Journal of Behavioral Decision Making, *J. Behav. Dec. Making*, **30**: 1066–1077 (2017)
Published online 4 July 2017 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/bdm.2024

Can People Identify "Deceptive" or "Misleading" Items that Tend to Produce Mostly Wrong Answers?

ASHER KORIAT*
University of Haifa, Haifa, Israel

ABSTRACT

In many domains, two-alternative forced-choice questions produce more correct responses than wrong responses across participants. However, some items, dubbed "deceptive" or "misleading", produce mostly wrong answers. These items yield poor calibration and poor resolution because the dominant, erroneous response tends to be endorsed with great confidence, even greater than that of the correct response. In addition, for deceptive items, group discussion amplifies rather than mitigates error while enhancing confidence in the erroneous response. Can participants identify deceptive items when they are warned about their existence? It is argued that people's ability to discriminate between deceptive and non-deceptive items is poor when the erroneous responses are based on the same process assumed to underlie correct responses. Indeed, participants failed to discriminate between deceptive and non-deceptive perceptual items when they were warned that some of the items (Experiment 1) or exactly half of the items (Experiment 2) were deceptive. A similar failure was observed for general-knowledge questions (Experiment 3) except when participants were informed about the correct answer (Experiment 4). Possibly, for these tasks, people cannot escape the dangers lurking in deceptive items. In contrast, the results suggest that participants can identify deceptive problems for which the wrong answer stems from reliance on a fast, intuitive process that differs from the analytic mode that is likely to yield correct answers (Experiment 5). The practical and theoretical implications of the results were discussed. Copyright © 2017 John Wiley & Sons, Ltd.

KEY WORDS metacognition; deceptive items; errors; confidence; dual-process theory

A great deal of research on subjective confidence has concerned the correspondence between confidence judgments and actual performance (see Dunlosky & Metcalfe, 2009; Koriat, 2016). Two aspects of correspondence have been studied, calibration and resolution. Calibration (or bias) refers to the absolute discrepancy between mean confidence and mean accuracy, and reflects the extent to which confidence judgments are realistic or exhibit an overconfidence bias or an underconfidence bias (Griffin & Brenner, 2004; Lichtenstein, Fischhoff, & Phillips, 1982). Resolution, in turn, refers to the within-person confidence–accuracy (C/A) correlation -- the extent to which confidence judgments discriminate between correct and incorrect answers.

The impetus for the study of calibration derived in part from observations suggesting an overconfidence bias for almanac questions. Specifically, for two-alternative forced-choice (2AFC) questions, mean reported probability typically exceeds the proportion of correct answers (Lichtenstein et al., 1982; McClelland & Bolger, 1994). Several theories have been proposed to account for this bias (Dunning, Heath, & Suls, 2004; Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980; Metcalfe, 1998). However, proponents of the ecological approach to judgments and decisions have argued that the overconfidence bias actually derives from researchers' tendency to oversample almanac items that are difficult or misleading, for which participants tend to choose the wrong answer (Björkman, 1994; Hoffrage & Hertwig, 2006; Juslin, 1994; Juslin, Winman, & Olsson, 2000). Indeed, several studies in which items were sampled randomly from their domains yielded little evidence for overconfidence (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994; Juslin et al., 2000). For example, Juslin (1994) asked participants ("selectors") to select items that provide a test of subjects' knowledge. The selected set was found to yield a strong overconfidence bias, whereas no overconfidence bias was observed when the items were selected randomly from their reference class.

These results suggest that the overconfidence phenomenon is tied to the properties of the items used, and that test-makers tend to construct tests that tax subjects' knowledge, oversampling items for which the wrong answer is alluring. The implication is that the selection of items by experimenters embodies access (perhaps unconscious) to the characteristics of items that tend to elicit erroneous answers. A question of interest is whether test-takers can also access these properties, identifying tricky items that are liable to contribute to overconfidence. The results so far suggest that test takers do not do that spontaneously. However, can they do so when they are alerted to the existence of such items in a list? If they can do so when they are put on guard, they may be able to avoid some of the deplorable effects of overconfidence, such as the overestimation of their knowledge, and the exaggeration of their ability to control events (see Arkes, Christensen, Lai, & Blumer, 1987; Dunning et al., 2004).

We turn next to resolution, which has received greater attention by metacognition researchers (see Dunlosky & Metcalfe, 2009; Koriat, 2007). Resolution and calibration are independent aspects of metacognitive correspondence (Fleming & Lau, 2014; Koriat & Goldsmith, 1996). However, like calibration, resolution also depends heavily on the nature of the items over which the C/A correlation is calculated. In fact, although people endorse correct answers

*Correspondence to: Asher Koriat, Department of Psychology, University of Haifa, Haifa 3498838, Israel. E-mail: akoriat@research.haifa.ac.il

with higher confidence than wrong answers, conditions have been reported for which the C/A correlation is *negative*: People are more confident when they are wrong.

To review briefly these results and their behavioral implications, we shall adopt the operational distinction of Koriat (1976, 1995, 2012) between consensually-correct (CC) items that yield a majority of correct answers across participants, and consensually-wrong (CW) items that yield a preponderance of wrong answers. In several studies, 2AFC items were divided ad-hoc, on the basis of the empirical results, into the CC and CW categories. It was found that whereas for CC items confidence was higher for correct answers than for wrong answers, for the CW items it was actually higher for the wrong answers than for the correct answer. This pattern was observed for a word-matching task (Koriat, 1976), for general-knowledge (Koriat, 2008), and for perceptual judgments (Koriat, 2011). A similar pattern was found by Brewer, Sampaio and their associates (Brewer & Sampaio, 2006, 2012; Brewer, Sampaio, & Barlow, 2005; Sampaio & Brewer, 2009) in comparing non-deceptive items with deceptive items that yield a high proportion of errors. For the deceptive items, the erroneous responses were associated with higher confidence than the correct responses. A similar pattern was reported by Roediger and DeSoto (2014) for an old-new recognition memory task: The C/A correlation was positive across studied words, but negative across lures that were strongly related to some of the studied words.

