Journal of Experimental Psychology: General
2015, Vol. 144, No. 5, 934–950

0096-3445/15/$12.00 http://dx.doi.org/10.1037/xge0000092

When Two Heads Are Better Than One and When They Can Be Worse: The Amplification Hypothesis

Asher Koriat
University of Haifa

According to the self-consistency model (Koriat, 2012a), confidence judgments in the responses to 2-alternative forced-choice items are correlated with the consensuality of the responses rather than with their correctness: For consensually correct (CC) items, for which the majority response is correct, accuracy is higher for the correct answer than for the wrong answer, whereas for consensually wrong items (CW), confidence is higher for the wrong answer. Assuming that group decisions are dominated by the more confident members, a maximum confidence slating (MCS) algorithm that was applied to virtual dyads outperformed the better member for CC items, but yielded worse performance than the worse member for CW items (Koriat, 2012b). We examined whether group deliberation also amplifies the tendencies that are exhibited by individual decisions, or rather improves performance for both CC and CW items. A perceptual task and a general-information task yielded very similar results. MCS applied to the individual decisions yielded a similar amplification as in Koriat (2012b), but dyadic interaction accentuated this amplification further. Thus, group deliberation had an added effect over confidence-based judgments, possibly due to the exchange of arguments within a dyad, but both confidence slating and group deliberation affected performance in the same direction, improving accuracy when individual accuracy was better than chance, but impairing it when individual accuracy was below chance. Notably, for CW items, group interaction not only impaired accuracy but also enhanced confidence in the erroneous decisions. The mechanisms underlying consensual amplifications were discussed.

Keywords: group decisions, subjective confidence, wisdom of crowds, self-consistency model, consensual amplification

Many decisions in everyday life are made jointly by several people. This is particularly true when the decisions are important, and when there is a great deal of uncertainty. A large number of studies explored the advantage of group-based decisions over individual decisions, and attempts have been made to devise techniques (such as the Delphi method, Dalkey, 1969) to improve group-based decisions. Two lines of investigation may be distinguished. The first relies on the wisdom-of-crowds phenomenon that information aggregated across a group of individuals is generally closer to the truth than the information provided by each individual (Armstrong, 2001; Larrick, Mannes, & Soll, 2012; Surowiecki, 2005). Research has examined different rules for combining judgments across individuals (see Ariely et al., 2000; Clemen, 1989; Wallsten, Budescu, Erev, & Diederich, 1997). The general finding is that sophisticated rules do not yield more accurate judgments than simple averaging (Armstrong, 2001). The idea underlying the wisdom-of-crowds has been extended to a within-person context (Herzog & Hertwig, 2009, 2014; Hourihan & Benjamin, 2010; Steegen, Dewitte, Tuerlinckx, & Vanpaemel, 2014; Vul & Pashler, 2008): When the same person provides several estimates, the aggregated estimate is more likely to be closer to the truth than any of the individual estimates.

The second line of research involves joint decisions reached by interacting group members. Several studies indicated that cooperative groups perform better than independent individuals on a wide range of tasks (e.g., Hill, 1982; Laughlin, 2011; Trouche, Sander, & Mercier, 2014). However, the "groupthink" phenomenon has been claimed to underlie some of the disastrous decisions made in U.S. history (Baron, 2005; Esser, 1998; Janis, 1982), and several studies suggest that group decisions can sometimes go astray (Lightle, Kagel, & Arkes, 2009; Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Sunstein & Hastie, 2015; Yaniv, 2011). Simulations of group-based forecasting, grounded in the social psychology of groups, indicated how group accuracy may vary with various factors such as group size, and the accuracy and distribution of individual forecasts (Kerr & Tindale, 2011).

Koriat (2012b; see also Bang et al., 2014) explored an algorithm for combining judgments across noninteracting individuals, which takes advantage of the subjective confidence of each group member in his or her judgment. This algorithm is based on two general findings. First, for many two-alternative-forced-choice (2AFC) tasks, confidence judgments in one's answer are generally accurate in discriminating between correct and wrong answers and solutions. Thus, the within-individual confidence/accuracy correlation is typically moderate to high for perceptual tasks and general-

This article was published Online First July 13, 2015.

The work reported in this article was supported by Grant 2013039 from the United States–Israel Binational Science Foundation.

I am grateful to Miriam Gil for her help in the analyses, and to Shiri Adiv and Etti Levran (Merkine) for their help in copyediting. Liat Braun, Hila Berkovitch, and Tamar Jermans helped in the collection of the data.

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

knowledge tasks. This correlation has its counterpart at the group level: For each item, the confidence/accuracy correlation tends to be high across participants: When the confidence judgments of the members of a group are standardized so as to neutralize chronic differences in mean confidence judgments, the more confident individuals for each item are more likely to be correct than the less confident individuals (Koriat, 2012a).

The second observation is that participants take the validity of their subjective confidence for granted and use it to guide their behavioral decisions (Fischhoff, Slovic, & Lichtenstein, 1977; Gill, Swann, & Silvera, 1998; Koriat, 2011; Pansky & Goldsmith, 2014). For example, in Koriat and Goldsmith's (1996) study on the strategic regulation of memory performance, when participants were allowed the option to withhold information that is likely to be wrong, the decision to volunteer or withhold an answer was based practically entirely on the subjective confidence in that answer. In general, judgments associated with higher confidence tend to have greater behavioral impact than those associated with lower confidence. The counterpart of this observation at the group level is that when groups attempt to reach a group decision, they rely more heavily on the more confident members: For each issue, group members who are more confident in their judgments tend to have greater impact on the decision that is endorsed by the group than those who are less confident (Aramovich & Larson, 2013; Bang et al., 2014; Cutler, Penrod, & Stuve, 1988; Moussaïd, Kämmer, Analytis, & Neth, 2013; Tormala & Rucker, 2007; Zarnoth & Sniezek, 1997).

