychology, 79,* 427–434. http://dx.doi.org/10.2307/1420883

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior* *and Human Decision Processes, 65,* 117–137. http://dx.doi.org/10.1006/obhd.1996.0011

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences, 21,* 971–986. http://dx.doi.org/10.1016/S0191-8869(96)00130-4

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

Sternberg, R. J. (1998). Metacognition, abilities, and developing expertise: What makes an expert student? *Instructional Science, 26,* 127–140. http://dx.doi.org/10.1023/A:1003096215103

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

Thompson, W. B., & Mason, S. E. (1996). Instability of individual differences in the association between confidence judgments and memory performance. *Memory & Cognition, 24,* 226–234. http://dx.doi.org/10.3758/BF03200883

Tulving, E. (1974). Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *American Scientist, 62,* 74–82.

Tulving, E., & Madigan, S. A. (1970). Memory and verbal learning. *Annual Review of Psychology, 21,* 437–484. http://dx.doi.org/10.1146/annurev.ps.21.020170.002253

Tulving, E., & Thomson, D. M. (1971). Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology, 87,* 116–124. http://dx.doi.org/10.1037/h0030186

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373. http://dx.doi.org/10.1037/h0020071

Unkelbach, C. (2006). The learned interpretation of cognitive fluency. *Psychological Science, 17,* 339–345. http://dx.doi.org/10.1111/j.1467-9280.2006.01708.x

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition, 18,* 22–38. http://dx.doi.org/10.1016/j.concog.2008.09.006

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

Vickers, D. (2001). Where does the balance of evidence lie with respect to confidence? In E. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), *Proceedings of the seventeenth annual meeting of the international society for psychophysics* (pp. 148–153). Lengerich, Germany: Pabst Science.

Winman, A. (1997). The importance of item selection in “Knew-It-All-Along” studies of general knowledge. *Scandinavian Journal of Psychology, 38,* 63–72. http://dx.doi.org/10.1111/1467-9450.00010

Winman, A., Juslin, P., & Björkman, M. (1998). The confidence–hindsight mirror effect in judgment: An accuracy-assessment model for the knew-it-all-along phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 415–431. http://dx.doi.org/10.1037/0278-7393.24.2.415

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025–1054. http://dx.doi.org/10.1037/a0020874

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L., III. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70,* 515–526. http://dx.doi.org/10.1037/a0039510

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18,* 10–65. http://dx.doi.org/10.1177/1529100616686966

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

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

Received April 19, 2017
Revision received September 17, 2017
Accepted November 19, 2017 ■