Resolution is important for tasks that require a choice between alternative courses of action. Because people rely heavily, even blindly, on their confidence in guiding their behavior (Goldsmith & Koriat, 2008), a situation in which confidence is counter-diagnostic of accuracy can result in preference for the wrong choices. Two studies illustrate this idea. In Koriat's (2011) study, participants who wagered money on the correctness of their answer, placed larger wagers on the correct answers for CC items, maximizing their cash earnings. For CW items, in contrast, they lost money by betting heavily on the wrong choices (see also Fischhoff, Slovic, & Lichtenstein, 1977). The results suggested that reliance on confidence judgments in making a choice can sometimes be detrimental.

The second study concerned the benefit of group-based decisions. Previous research has indicated that groups perform better than independent individuals on many tasks (e.g., Hill, 1982; Sunstein & Hastie, 2015), but conditions have been documented in which group decisions sometimes go astray (Baron, 2005; Janis, 1982). Koriat (2015) compared individual and group decisions using 2AFC items that were divided into CC and CW categories. For both a perceptual task and a general-knowledge task, group decisions were more accurate than individual decisions for the CC items, whereas for the CW items dyadic interaction actually yielded even less accurate decisions than the decisions made individually. These results naturally raise the question whether people can avoid some of the perils lurking in group decisions when they are warned about the existence of deceptive items and issues. Such identification can at least save the cost involved in group deliberation, particularly if that deliberation can only be detrimental.

The foregoing review indicates that both calibration and resolution vary strongly with characteristics of the items included in a test. In many domains, some of the items tend to draw a high proportion of erroneous responses for a variety of reasons (see e.g., Brewer & Sampaio, 2012; Fischhoff et al., 1977; Kelley & Lindsay, 1993; Roediger & McDermott, 1995). These items, which have been referred to as "deceptive," "misleading," or "tricky", tend to yield erroneous responses that are endorsed with high confidence, thus impairing both overall calibration and overall resolution. The question addressed in this study is whether participants can identify such detective items when they are warned about their existence. As noted earlier, the claim that experimenters tend to oversample deceptive items implies that they can somehow identify these items. Can test-takers likewise do so?

We propose that the answer to this question depends on whether the psychological processes that lead to erroneous responses are the same or different than those underlying correct responses. Consider the studies carried out within the dual-process theoretical framework. In this framework, a distinction is posited between two modes of processing, referred to as System 1 and System 2 (Stanovich & West, 2000; see Kahneman, 2011). The former is assumed to be intuitive, heuristic, fast and effortless, whereas the latter is analytical, deliberate, slow, and effortful. A variety of errors and illusions have been assumed to stem from reliance on the fast, intuitive mode of System 1, and it has been proposed that manipulations that induce a slow, analytic processing may help mitigate these errors. Indeed, several studies indicated that participants can be led to detect "tricky" questions and avoid errors. For example, when asked, "How many animals of each kind did Moses take on the Ark?" most people respond "two". However, participants were more likely to notice the distortion when the question was printed in a difficult-to-read font (Song & Schwarz, 2008). Other studies also suggest that disfluency experience induces closer scrutiny and activates analytic forms of reasoning that can correct the output of intuitive forms of reasoning (Alter, 2013; Alter, Oppenheimer, Epley, & Eyre, 2007; but see Meyer et al., 2015; Thompson et al., 2013).

These and other studies assume that in attempting to solve a problem that involves a conflict between System 1 and System 2 processes, people initially produce a fast but incorrect response (System 1), which may then be overridden by a deliberative process (System 2) that often yields the correct response. However, De Neys (2014, see also, Bago and De Neys, 2017) proposed that the ability to detect a conflict between System 1 and System 2 responses implies that both heuristic intuitions and logical intuitions are activated from the start. Thus, when faced with conflict problems, the generation of both a logical and a heuristic intuition allows people to detect conflict at a subconscious level, sometimes motivating a shift to System-2 processing (see also Pennycook, Fugelsang, & Koehler, 2015; Thompson & Morsanyi, 2012).

The idea that people are sensitive to the conflict between heuristic and analytic responses has received support in a

DOI: 10.1002/bdm

large number of studies on conflict detection. These studies compared conflict problems in which an intuitive response conflicts with the correct logical response, with control problems that do not involve such conflict. The results suggested that even participants who failed to give the correct response demonstrated longer response latencies and lower confidence for the conflict problems than for the control problems (Bonner & Newell, 2010; De Neys, 2012; De Neys, Cromheeke, & Osman, 2011; De Neys & Glumicic, 2008; Mevel et al., 2015).

What are the implications of the conflict detection studies for the ability of people to identify deceptive items? The answer depends on whether the response to the items that have been referred to as "misleading" or "deceptive" derives from a process such as that associated with the System 1 mode of thinking. In fact, several authors have claimed that the erroneous response to so-called misleading or deceptive questions depends on the same process that generally leads to the correct answer. For example, it has been proposed that the choice of an answer to general-information questions is based on inference from cues that are retrieved from memory. These cues generally lead to the correct answer by virtue of people's general adaptation to the real world (see Dhami, Hertwig, & Hoffrage, 2004; Hoffrage & Hertwig, 2006). However, because the validity of these cues is limited, they may sometimes support the wrong answer (Gigerenzer et al., 1991; Koriat, 2012). Juslin (1994), for example, noted that "the existence of so called "misleading items," i.e., items for which most subjects select the wrong answer does not signify irrationality, bad calibration, or in any way a "pathological" condition" (p. 232). Similarly, Brewer and Sampaio (2012), who investigated confidence for different types of deceptive and non-deceptive items, argued that both types of items make use of the same cognitive products and processes, and therefore participants cannot be aware that an item is deceptive.

In this study, we used deceptive and non-deceptive items from different domains. Participants were warned that for some of the items most participants tend to choose the wrong answer. Their task was to decide for each item whether it was non-deceptive, drawing mostly correct answers, or whether it was deceptive, yielding a preponderance of wrong answers. Confidence and response speed were also measured to examine the possibility that participants have implicit access to the conflict involved in deceptive items.

Experiments 1 and 2 used CC and CW perceptual items, whereas Experiment 3 used general-information questions. In both tasks, erroneous responses have been assumed to derive from the same process that generally leads to correct responses. Thus, our hypothesis was that for these tasks, participants cannot discriminate between deceptive and non-deceptive items and hence cannot escape the dangers lurking in deceptive items. They can do so only when they know which the correct answer is, relying on their own tendency to choose it (Experiment 4). Experiment 5, in turn examined whether participants can spot deceptive problems that tend sometimes to induce fast erroneous responses because of impulsive reliance on System 1 mode of processing.