On the basis of these two observations, the maximum-confidence slating (MCS) algorithm was applied to dyadic decisions. Participants, who took part in the experiments individually, were paired ad hoc to form virtual dyads, and their confidence judgments were standardized to neutralize chronic individual differences in mean confidence. For each item, the decision that was made with higher confidence by one member of the dyad was selected, and all selected decisions were compiled to form a dummy high-confidence participant. In two studies, one (Study 1) using a perceptual task taken from Bahrami, Olsen, Latham, Roepstorff, Rees, and Frith (2010), and another using a general knowledge task (Study 2), MCS was found to yield a two-heads-better-than-one (2HBT1) effect. In fact, it yielded better decisions than the best member of a dyad.

These results document the benefits that may ensue from group decisions in comparison with individual decisions. However, the theoretical model proposed by Koriat (2012a, see also Koriat & Adiv, 2011) brings to the fore some of the perils lurking in confidence-based decisions. Koriat's self-consistency model (SCM) of subjective confidence addressed the question why confidence judgments are accurate in discriminating between correct and wrong decisions. It was proposed that confidence judgments do not monitor directly the accuracy of an answer or a decision, but are based on the reliability with which that answer or decision is supported across the clues that are consulted in making a choice. In responding to a 2AFC item, participants were assumed to draw a sample of clues from a population of item-related clues, and base their confidence on self-consistency—the balance of evidence in favor of the chosen response. Thus, they behave like intuitive statisticians who have to infer the central tendency in a population on the basis of a sample of observations. Like statistical level of confidence, subjective confidence depends on the extent to which the chosen answer is consistently supported across the sampled clues, so that reliability is used as a cue for validity.

An important assumption of SCM is that the population of item-related clues from which participants draw their clues on each occasion is commonly shared by people with the same experience. For 2AFC general-knowledge and perceptual items, most of the shared clues are assumed to favor the correct answer. This assumption accords with ecological approaches to cognition (Dhami, Hertwig, & Hoffrage, 2004; Gigerenzer, 2008; Juslin, 1994) and is also consistent with the wisdom-of-crowds phenomenon. Hence, although confidence monitors self-consistency, it is generally diagnostic of the correctness of the answers because the answer chosen, as well as confidence in that answer, are based on clues that tend to support the correct answer. Such should be the case when a set of 2AFC general-knowledge or perceptual items are selected randomly so that they are representative of their domain (see Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994). For such a set of items, the majority of items are CC, yielding more correct answers than wrong answers. Indeed, this was the case for most tasks that yielded a positive confidence/accuracy correlation (see, e.g., Study 2, Koriat, 2012b).

What happens when consensuality and correctness are disentangled by including a set of consensually wrong (CW) items for which the majority of participants choose the wrong answer? There are many such items that for one reason or another lead most people to opt for the wrong answer or solution (see Brewer & Sampaio, 2012; Fischhoff et al., 1977; Gigerenzer et al., 1991; Koriat, 1995; Sampaio & Brewer, 2009). Such items have been variously labeled "deceptive," "tricky," "misleading," or "non-representative." The results of several studies that included such CW items have been rather consistent: The typical positive confidence/accuracy correlation was observed only across the CC items. The CW items, in contrast, yielded a negative correlation so that confidence was higher for the wrong answers than for the correct answers. This pattern has been observed for a word-matching task (Koriat, 1976), general-knowledge (Koriat, 2008b), semantic memory (Brewer & Sampaio, 2012), perceptual judgments (Koriat, 2011), episodic memory (Brewer & Sampaio, 2006; DeSoto & Roediger, 2014) and the predictions of others' responses (Koriat, 2013).

What are the implications for group decisions? Koriat (2012b) applied the MCS algorithm to the results of two studies (Studies 3 and 4) in each of which a set of CW items was deliberately included. Study 3 used two perceptual tasks (deciding which of two lines is longer and which of two shapes has a larger area), whereas Study 4 involved 2AFC general information questions. Virtual dyads were formed as before. For CC items, the MCS algorithm yielded better accuracy than the best member in each dyad. In contrast, for CW items, two heads were significantly worse than one. In fact, for these items, the best accuracy was achieved by selecting for each item the response of the less confident member of a dyad. Thus, for situations in which individuals' performance is likely to be below chance, dyadic interaction would be expected to exacerbate the situation, increasing the likelihood of reaching the wrong decision.

These results are consistent with the consensual amplification hypothesis (see Sunstein & Hastie, 2015). Group interaction is expected to amplify the trend that is exhibited by individual decisions: It should improve accuracy for 2AFC items when indi-

vidual accuracy is better than chance, but should impair it when accuracy is below chance.

The work of Koriat (2012b) suggests one mechanism by which group interaction may be beneficial for representative issues or items. In addition, SCM implies a moderator variable that can distinguish between situations in which groups are expected to outperform individuals and those liable to yield a pattern that mirrors the groupthink phenomenon.

The assumptions underlying the MCS algorithm have been challenged recently by Trouche et al. (2014), who argued that in intellective tasks, "arguments, more than confidence, explain the good performance of reasoning groups" (p. 1958). In their study, when participants were given arguments against their answer in intellective tasks, many participants changed their minds to adopt the correct answer regardless of their initial confidence. Also, when participants solved intellective tasks individually and then in groups, they tended to adopt the correct answer when it was present in the group. This was true even when the correct answer was held with less confidence than the wrong answer. These results are consistent with previous findings indicating that in tasks involving reasoning, groups consistently outperform individuals (e.g., Moshman & Geil, 1998; Laughlin, 2011; Nussbaum, 2008). The results were taken to indicate that argument quality can overcome confidence.

Clearly, there are many tasks for which you can prove to yourself or to another person that a given answer is wrong (Mercier & Sperber, 2011, 2012). For some of these tasks, participants may not only discover the correct answer, but may also understand why other people (or even themselves) could fall into the trap of the wrong answer (Mata & Almeida, 2014). However, even for tasks with a demonstrably correct answer, several conditions must exist to enable the group to reach the correct solution. The member who has the correct solution must understand that solution, and must have the ability and motivation to demonstrate the correct solution to the incorrect members, and these members must have sufficient knowledge to accept the correct solution proposed (Laughlin & Ellis, 1986). Not surprisingly, Aramovich and Larson (2013) observed that although confidence and accuracy were correlated, confidence played a significant role in group discussions beyond its correlation with accuracy. Zarnoth and Sniezek (1997) also found that the confidence of group members predicted their influence on the group decision for intellective tasks. Nevertheless, there is no question that sound arguments can sometimes cause people to change their mind.

In the present study, we examined whether the consensual amplification hypothesis suggested by the simulation of Koriat (2012b) will hold true for collective group decisions. It should be noted that the predictions from SCM were tested under conditions that circumvented some of the complexities of real-life group decisions. Thus, in forming virtual dyads, participants were paired on the basis of their percent correct, and in addition, the confidence judgments were standardized in order to nullify chronic individual differences in confidence. In real-life, however, people differ in their chronic confidence (Kleitman & Stankov, 2001; Stankov & Crawford, 1997), and individual differences in mean confidence are only weakly correlated with degree of knowledge and expertise (see Kruger & Dunning, 1999; Williams, Dunning, & Kruger, 2013). Therefore, it is not clear that the predictions from SCM will hold true for empirical dyads (see Bang et al., 2014; Massoni & Roux, 2012). In addition, group discussion brings to the fore processes beyond those that follow from the members' relative confidence (e.g., Aramovich & Larson, 2013). An attractive hypothesis is that group deliberation may actually help mitigate the faulty decisions associated with CW items (Trouche et al., 2014), enhancing accuracy particularly for CW items. Indeed, recent studies indicated that group decisions are more accurate than what would be predicted by the MCS algorithm for a visual discrimination task (see Bang et al., 2014) and for clinical diagnosis (Hautz, Kämmer, Schauber, Spies, & Gaissmaier, 2015). These studies, however, did not include a distinction between different types of items.

In this study, we compared individual and dyadic decisions using a perceptual task (Experiment 1) and a general knowledge task (Experiment 2). Neither of these tasks would be classified as "intellective" (see Laughlin, 2011). In both tasks, members of a dyad first made their decisions individually and then interacted to reach a joint decision. Two questions were addressed. First, can the outcome of group discussion be predicted, at least in part, by the members' confidence in their individual decisions? To examine this question the accuracy of joint decisions will be compared to the accuracy that follows solely from the application of the MCS algorithm to the individual confidence judgments and decisions.

Second, does group interaction indeed yield an amplification pattern, improving performance for CC items but impairing performance for CW items? Alternatively, does the exchange of arguments within a dyad help mitigate the faulty decisions associated with CW items, similar to what has been observed for intellective tasks (Trouche et al., 2014)? Perhaps collaborative interaction should at least raise some doubts about the correctness of the answers to CW items, leading to reduced confidence in these answers.

In Experiment 1, participants decided which of two irregular lines was longer. The items were selected on the basis of the results of Koriat (2011, Experiment 1) to represent CC and CW items. In each trial, members of a dyad first made their decision individually and indicated their confidence in that decision, and then interacted with each other with the goal of reaching a joint decision. After reaching a joint decision, each of the two members indicated his or her degree of confidence in that decision. The advantage of the task used in Experiment 1 is that all the information was available to the participants, and they were free to examine the stimuli as long as they needed. In principle, participants could reach the correct answer had they been permitted to make use of technical devices (e.g., an opisometer).

Experiment 2, in turn, involved general-knowledge. True/false geographical questions were used, some of which were expected to yield predominantly correct answers whereas others were expected to yield predominantly wrong answers (see Brewer & Sampaio, 2012; Tversky, 1981). In Block 1, the two members of a dyad indicated their decision and confidence individually. In Block 2, they saw the same items again, and negotiated with each other to reach a joint decision, and each member then indicated his or her confidence in that decision. Block 3, which was a replication of Block 1, was intended to examine whether individual decisions were affected by the group decision that had been reached in Block 2.

A note on terminology. In previous studies in which some of the items have been found to draw a high proportion of erroneous

responses, these items have been variously labeled "deceptive," "misleading," or "tricky." Because some of these labels carry the connotation that participants were intentionally tricked (as might occur in exams), we will use the terminology of CC versus CW items, which is primarily descriptive.

Experiment 1

Participants in Experiment 1 were presented with pairs of irregular lines and were required to decide which of two lines was longer. In each trial, they indicated their individual decision for a given pair of lines before making a joint decision.