EXPERIMENT 1

Experiment 1 used two perceptual tasks that had been used by Koriat (2011). The first task required deciding which of two lines is longer whereas the second task required deciding which of two geometrical shapes had a larger area. For both tasks, the items had been found to differ extensively in the percentage of correct responses so that for some items (CC) the majority of responses were correct, whereas for others (CW) the majority of responses were wrong. For each item, participants first predicted the response of the majority of participants. They saw the same pairs again and performed a deceptiveness discrimination task: They were warned that for some of the items, most participants had been found to choose the wrong answer, and their task was to decide for each pair, whether it is likely to create the illusion that the wrong answer is the correct answer, and to indicate their confidence in their decision. They predicted again the majority choice for people who had not been warned about the possibility of deceptive items, and then indicated their own choice of the correct answer. The question was whether their choice would now depart from what they had judged to be the popular choice.

Method

Stimuli and procedure

Two perceptual tasks from Koriat (2011) were used, one (Lines) required deciding which of two irregular lines was longer, and another (Shapes) required deciding which of two geometric shapes had a larger area. There were 40 pairs in each task. Eight items for the Lines task, and 15 items for the Shapes task were classified by Koriat (2011) as CW items on the basis of participants' performance. For examples of the stimuli see Figure 2 in Koriat (2011).

Apparatus and procedure

The experiment was conducted on a personal computer. In Block 1, the 40 line pairs (preceded by two practice pairs) were presented in turn. When participants clicked a *show line drawings* box, the two stimuli appeared side by side, with the question "*What was the majority answer?*" Participants predicted the response of the majority of participants by clicking one of the two lines with the mouse, and then clicked a *confirm* box. Participants then indicated their estimate of the percentage of participants who chose the selected answer by sliding a pointer on a 51%-100% scale using the mouse (a number in the range 51-100 corresponding to the location of the pointer on the screen appeared in a box). After clicking a second *confirm* box, the *show line drawing* box appeared on the screen, and the next trial began. When the block ended, participants were asked to make an aggregate estimate. The prompt was, "You were presented with 40 pairs. For how many of them do you think you were correct in choosing the majority answer?"[1]

[1]The aggregate judgments did not yield interesting results and will not be reported.

DOI: 10.1002/bdm

The procedure for Block 2 was identical except that the stimuli consisted of the geometric shapes, and the task was to guess which of the two members of each pair would be judged as having a larger area.

In Block 3, participants were presented again with the 40 pairs of line drawing but performed three tasks successively: deceptiveness judgment, majority choice, and own choice. For the deceptiveness judgment task, participants were told: "You will be presented with the same stimuli again, but we would like you to think carefully. We warn you that for some of the pairs, most participants have been found to make the *wrong* decision. That is, the line that they chose as being longer was actually shorter. We want to know whether you can identify the deceptive pairs, those that create an illusion that induces people to make the wrong choice". The pairs were then presented one after the other followed by the question "*Does the pair create a deceptive feeling*?" Participants clicked *Yes* or *No*, and then a *confirm* box. Response latency was measured. A confidence scale (50–100) was then added beneath the drawings, and participants marked their confidence 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).

After clicking a second *confirm* box, the pair appeared again. Participants were instructed to guess the choice of the majority of participants among those who had not been warned about the possibility of deceptive items. They clicked one of the two lines and then a *confirm* box. A confidence scale (50–100) was added on the screen, and participants marked their confidence by sliding a pointer on the scale.

After clicking the *confirm* box, the pair appeared for the third time. Participants were now asked to indicate their own judgment of which of the two lines was longer by clicking one of the lines and then a *confirm* box. Response latency was measured. Participants indicated their confidence on a 50–100 scale. After clicking a *confirm* box, the next pair appeared.

When the block ended, participants were asked to make an aggregate estimate for their predictions of others' responses[1].

The procedure for Block 4 was identical except that the Shapes task was used.

The order of the pairs as well as the order of the left-right arrangement of the members within each pair were determined randomly for each participant and task in Blocks 1 and 2. However, the same orders and arrangements were retained in Blocks 3 and 4.

Participants

Twenty undergraduate students from the University of Haifa (7 males) participated in the experiment, 18 for pay and 2 for course credit.

Results

Predicting others' responses

Examination of participants' predictions of others' responses in Blocks 1 and 2 indicated that they expected others' responses to be correct in 79.21% of the cases for the CC items, and in 19.13% of the cases for the CW items, $t(19) = 17.49$, $p < .0001$, $d = 6.43$, Thus, participants' predictions implied that others would be much more likely to be correct for CC items than for CW items.

Deceptiveness judgments

We now examine the results for deceptiveness judgments. The percentage of items judged as deceptive was calculated for each participant for the CC and CW items. The means and SDs of deceptiveness judgments, and of confidence and response latency for these judgments appear in Table 1.

Table 1. Mean deceptiveness judgments and confidence and response latency for these judgments for CC and CW items (classified as in Koriat, 2011) (STD in parentheses). (Experiment 1)

Task		Item Type	Percentage/Confidence/Latency	*t* -test
Lines	Deceptiveness Judgments	CC items	55.47% (22.44)	$t(19) = 0.46$, $p < .66$
		CW items	56.88% (28.53)	
	Confidence	CC items	73.79% (8.61)	$t(19) = 4.41$, $p < .0005$
		CW items	71.42% (9.01)	
	Response Latency	CC items	5.35s (2.48)	$t(19) = 1.72$, $p < .11$
		CW items	5.84s (2.57)	
Shapes	Deceptiveness Judgments	CC items	45.40% (25.11)	$t(19) = 2.02$, $p < .07$
		CW items	51.33% (25.51)	
	Confidence	CC items	73.85% (7.91)	$t(19) = 2.07$, $p < .06$
		CW items	72.27% (8.00)	
	Response Latency	CC items	5.06s (3.39)	$t(19) = 0.59$, $p < .57$
		CW items	5.37s (5.46)	
All	Deceptiveness Judgments	CC items	51.05% (22.89)	$t(19) = 1.04$, $p < .32$
		CW items	53.26% (25.31)	
	Confidence	CC items	73.82% (8.05)	$t(19) = 3.05$, $p < .01$
		CW items	71.98% (8.01)	
	Response Latency	CC items	5.22s (2.53)	$t(19) = 0.65$, $p < .53$
		CW items	5.57s (4.11)	

DOI: 10.1002/bdm

The results indicate that participants largely failed to discriminate between CC and CW items. Although the Shapes task yielded a trend suggesting higher ratings for the CW items, the results across the Lines and Shapes tasks indicated that participants judged as deceptive 51.05% of the CC items and 53.26% of the CW items when the percentages of correct responses in Koriat (2011) were 82.30% and 25.32%, respectively.