Method

Participants. Eighty University of Haifa undergraduates (62 women) participated in the experiment, 70 were paid and 10 received course credit. Same-gender participants were paired to perform the task according to their registration for the experiment.

Stimulus materials. The experimental materials consisted of pairs of line drawings that had been used in the study of Koriat (2011, Experiment 2). In that study, 40 pairs of line drawings were used. The pairs were constructed so as to yield a sufficiently large number of pairs for which participants would be likely to agree on the wrong answer. On the basis of these results, 16 pairs were used in the present study, eight CC items (77.50% correct responses across the five blocks in Koriat, 2011) and eight consensually wrong (CW) items (24.13% correct responses across the five blocks). Items were selected so that the CC and CW items matched as closely as possible in terms of the percentage of consensual choices.

Apparatus and procedure. The experiment was conducted in dyads. Dyad members sat in the same testing room, each viewing his or her own display screen. The two screens were placed on separate tables at a right angle to each other, and were connected to the same IBM computer.

The 16 pairs were presented in random order for each dyad, preceded by two practice pairs. Participants were told that for each pair, they should first judge individually which of the two lines was longer. The trial began when each of the two participants clicked a box labeled *Show line drawing*. The two lines then appeared side-by-side on each screen, with two circles underneath. Each participant indicated his or her response by clicking one of the two circles with the mouse (the clicked circle then changed its color), and then clicking a *Confirm* box (participants could change their response but not after clicking *Confirm*). After participants clicked the *Confirm* box, they indicated their confidence (the chances that their response was correct). A confidence scale (50–100) was added beneath the two lines, preceded by the question *How confident are you?* Each of the two participants marked his or her 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), and were encouraged to use the full range of the confidence scale.

The instructions at the beginning of the experiment indicated that the individual decisions of the two participants might differ; therefore, they should discuss their decisions, try to persuade each other if necessary, and come to an agreement about the joint decision. After both members clicked a second *Confirm* box, a sentence was added at the bottom of the screen: *Beginning of discussion; when you are through, press the space bar*. Participants were encouraged to communicate with each other even when they gave the same response, and when they disagreed, they had to negotiate the situation in order to agree on a joint decision. There was no time constraint on the dialogues between the two members. When the members reached an agreement, they pressed the space bar, which cleared the screen. The stimuli appeared again with the two empty circles below, and participants indicated their joint response by clicking one of the two circles (the program made sure that they had clicked the same response). A confidence scale then appeared, as before, and each participant marked his or her confidence in the joint decision. After participants clicked a second *Confirm* box, the *Show line drawings* box appeared on the screen, and the next trial began. The order of the 16 experimental pairs was determined randomly. The discussion between the two members during the dyadic interaction was recorded using wireless microphones (but only for the last 21 dyads).[1]

Results and Discussion

On average, participants spent 25.0 s reaching each individual decision, 23.5 for CC items, and 26.4 for CW items. The dyadic discussions lasted 28.8 s on average, 5.8 s when participants agreed on the response, and 72.9 s when they disagreed. During the dyadic discussions, participants typically voiced their considerations while pointing to different features of the lines on one of the two screens. Examples of the considerations: "I counted the number of segments; this line includes more segments than that"; "the two lines are roughly similar but this line has an additional segment."

Individual decisions: Confidence and accuracy for CC and CW items. Mean accuracy of the individual decisions for each item ranged from 70.0% to 90% for the CC items, and from 21.25% to 45.00% for the CW items. Across participants, mean accuracy (81.88) was higher than 50% for the CC items, $t(79) = 18.84$, $p < .0001$, $d = 2.11$, and lower than 50%, (33.44) for the CW items, $t(79) = 7.31$, $p < .0001$, $d = 0.82$. Note that consensus, calculated for each item across all participants, was higher for the CC items (averaging 82% across items) than for the CW items (67%).

Figure 1 presents mean accuracy and confidence for the CC and CW items. Confidence in the CW items was strongly inflated, averaging 72% when accuracy averaged only 33%, $t(79) = 14.84$, $p < .0001$, $d = 2.48$. For the CC items, in contrast, confidence exhibited an underconfidence bias, $t(79) = 3.04$, $p < .005$, $d = 0.47$. Thus, judging from confidence judgments, it would seem that participants were not aware of the deceptiveness of the CW items (see Brewer & Sampaio, 2012).

Agreement between the members and its effects on the joint decision and confidence. Both members of a dyad chose the same answer in 66.56% of the trials. In all of these trials, the joint decision was the same as the individual decision. For these "agreement" trials, confidence in the joint decision was higher (averaging 83.15 across the two members) than mean confidence in the individual decisions (75.01), $t(39) = 12.65$, $p < .0001$, $d = 1.18$.

[1] The results for the content of the dialogues are not reported in this article.

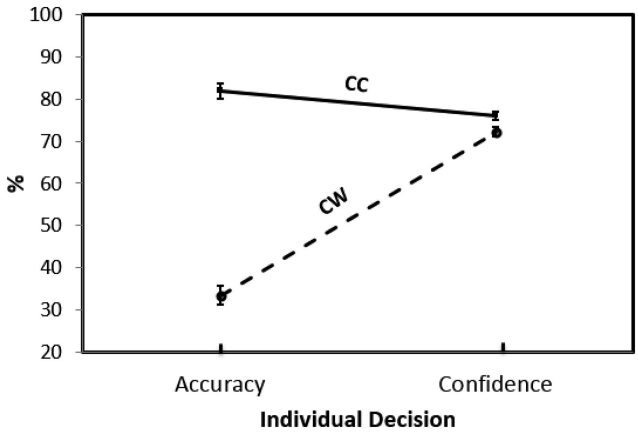

Figure 1. Mean accuracy and confidence for individual decisions plotted separately for the consensually correct (CC) and consensually wrong (CW) items (Experiment 1).