However, participants were significantly less confident in their judgment for CW items (71.98) than for CC items (73.82). This result may be seen to accord with the idea that people are sensitive to the presence of conflict at an implicit level (De Neys et al., 2011; Mevel et al., 2015). We shall return to this point in Experiment 2.

Response latencies above or below 2.5 STDs from each participant's mean were eliminated (3.56% across tasks). As can be seen in Table 1, there were little differences in response latency between CC and CW items.

Participants' predictions of others' answers and their own answers

Asking participants to predict others' answers again before indicating their own answer was intended to sharpen the contrast between what they believed to be others' answer and the answer that they themselves might choose after having performed the deceptiveness discrimination task.

Examination of participants' predictions (Other) indicates that they expected others to be correct in 74.47% of the cases for CC items, and in 21.09% of the cases for the CW items. The results for participants' own answers were quite similar, averaging 76.32 and 30.22, respectively. An analysis of variance (ANOVA) comparing the accuracy of Other and Self answers indicated higher accuracy for Self (53.27) than for Other (47.78), $F(1, 19) = 5.16$, $MSE = 116.60$, $p < .05$, but the interaction was not significant, $F(1, 19) = 1.41$, $MSE = 188.50$, $p = .25$.

Confidence in one's own answers averaged 70.64 and 69.97 respectively, for CC and CW items, $t(19) = 1.40$, $p = .18$, $d = 0.08$. Thus, there was no indication that confidence judgments discriminated between CC and CW items. In addition, despite the warning that some of the items were deceptive, confidence yielded the same interactive pattern as in Koriat (2011; see Koriat 2012): For CC items, confidence was higher for correct answers (71.75) than for wrong answers (66.33), $t(19) = 5.07$, $p < .0001$, whereas for CW items, it was higher for wrong answers (70.89) than for correct answers (68.09), $t(19) = 2.71$, $p < .05$.

EXPERIMENT 2

Although participants in Experiment 1 failed to discriminate between CC and CW items, they were less confident about their judgment for the CW items. This observation is consistent with the idea that even when participants may not explicitly detect that they are erring, implicit indexes can still reveal sensitivity to erroneous responses (De Neys, 2012).

Note, however, that even in Koriat's study (2011), confidence was higher for the CC items (72.30) than for the CW items (67.83), possibly because confidence generally increases with inter-participant consensus (Koriat, 2012), and consensus was in fact higher for CC items (82.30) than for CW items (74.68) in Koriat (2011).

Two changes were introduced in Experiment 2. First, CC and CW items were matched in degree of inter-participant consensus. Second, an equal number of CC and CW items was used in each task (Lines or Shapes), and participants were informed that exactly half of the items in each task had been found to yield mostly wrong answers.

Method

Stimulus materials

The stimuli were selected from Koriat (2011) so that for the Lines task, the 8 CW items were used in addition to 8 CC items that matched them closely in terms of the percentage of consensual choices. Mean correct responses for these items in Koriat (2011) were 80.13% and 25.96%, respectively. For the Shapes task, 15 CC items and 15 matching CW items were used. Their mean percent accuracy in Koriat (2011) was 78.05% and 24.72%, respectively.

Apparatus and procedure

In Block 1, Participants first predicted others' responses to each item. All pairs were presented again for deceptiveness judgments. The instructions were the same as in Experiment 1, but participants were warned that exactly 8 pairs out of the 16 pairs had been found to yield a majority of wrong answers. Finally, participants saw the same pairs again, and were asked to indicate their own answers. Confidence and response latency were measured.

The procedure for Block 2 was similar except that the 30 Shapes pairs were used. In the deceptiveness judgment task, participants were warned that exactly 15 of the 30 pairs had been found to be deceptive, yielding a majority of wrong answers.

The order of the pairs as well the order of the left-right arrangement of the members within each pair were determined randomly for each participant for the two types of stimuli.

Participants

Twenty University of Haifa undergraduate students (7 males) participated in the experiment, 17 for pay and 3 for course credit.

Results

Two participants classified as "deceptive" all items in Block 1 and 29 out of the 30 items in Block 2. They were dropped from the analyses.

DOI: 10.1002/bdm

Predicting others' responses

Examination of participants' predictions of others' responses indicated that they expected others' responses to be correct in 71.48% of the cases for CC items and in 25.60% of the cases for CW items, $t(17) = 13.82$, $p < .0001$, $d = 4.94$, again implying higher accuracy for CC than for CW items.

Deceptiveness judgments

Table 2 presents the percentage of items judged as deceptive for the CC and CW items. The table also presents the means for confidence and response latencies, after eliminating all latencies below or above 2.5 SDs from each participant's mean (3.26% across tasks).

As in Experiment 1, participants failed to discriminate between deceptive and non-deceptive items. They judged as deceptive 48.79% of the CC items and 51.45% of the CW items, $t(17) = 0.60$, $p = .56$, $d = 0.14$. However, there was again a slight indication that participants were less confident in their response to CW items (77.39) than in their response to CC items (79.23), $t(17) = 2.86$, $p < .05$, $d = 0.67$. Response latency did not differ between CC (5.57s) and CW items (6.13s).

Participants' own answers

The mean percentage of correct answers in the Lines task averaged 79.17 and 27.78 for the CC and CW items, respectively, $t(17) = 8.05$, $p < .0001$, $d = 2.74$. The respective means for the Shapes task were 68.89 and 29.26 for the CC and CW items, respectively, $t(17) = 8.76$, $p < .0001$, $d = 3.08$. The percentage of correct answers across the two tasks, was 72.46 for CC items and 28.74 for CW items. These percentages hardly differ from those predicted for others, suggesting that the deceptiveness judgments task did not improve participants' discrimination between CC and CW items. However, like for deceptiveness judgments, confidence was somewhat lower for CW items (72.92) than for CC items (75.16), $t(17) = 2.57$, $p < .05$, $d = 0.26$.

In sum, people did not succeed in identifying deceptive items even when they were warned that 50% of the presented items were deceptive. However, confidence in deceptiveness judgments was still slightly lower for CW than for CC items.

EXPERIMENT 3

Experiment 3 extended investigation to a task involving general knowledge. The task included 40 2AFC geography questions. The questions were chosen on the basis of previous results (e.g., Brewer & Sampaio, 2012) so that some were likely to yield mostly wrong answers. In Block 1, participants chose the correct answer to each question. The results from this block provided the basis for distinguishing between CC and CW items. In Block 2, participants were presented with the same questions again but were required to make deceptiveness judgments.

Method

Materials and procedure

There were 40 questions, all of which concerned the spatial relationship between two cities. All questions, preceded by two practice questions, were presented in turn in Block 1. The question (e.g., "*which of the two cities is more to the north*?") appeared on top, followed by the names of two cities (e.g., *Toronto, Canada; Venice, Italy*). Participants chose the correct answer by clicking it, and indicated their confidence on a 50-100% scale.

In Block 2, participants made deceptiveness judgments. As in the previous experiments, they were warned that for some of the questions most participants tend to choose the wrong answer. Their task was to guess which of the questions was deceptive, so that most participants would be

Table 2. Mean deceptiveness judgments and confidence and response latency of these judgments for CC and CW items (classified as in Koriat, 2011) (STD in parentheses). (Experiment 2)

Task		Item Type	Percentage/Confidence/Latency	*t* -test
Lines	Deceptiveness Judgments	CC items	52.08% (20.67)	$t(17) = 0.00$
		CW items	52.08% (21.11)	
	Confidence	CC items	79.40% (10.52)	$t(17) = 2.59$, $p < .05$
		CW items	76.61% (9.42)	
	Response Latency	CC items	6.20s (2.50)	$t(17) = 1.20$, $p < .26$
		CW items	8.29s (7.43)	
Shapes	Deceptiveness Judgments	CC items	47.04% (12.62)	$t(17) = 0.76$, $p < .47$
		CW items	51.11% (18.01)	
	Confidence	CC items	79.14% (9.30)	$t(17) = 1.73$, $p < .11$
		CW items	77.80% (10.30)	
	Response Latency	CC items	5.23s (1.87)	$t(17) = 0.98$, $p < .35$
		CW items	4.98s (2.03)	
All	Deceptiveness Judgments	CC items	48.79% (9.50)	$t(17) = 0.60$, $p < .57$
		CW items	51.45% (14.48)	
	Confidence	CC items	79.23% (9.32)	$t(17) = 2.86$, $p < .05$
		CW items	77.39% (9.68)	
	Response Latency	CC items	5.57s (1.94)	$t(17) = 0.91$, $p < .39$
		CW items	6.13s (3.10)	

DOI: 10.1002/bdm

likely to choose the wrong answer. Participants clicked *yes/no* and then indicated their confidence in their judgment. Response latency was measured.

Participants

Fifty Hebrew-speaking University of Haifa undergraduates (41 women) participated in the experiment, 32 for pay and 18 for credit.

Results

The responses in Block 1 provided the norms against which the deceptiveness judgments were evaluated. Excluding 2 items for which exactly 25 participants gave the correct answer in Block 1, the remaining items were split between 21 CC items for which the percentage of correct answers averaged 60.67, and 17 CW items for which that percentage averaged 38.47. Across the 38 items, mean deceptive judgments in Block 2 correlated .01, $p = .96$ with mean correct responses in Block 1. In Block 2, participants judged each CC item as deceptive in 47.24% of the cases compared to 45.76% for CW items, $t(49) = 0.60$, $p = .56$, $d = 0.08$. Confidence in deceptiveness judgments averaged 72.96 and 72.70 for the CC and CW items, respectively, $t(49) = 0.46$, $p = .66$, $d = 0.07$. Response latency, after eliminating latencies above or below 2.5 STDs from each participant's mean (3.30%), averaged 6.57 s and 6.35 s, respectively, $t(49) = 1.92$, $p < .07$, $d = 0.27$. Thus, there was no indication that participants succeeded in discriminating between the two types of items.

Note that the results for confidence judgments were consistent with the consensuality principle (Koriat, 2008): Across the CC items, confidence was higher for those who chose the correct answer ($M = 68.83$) than for those who chose the wrong answer ($M = 66.30$), $t(49) = 2.35$, $p < .05$, $d = 0.33$. For the CW items, in contrast, confidence was higher for those who chose the wrong answer ($M = 69.74$) than for those who chose the correct answer ($M = 66.45$), $t(49) = 3.50$, $p < .001$, $d = 050$.

We repeated the analyses using a more conservative criterion, dividing items into 9 items that produced 32 correct answers or more, versus 8 items that produced 18 correct answers or less. On average, participants judged each CC item as deceptive in 40.67% of the cases compared with 36.25% for CW items, $t(49) = 1.18$, $p < .25$, $d = 0.17$. Confidence in deceptiveness judgments averaged 73.34 and 73.74 for the CC and CW items, respectively, $t(49) = 0.48$, $p < .65$, $d = 0.07$. The respective means for response latency were 6.47 s and 6.48 s, respectively, $t(49) = 0.03$, $p < .98$, $d = 0.004$.

In sum, there was no indication that participants succeeded in discriminating between the two types of items either explicitly or implicitly. These results are consistent with the claim of Brewer and Sampaio (2012; see also Brewer et al., 2005), who used a similar general-knowledge task, that participants are not aware that an item is deceptive.

EXPERIMENT 4

As noted earlier, the claim that experimenters and test makers tend to over-select deceptive or misleading items implies that they can access the properties that distinguish these items from non-deceptive items. If experimenters and subjects share the same knowledge structure that determines choice and confidence, then subjects might be able to identify misleading items when they are specifically alerted to their existence. The results of Experiments 1-3, however, did not support that possibility. It is our conjecture that experimenters and test-makers are able to identify misleading items only because they generally have access to the correct answer (as was the case for the "selectors" in Juslin's study, 1994).