For the disagreement trials, individual confidence averaged 72.46 across the two members, and was lower than that for the agreement trials, $t(39) = 3.81$, $p < .0005$, $d = 0.36$. It might have been expected that for the disagreement trials, confidence in the joint decision would be lower than the average confidence in the individual decisions, but it was in fact higher, averaging 76.23, $t(39) = 4.00$, $p < .0005$, $d = 0.45$. Note that for these trials, confidence in the joint decision was higher for the participant whose individual judgment became the group judgment (80.72) than for the participant whose initial judgment was rejected (71.74), $t(39) = 6.47$, $p < .0001$, $d = 0.91$.

However, the increase in confidence was still lower when the two members initially disagreed on the response than when they agreed. A two-way analysis of variance (ANOVA) comparing mean individual and joint confidence for agreement and disagreement trials yielded higher confidence in the joint decision than in the individual decision, $F(1, 39) = 72.77$, $MSE = 19.47$, $p < .0001$, $\eta p^2 = .65$, and higher confidence for agreement than for disagreement trials, $F(1, 39) = 53.29$, $MSE = 16.82$, $p < .0001$, $\eta p^2 = .58$. The interaction, however, was also significant, $F(1, 39) = 29.04$, $MSE = 6.57$, $p < .0001$, $\eta p^2 = .43$.

Focusing on the disagreement trials, how did confidence in the individual decisions predict which of the two decisions was endorsed as the joint decision? Ignoring ties (4.2%), in 58.12% of the disagreement trials, the joint decision was the individual decision that had been associated with higher confidence, $t(39) = 2.07$, $p < .05$, $d = 0.33$, for the difference from 50%. When the confidence judgments of the two members were first standardized in order to control for chronic differences in confidence (see Kleitman & Stankov, 2001; Stankov & Crawford, 1997), the respective value was 61.41%, $t(39) = 2.75$, $p < .01$, $d = 0.43$. This result accords with the idea that joint decisions are generally dominated by the more confident members (Aramovich & Larson, 2013; Bang et al., 2014; Cutler et al., 1988; Tormala, & Rucker, 2007). However, this dominance was considerably lower than what was postulated in the simulation of Koriat (2012b).

Comparing accuracy and confidence for individual and dyadic decisions. We first compared the accuracy of individual and group decisions across all items. The mean individual accuracy (averaged across the two members of a dyad) averaged 57.66%, whereas the mean accuracy of the joint decision was 57.19%, $t(39) = 0.51$, $p < .62$, $d = 0.05$. Thus, across items there was no advantage to group decisions.

The results, however, differed for CC and CW items. The accuracy of individual decisions (averaged across the two members) and that of the joint decisions are presented in Figure 2A across the 40 dyads. A two-way ANOVA on these means yielded $F(1, 39) = 222.53$, $MSE = 510.29$, $p < .0001$, $\eta p^2 = 1.09$, for item type (CC vs. CW), and $F < 1$, for individual versus dyadic decisions. The interaction, however, was significant, $F(1, 39) = 12.93$, $MSE = 72.59$, $p < .001$, $\eta p^2 = .25$. For CC items, joint decisions were significantly more accurate than individual decisions, $t(39) = 2.82$, $p < .01$, $d = 0.37$. In contrast, for the CW items, joint decisions were significantly less accurate than individual decisions, $t(39) = 3.13$, $p < .005$, $d = 0.30$. It is noteworthy that both effects were significant despite a limiting ceiling effect for CC items, and a floor effect for CW items. The interaction observed is consistent with what was predicted by MCS (Koriat, 2012b), suggesting that the amplification hypothesis holds true for collaborative groups as well.

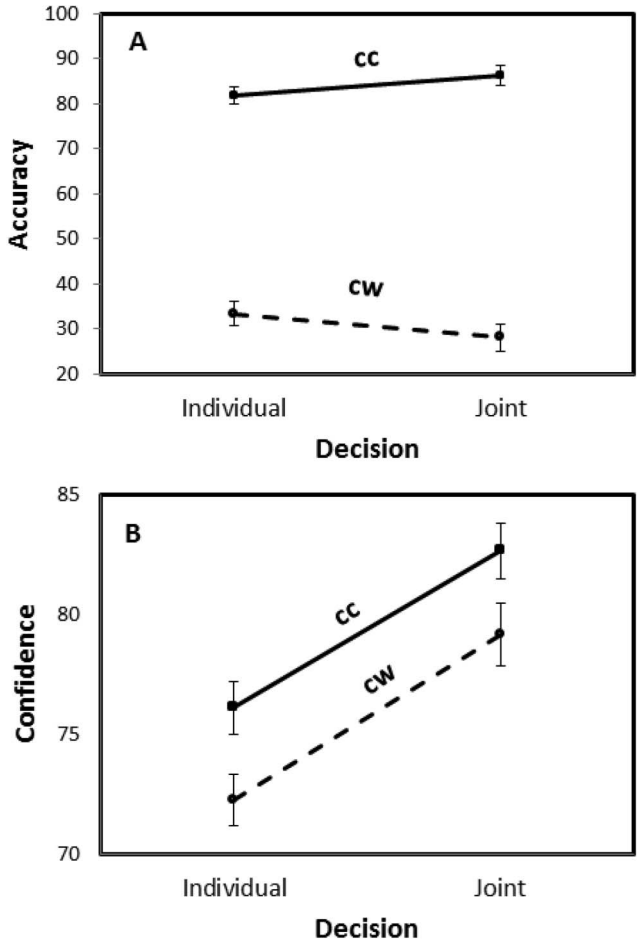

Figure 2. Panel A presents mean accuracy of individual decisions and joint decisions plotted separately for the consensually correct (CC) and consensually wrong (CW) items. Panel B presents the same results for mean confidence judgments (Experiment 1).