To examine this possibility, we had participants make deceptiveness judgments for the geography questions used in Experiment 3 when they were informed about the correct answer. The design was similar to that of Block 2 of Experiment 3. Participant first saw the question with the two response options, and shortly thereafter were presented with an indication of which of them was the correct answer. They were asked then to decide whether the question is deceptive or not.

Method

Materials and procedure

The procedure was similar to that of Block 2 of Experiment 3. The items were also the same as those used in Experiment 3. These items, preceded by two practice items, were presented in turn; the question appeared on top, followed by the names of two cities. The question and the response options remained on the screen for 5 seconds. Participants were urged to try to guess the correct answer. The correct answer then changed its color to blue and flickered twice. Shortly thereafter the statement "*is the question deceptive?"* appeared on the screen with *yes/no* underneath. Participants clicked one of the two options and indicated their confidence in the deceptiveness judgments on a 50-100% scale. Response latency was measured.

The instructions for the deceptive judgments were the same as in Experiment 3. The order of the 40 questions was random for each participant.

Participants

Twenty-three Hebrew-speaking University of Haifa undergraduates (17 women) participated in the experiment, 8 were paid, and 15 received course credit.

Results

The items were divided into 3 categories based on the conservative criterion used in Experiment 3: Nine items for which accuracy averaged 64% or more (CC), eight items for which accuracy was 36% or less (CW) and the remaining non-consensual (NC) items. Deceptiveness ratings for these items in Experiment 4 averaged 35.27% for CC items, and 64.13% for CW items, $t(22) = 4.15$, $p < .0005$, $d = 0.87$.

DOI: 10.1002/bdm

Recall that the respective means in Experiment 3 were 40.67% and 36.25%, respectively. A two-way ANOVA comparing deceptiveness ratings for the CC and CW items in the two experiments yielded $F(1, 71) = 11.16$, $MSE = 412.23$, $p < .005$, for item type, $F(1, 71) = 10.37$, $MSE = 383.97$, $p < .005$ for Experiment, and $F(1, 71) = 21.16$, $MSE = 412.23$, $p < .0001$ for the interaction. Mean deceptiveness ratings for the NC items were 52.35% compared to 50.66% in Experiment 3.

Confidence in the deceptiveness ratings averaged 80.49 for the CC items, and 78.24 for the CW items, $t(22) = 1.28$, $p < .22$, $d = 0.27$. The respective means for response latency in the deceptiveness ratings were 2.98 s and 3.37 s, respectively, $t(22) = 1.35$, $p < .20$, $d = 0.28$.

We repeated the analyses using the less conservative criterion that had been used in Experiment 3. Deceptiveness ratings averaged 42.03% for the 21 items with above chance accuracy, and 60.87% for the 17 items with below-chance accuracy, $t(22) = 6.24$, $p < .0001$, $d = 1.28$. Confidence and response latency did not differ for the two categories of items, $t(22) = 1.45$, $p < .17$, $d = 0.30$, and $t(22) = 1.41$, $p < .18$, $d = 0.30$, respectively.

In sum, taken together, the results of Experiments 3 and 4 suggest that participants can discriminate between deceptive and non-deceptive questions only when they know which of the two answers is correct. Presumably, in that case they rely on their own responses in judging whether an item is likely to elicit mostly correct or mostly wrong responses among other participants (see Kelley & Jacoby, 1996). These results can explain why researchers and test-makers have been able to oversample almanac items for which participants tend to choose the wrong answer (Björkman, 1994; Hoffrage & Hertwig, 2006; Juslin, 1994).

EXPERIMENT 5

A final experiment was intended to obtain comparative results for participants' ability to discriminate between problems that tend to elicit a quick wrong solution and control problems that generally yield correct solutions. Ten experimental problems were used for which participants' solutions tend to be wrong. Among these were the three problems included in the Cognitive Reflection Test (CRT; Frederick, 2005), which have been assumed to induce errors due to reliance on System 1 processing (see Travers, Rolison, & Feeney, 2016). For these problems, an intuitive but wrong solution tends to spring quickly to mind, but participants can catch the error after some reflection. The remaining problems in this category were selected on the basis of an exploratory study in which each of them had been found to bring to mind a particular wrong solution. The 10 problems in the experimental category were compared with ten control problems that do not offer an immediate intuitive solution but require computation. Each of the 20 problems was presented for up to 20 seconds. Participants were asked to indicate whether it tends to yield primarily correct answers or primarily wrong answers. All problems were presented once again in a second block, and participants were asked to try to solve each of them.

Method

Stimuli

Twenty problems were compiled from different sources. They were translated to Hebrew and adapted to the Israeli population. All problems were open ended, requiring a numerical answer except for three that required a choice between two answers. The experimental problems, which included the three CRT problems (Frederick, 2005), had all been found to elicit quick wrong solutions in an exploratory study. The remaining 10 problems, required computation to reach the solution.

Examples:

Experimental: A frog fell into a well thirty meters deep. Each day he jumped two meters up the wall and slid back down one meter each night. How many days did it take him to jump out of the well?

Control: There are several books on a bookshelf. If one book is the 4th from the left and 6th from the right, how many books are on the shelf?

Apparatus and procedure

In Block 1, participants were told that "in previous studies we presented people with several problems. Some of these were found to be deceptive so that the majority of people gave the wrong solution. Others were found to be non-deceptive, eliciting mostly correct solutions. The same problems will be presented to you. Your task is *not* to solve these problems. Rather, we want you to guess whether each problem is deceptive or non-deceptive; that is, whether most people are likely to provide a wrong answer or most people are likely to give the correct answer". Participants were also told that the deceptiveness of a problem is unrelated to its difficulty; both difficult and easy problems can be either deceptive or non-deceptive. They were instructed that each problem will appear on the screen for only 20 seconds and they have to judge quickly whether it is deceptive or non-deceptive.

The 20 problems were presented in a random order, preceded by two practice problems. Each trial began by the participant clicking a *show problem* box. The problem then appeared followed by the question "*Is the problem deceptive?"* Participants clicked *Yes* or *No* and then *confirm*. The problem remained on the screen for up to 20 seconds. Response time was measured from the presentation of the problem to the *confirm* click. Participants then indicated their confidence in their decision on a 0-100% scale. After clicking a *confirm* box, a *show problem* box appeared on the screen and the next trial began.