Turning next to confidence judgments, as noted earlier, dyadic interaction enhanced confidence in the joint decision. It is notable, however, that this was true for both the CC items and the CW items (see Figure 2B). A two-way ANOVA, item type (CC vs. CW) × Condition (Individual vs. Joint) yielded $F(1, 39) = 49.26$, $MSE = 10.85$, $p < .0001$, $\eta p^2 = .56$, for item type, $F(1, 39) = 112.23$, $MSE = 16.23$, $p < .0001$, $\eta p^2 = .74$, for condition, and $F < 1$ for the interaction. Confidence was higher for CC items than for CW items, but dyadic interaction enhanced confidence for both types of items. Thus, dyadic interaction improved accuracy and enhanced confidence for the CC items, whereas for the CW items, it impaired accuracy while enhancing confidence. Indeed, for the latter items, a two-way ANOVA comparing accuracy and confidence for individual and joint decisions yielded $F(1, 39) = 44.65$, $MSE = 33.62$, $p < .0001$, $\eta p^2 = .53$, for the interaction.

We examined whether the enhancement of confidence occurred even when the joint decision was wrong. For the CC items, there were 27 dyads that produced wrong joint decisions on some trials. Focusing on these trials, the individual confidence judgments averaged 72.34, whereas the confidence in the joint decision averaged 76.22 across the two members, $t(26) = 2.17$, $p < .05$, $d = 0.39$. The respective means for the CW items (using all 40 dyads) were 72.29 and 80.00, $t(39) = 11.13$, $p < .0001$, $d = 1.03$. Thus, dyadic interaction enhanced the members' confidence in their joint decision even when that decision was wrong.

The relationship between confidence and accuracy. As noted earlier, when the two members of a dyad disagreed, the joint decision was more likely to be the one that had been endorsed with higher confidence. To examine how the selection of the high confidence decision may have contributed to the accuracy of the joint decision, we analyzed the confidence-accuracy relationship for the individual decisions. Figure 3A presents mean confidence in individual decisions for correct and wrong decisions. The results are presented separately for CC and CW items using 56 participants who had data for all cells. For these participants, a two-way ANOVA on confidence judgments, Item Type (CC vs. CW) × Accuracy (Correct vs. Wrong) yielded $F(1, 55) = 10.23$, $MSE = 58.78$, $p < .0001$, $\eta p^2 = .16$, for the interaction. The results conform to the consensuality principle (Koriat, 2008b): For CC items, confidence was higher for correct decisions than for wrong decisions, $t(55) = 4.50$, $p < .0001$. For CW items, in contrast, confidence was higher for wrong decisions than for correct decisions but not significantly so, $t(55) = 1.05$, $p < .31$. The interaction was also supported by the within-individual confidence/accuracy gamma correlation across items (Nelson, 1984). The correlation was significantly positive across the CC items, averaging +.27 for 59 participants with complete data, $t(58) = 3.99$, $p < .0005$. For the CW items, in contrast, it was negative, averaging −.13 for 74 participants with complete data, $t(73) = 1.95$, $p < .06$.

Figure 3B presents the same results for the joint decisions. For each dyad, we first averaged the confidence of the two members in the joint decision, and then examined the relationship between average confidence and the accuracy of the joint decision for 24 dyads who had data for all cells. A similar ANOVA as before, yielded $F(1, 23) = 16.10$, $MSE = 47.04$, $p < .0005$, $\eta p^2 = .41$, for the interaction. For CC items, confidence was higher for correct decisions than for wrong decisions, $t(23) = 3.67$, $p < .005$, whereas for CW items, confidence was higher for wrong decisions than for correct decisions, $t(23) = 2.93$, $p < .01$. The gamma correlation between mean confidence in the joint decision and the accuracy of that decision was positive for CC items, averaging +.47 for 27 dyads with complete data, $t(26) = 4.79$, $p < .0001$, whereas for CW items it was negative, averaging −.29 for 34 dyads with complete data, $t(33) = 3.31$, $p < .005$. Thus, the consensuality principle also holds true for the confidence of the members of a dyad in their joint decision.