In Block 2, the 20 problems were presented in turn again. However participants were asked to solve each problem and to type in the numerical answer and then to click *confirm*. Participants were allowed to use a pen and paper to carry out their calculations and were requested to provide an

DOI: 10.1002/bdm

answer to each problem. Response time from the appearance of the problem to the *confirm* click was measured. Participants then indicated their confidence in the correctness of the answer on a 0-100% scale. After clicking a *confirm* box, the following question appeared: *Did you try to solve this question when it appeared in block 1?* Participants chose one of the following response options: *Yes I tried and got the same answer, Yes, I tried and got a different answer, yes, I tried, but only superficially* and *No, I did not try*. After pressing a *confirm* box, the next problem appeared. The order of the problems was determined randomly for each participant and for each block.

Results

We first examine the results from Block 2, which can provide a check on the presumed qualitative differences between the two categories of problems used in this study. A critical feature of the 3 CRT problems is that each is assumed to bring to mind quickly an intuitive but wrong solution. This was indeed the case even in Block 2. For these problems, in 65.1% of the cases participants gave a wrong solution, and among those who gave a wrong solution, in 81.4% of the cases that solution was the same dominant solution across participants. A similar pattern was observed for the remaining experimental problems. For these problems, participants gave the wrong solution in 71% of the cases, and that solution was the consensual solution in 79% of the cases.

In contrast, for the control problems, participants gave a wrong solution only in 14.5% of the cases, and that solution was the dominant solution across participant only in 3.1% of the cases. All 22 participants yielded a lower percentage of correct solutions for the experimental problems than for the control problems. In particular, all 10 control questions yielded more that 50% correct answers, whereas only 2 of the experimental problems yielded more than 50% correct answers. Note that solution time averaged 34.94 s for the experimental problems, and 44.75 s for the control problems. Confidence in the correctness of the solution averaged 77.10 for the experimental problems, and 90.17 for the control problems, $t(18) = 4.11$, $p < .001$.

Turning to the results from Block 1, the experimental problems were rated as deceptive in 45.45% of the cases across participants, compared with 20.00% for the control problems, $t(21) = 5.99$, $p < .0001$. Mean deceptiveness ratings for the three CRT problems (48.48%) was also significantly higher than that for the control problems, $t(21) = 4.05$, $p < .001$, $d = 0.86$.

Confidence in the deceptiveness judgments averaged 79.55 for the experimental problems and 83.95 for the control problems, $t(21) = 2.48$, $p < .05$. Response latency (after eliminating latencies above or below 2.5 STDs from each participant's mean, 2.27% in total) averaged 17.63 s for the experimental problems and 17.13 s for the control problems, $t(21) = 1.12$, $p < .29$.

The results for individual participants indicated that of the 22 participants, 18 rated the experimental problems as more deceptive than the control problems, and 2 yielded a tie, $p < .0005$ by binomial test. We also repeated the analysis using only participants who reported that they had not attempted to solve the problem in Block 1. Using 19 participants with complete data, the experimental problems were rated as deceptive in 55.91% of the cases compared with 24.78% for the control problems, $t(18) = 2.77$, $p < .05$.

We repeated the analyses after eliminating the two experimental questions that yielded more than 50% correct solutions in Block 2. In this manner, the percentage of correct solutions varied across questions from 0% to 45.4% for the remaining 8 experimental problems, and from 59.1% to 100% for the 10 control problems. The 8 experimental problems were rated as deceptive in 44.32% of the cases across participants in Block 1 compared with 20.00% for the control problems, $t(21) = 5.19$, $p < .0001$, $d = 1.11$

In sum, the results of Experiment 5 are different from those of the previous experiments. They suggest that people can sense that there is something tricky about problems that are liable to yield erroneous solutions (see De Neys, 2014). However, because there is no clear operational definitions by which the two types of problems used in this study can be distinguished in terms of the process leading up to errors, the comparison between them is only suggestive.

DISCUSSION

The question addressed in this study was whether people can identify so-called "deceptive" or "misleading" items for which most participants opt for the wrong answer. These items have been assumed to impair both calibration and resolution. In fact, across deceptive, CW items, participants' confidence tends to be counter-diagnostic of accuracy.

The question whether people can distinguish between deceptive and non-deceptive items is important both practically and theoretically. The practical implications derive from the fact that people rely heavily on their confidence in translating their beliefs into action (Gill, Swann, & Silvera, 1998; Koriat & Goldsmith, 1996). The erroneous responses to deceptive, CW items are associated with inordinately high confidence, which should increase the likelihood of acting on these responses (Koriat, 2011). In addition, it was observed that for CW items, group discussion, rather than mitigating errors, actually amplified them while enhancing confidence in the wrong answers (Koriat, 2015).

Two observations suggest that people can discriminate between deceptive and non-deceptive items. First, the claim that experimenters and test makers tend to over-select deceptive or misleading items implies that they have access to the distinctive properties of these items. It was suspected, however, that people (e.g., test-makers and "selectors" in Juslin, 1994) are able to identify misleading items only when they have access to the correct answers to these items. This possibility was supported by the results of Experiment 4. It seems that participants judge the response of others on the basis of their own responses, by assuming that others respond like themselves (Kelley & Jacoby, 1996; Krueger, 1998).

DOI: 10.1002/bdm

Second, conflict detection studies suggest that people can detect that their heuristic response is questionable (De Neys, 2012, 2014; see Bago & De Neys, 2017). We argued, however, that this may be true of situations in which the process that yields an erroneous answer differs from that underlying correct answers. Indeed, in Experiment 5, the CRT problems, in which errors have been assumed to derive from a tempting, heuristic-driven response, were rated as more deceptive than the control problems.

In many cases, however, errors derive from the same process that underlies correct answers. For example, it was argued that responses to almanac questions is based generally on inference from cues (Gigerenzer et al., 1991; Koriat, 2012). These cues often support the correct answer but sometimes lead to the wrong answer. Because the cues for each item are largely shared by individuals with the same experience (Juslin, 1994; Juslin & Olsson, 1997; Koriat, 2012), reliable inter-item differences exist in the tendency of items to elicit the correct answer or wrong answer across participants. Furthermore, according to the self-consistency model (Koriat, 2012), confidence is based on the agreement among the sampled cues in favoring that answer (Alba & Marmorstein, 1987; Brewer & Sampaio, 2012; Slovic, 1966). That is, it is based on the reliability with which the cues support a given answer rather than on the validity of these cues. Therefore, confidence correlates with the consensuality of the answer – its likelihood to be chosen across participants. The implication is that the choice likelihood and confidence associated with an answer are based on the same process regardless of the accuracy of the answer. Hence participants should fail to tell whether an item is deceptive (CW) or non-deceptive (CC).

A similar argument was made by Brewer and Sampaio (2006, 2012) who compared confidence for deceptive and non-deceptive items using episodic and semantic memory tasks. In their metamemory approach to confidence, they proposed that "individuals are not aware of the nature of deceptive items and use the same processes and products for these items that they use in responding to non-deceptive items" (Brewer & Sampaio, 2012, p. 68).

The result of Experiments 1-3 were consistent with this conclusion, suggesting that the process that leads to wrong responses is the same as that underlying correct responses. In all of these experiments participants failed to discriminate between CC and CW items. In Experiments 1 and 2, the same CC-CW difference in accuracy as in Koriat (2011) was observed for participants' own answers even after having performed the deceptiveness judgment task. The implication is that for the type of items used in these experiments, participants cannot escape the dangers lurking in deceptive items, and cannot predict whether group discussion would be beneficial or detrimental.

The conclusion suggested by the results is that the ability of participants to detect misleading items should differ for different types of items depending on the process underlying erroneous answers to these items. Koriat (2012) proposed that in many tasks metacognitive errors are intimately linked to cognitive errors: What makes people confident about a correct or erroneous answer is what makes them choose that answer in the first place. Hence for these tasks, errors are difficult to detect and are hard to escape.

Can participants nevertheless detect misleading items at an implicit level? Confidence in deceptiveness judgments was somewhat lower for CW items than for CC items, and this was true even in Experiment 2 in which the two classes of items were matched in cross-person consensus. Although the size of the difference was small, this difference is consistent with the idea that people are able to detect an error at an implicit level even when they fail to detect it at the explicit level (see De Neys, 2014). In the conflict-detection literature, the lower confidence associated with heuristic-based errors was seen to reflect disfluent processing and a low feeling of rightness (Thompson, 2009; Thompson & Morsanyi, 2012). Consistent with this view is the proposal of Mata, Ferreira, and Sherman (2013) that deliberative thinkers are more confident in their answers than intuitive thinkers because they are aware of both the deliberative solution and the intuitive solution (see also De Neys, Rossi, & Houdé, 2013). However, it is not clear that the process underlying confidence in deceptiveness judgments is the same as that underlying the feelings of rightness in studies of conflict detection. More research is needed on this issue.

ACKNOWLEDGEMENTS

This work was supported by Grant 2013039 from the United States–Israel Binational Science Foundation. Thanks to Tamar Jermans, Mor Peled and Shai Raz for data collection, to Miriam Gil for her help in the analyses, and to Etti Levran (Merkine) for her help in copyediting.

REFERENCES

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

Alter, A. L. (2013). The benefits of cognitive disfluency. *Current Directions in Psychological Science*, *22*, 437–442.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*, 569–576.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, *39*, 133–144.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109.

Baron, R. S. (2005). So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. *Advances in Experimental Social Psychology*, *37*, 219–253.

Björkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes*, *58*, 386–405.

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition*, *38*, 186–196.

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

DOI: 10.1002/bdm

Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language*, *67*, 59–77.

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.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*, 28–38.

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking and Reasoning*, *20*, 169–187.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS One*, *6*, e15954.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*, 1248–1299.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*, 269–273.

Dhami, M., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*, 959–988.

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

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

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.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*, 25–42.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

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.

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

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

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Hill, G. W. (1982). Group versus individual performance: Are N+1 heads better than one? *Psychological Bulletin*, *91*, 517–539.

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). New York: Cambridge University Press.

Janis, I. L. (1982). *Victims of groupthink* (2nd ed.,). Boston: Houghton-Mifflin.

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.

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.

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.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus & Giroux.

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

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.

Koriat, A. (1976). Another look at the relationship between phonetic symbolism and the feeling of knowing. *Memory & Cognition*, *4*, 244–248.

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

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

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

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

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

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

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

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

Krueger, J. (1998). On the perception of social consensus. *Advances in Experimental Social Psychology*, *30*, 164–240.

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.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probability: Theories and models 1980-94. In G. Wright (Ed.), *Subjective probability* (pp. 453–482). Chichester, England: Wiley.

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology*, *105*, 353–373.

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

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., et al. (2015). Bias detection: response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*, 227–237.

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., et al. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General*, *144*, e16–e30.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–32.

Roediger, H. L. III, & DeSoto, K. A. (2014). Confidence and memory: Assessing positive and negative correlations. *Memory*, *22*, 76–91.

DOI: 10.1002/bdm

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.

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

Slovic, P. (1966). Cue-consistency and cue-utilization in judgment. *The American Journal of Psychology*, *79*, 427–434.

Song, H., & Schwarz, N. (2008). Fluency and the detection of misleading questions: Low processing fluency attenuates the Moses illusion. *Social Cognition*, *26*, 791–799.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, *23*, 645–665.

Sunstein, C. R., & Hastie, R. (2015). *Wiser: Getting beyond groupthink to make groups smarter*. Boston: Harvard Business Review Press.

Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In J. B. S. T. Evans, & K. Frankish (Eds.), *In two minds: Dual processes and beyond*. Oxford, UK: Oxford University Press.

Thompson, V. A., & Morsanyi, K. (2012). Analytic thinking: do you feel like it? *Mind & Society*, *11*, 93–105.

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., et al. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*, 237–251.

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109–118.

Authors' biography:

Asher Koriat is Professor of Psychology and member of the Institute of Information Processing and Decision Making at the University of Haifa, Israel. He has degrees from Hebrew University, Jerusalem, and from University of California Berkeley. His research interests are memory and metacognition.

Authors' address:

Asher Koriat, University of Haifa, Haifa, Israel.

DOI: 10.1002/bdm