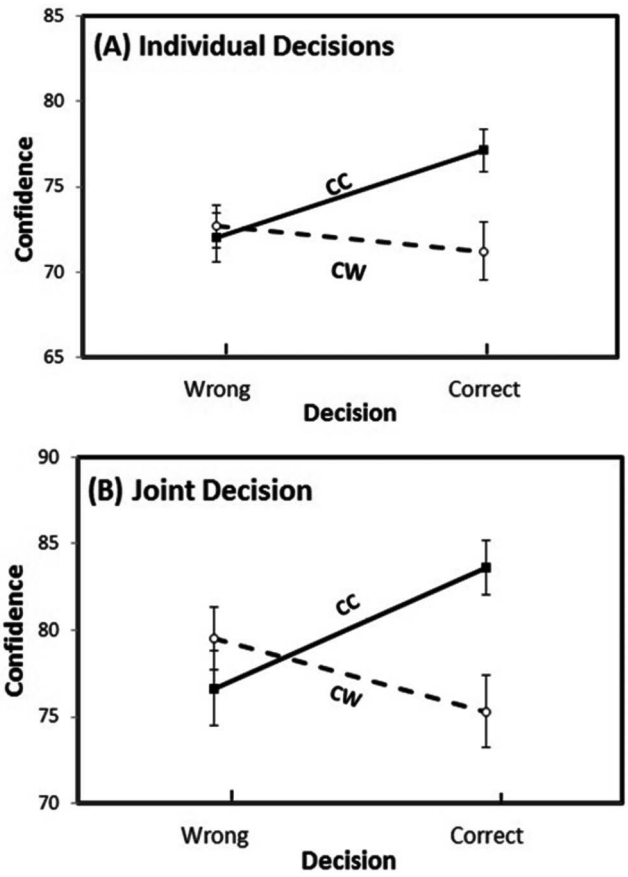

Figure 3. Panel A presents mean confidence for correct and wrong individual decisions, plotted separately for the consensually correct (CC) and consensually wrong (CW) items. Panel B presents the same results for mean confidence in the joint decisions (Experiment 1).

Experiment 2

Whereas Experiment 1 used a perceptual task, Experiment 2 used a general knowledge task for which the information is possibly not equally available to all group members. Therefore, the exchange of information between the members might be expected to be mostly beneficial to accuracy (Yaniv, 2004). Unlike Experiment 1, in which individual and joint decisions interspersed for each item, in Experiment 2 they were blocked so that individual decisions for all items were followed by joint decisions for these items.

Method

Participants. Eighty Hebrew-speaking University of Haifa undergraduates (40 women, and 40 men) participated in the experiment. Same-gender participants performed the task in dyads. For 36 dyads, participants were paid for their participation, and for 4 dyads participants received course credit.[2]

Stimulus materials. The experimental materials consisted of true/false geographical questions used in previous research. Example: *Hamburg, Germany is west of Casablanca, Morocco*. On the basis of previous results (e.g., Brewer & Sampaio, 2012), 14 CC items and nine CW items were selected. The CW items were questions that produce a high proportion of errors stemming mostly from alignment errors (Tversky, 1981; e.g., "Lima, Peru is west of Miami, Florida").

Apparatus and procedure. The apparatus and the sitting positions of the two members were the same as in Experiment 1. The experiment included three blocks. In Block 1, each participant performed the task individually. Participants were told that for each item, they should decide individually whether a geographical sentence was correct or wrong. After clicking a box labeled *Show the sentence*, a sentence appeared with *true* and *false* underneath. The sentence remained on the screens until the participant clicked *true* or *false* with the mouse (a circle underneath the clicked response was then filled). A *Confirm* box then appeared, which participants were expected to click (they could change their response but not after clicking *Confirm*). After clicking the *Confirm* box, a confidence scale was added, and the participant marked his or her confidence in the answer as in Experiment 1. After clicking a second *Confirm* box, the next sentence appeared. The first two sentences were used for practice; the remaining 23 experimental sentences, were ordered randomly, using the same random order for both participants.

When Block 1 ended, participants were told that they would be presented again with the same sentences but now they have to reach a joint decision. They were instructed that they may disagree on the answer, but they should discuss their decisions, try to persuade each other if necessary, and come to an agreement about the final decision. They were encouraged to communicate with each other even when they gave the same response.

The experimental sentences appeared in a new random order, preceded by the same two practice items. Each trial began when each of the two participants clicked a box labeled *Show the sentence*. The sentence then appeared, with *true* and *false* below. Each participant indicated his or her response. They attempted to come to an agreement, and only then each of them clicked the joint response. There was no time constraint on the dialogues between the two members, and these dialogues were recorded. After both participants clicked the same response and then the *Confirm* box, each of them indicated his or her confidence in the joint answer as in Experiment 1. After they clicked a second *Confirm* box, the next sentence appeared.

The procedure for Block 3 was the same as that used in Block 1 except that a new random order was used for the experimental sentences.

Results

The 23 items used in this study were selected on a priori grounds to represent CC and CW items. However, two CW items yielded better than 50% accuracy in the individual decisions: *Havana, Cuba is west to San-Jose, Costa Rica* (52.50%) and *Istanbul, Turkey is south of Lisbon, Portugal* (57.50%). These items were deleted from the analyses. Thus, the analyses will be based on 14 CC items and 7 CW items.

On average, participants spent 9.9 s reaching each individual decision. The dyadic discussions lasted 30.8 s, 26.9 s when participants initially agreed on a response, and 38.4 when they disagreed.

Individual decisions: Confidence and accuracy for CC and CW items. Figure 4 presents mean confidence and accuracy for the CC and CW items. Like in Experiment 1, confidence in the CW items was strongly inflated, averaging about 80 when accuracy averaged only 34%, $t(79) = 14.82$, $p < .0001$, $d = 2.71$. Confidence in the CC items also exhibited a certain degree of overconfidence $t(79) = 5.29$, $p < .0001$, $d = 0.50$. The overconfidence pattern differs from the underconfidence bias observed in Experiment 1 for the CC items. This difference is consistent with previous results suggesting that perceptual tasks tend to yield an underconfidence bias, unlike general-knowledge tasks, which tend to yield an overconfidence bias (see Björkman, Juslin & Winman, 1993; Juslin & Olsson, 1997; Winman & Juslin. 1993).

Confidence was higher for the CC than for the CW items possibly because the two classes of tems differed in their mean item consensus (79.5% for CC items and 65.9% for CW items; see Koriat, 2012a). When 4 CC items and 4 CW items were selected that matched roughly in mean consensus (74.4% for the CC items and 72.2% for the CW items), their mean confidence judgments averaged 83.9 and 80.5, respectively.

Determinants of the joint decision and the confidence in that decision. We examine next the potential determinants of the joint decision. Both members of a dyad chose the same answer in 64.40% of the trials. In 90.28% of these trials, the joint decision was the same as the individual decision. Thus, it is interesting that in about 10% of the trials the dyadic interaction resulted in both members changing their initial, individual decisions (The dyadic discussion for these trials averaged 47.4 s.). For the "agreement" trials (for which both members made the same individual judgment), confidence in the individual and joint decisions averaged 86.25 and 89.39, respectively, $t(39) = 4.93$, $p < .0001$, $d = 0.38$. Confidence for "disagreement" trials averaged 80.47 in Block 1, and was lower than that for the agreement trials, $t(39) = 4.69$, $p < .0001$, $d = 0.69$. For the disagreement trials, confidence in the joint decision (84.67) was also higher than for the individual decisions, $t(39) = 3.24$, $p < .005$, $d = 0.50$. As in Experiment 1, for the disagreement trials, confidence in the joint decision was higher for the participant whose individual judgment became the group judgment (86.54) than for the participant whose initial judgment was rejected (82.80), $t(39) = 4.52$, $p < .0001$, $d = 0.45$. Thus, dyadic interaction enhanced confidence for both agreement and disagreement trials, as was the case in Experiment 1.

Focusing on the disagreement trials, in 65.38% of these trials (ignoring 7.4% ties), the joint decision was the individual decision that had been associated with higher confidence in Block 1, based

[2] There were two additional dyads that were replaced. In one dyad, one member used 100% confidence judgments for all items except one, and in another dyad, one members used 50% confidence judgments for all items.

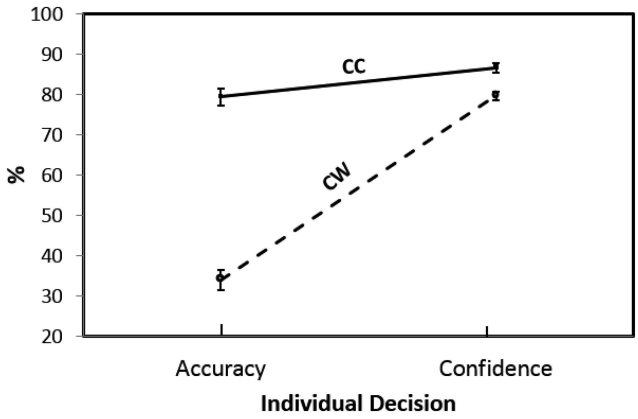

Figure 4. Mean accuracy and confidence for individual decisions plotted separately for the consensually correct (CC) and consensually wrong (CW) items (Experiment 2).

on 39 dyads (one dyad had only one disagreement trial for which both members gave the same confidence judgment), $t(38) = 4.33$, $p < .0001$, $d = 0.69$, for the difference from 50%. When the confidence judgments of the two members were first standardized in order to control for chronic differences in confidence, the respective value was 59.52%, $t(38) = 2.79$, $p < .01$, $d = 0.45$. This result accords with the idea that joint decisions are generally dominated by the more confident member.

Accuracy and confidence for individual and dyadic decisions. We first compare the accuracy of individual and group decisions across all items. The mean individual accuracy of the two members of a dyad averaged 64.35%, whereas the mean accuracy of the joint decision was 66.19%, $t(39) = 1.26$, $p < .22$, $d = 0, .25$. Thus, as in Experiment 1, there was no overall advantage of group decisions over individual decisions.

We then repeated this comparison separately for CC and CW items. As in Experiment 1, this was done using dyads as the unit of analysis. The results are presented in Figure 5A. A two-way ANOVA, Item Type (CC vs. CW) × Condition (Individual vs. Dyad) yielded $F(1, 39) = 143.35$, $MSE = 891.31$, $p < .0001$, $\eta p^2 = .79$, for Item Type, and $F(1, 39) = 1.17$, $MSE = 120.18$, $p < .30$ $\eta p^2 = .03$, for individual versus dyadic decisions. The interaction, however, was highly significant, $F(1, 39) = 32.01$, $MSE = 155.64$, $p < .0001$, $\eta p^2 = .45$. For CC items, dyadic decisions were significantly more accurate than individual decisions, $t(39) = 5.36$, $p < .0001$, $d = 0.66$, whereas for CW items the accuracy of the joint decision was significantly lower than the mean accuracy of the individual decisions, $t(39) = 3.97$, $p < .0005$, $d = 0.66$. This pattern is consistent with the results of Koriat's simulation (2012b) and replicates the amplification pattern observed in Experiment 1.

The respective results for confidence are presented in Figure 5B. A similar ANOVA to that for accuracy yielded $F(1, 39) = 58.21$, $MSE = 31.44$, $p < .0001$, $\eta p^2 = .60$, for item type, $F(1, 39) = 26.83$, $MSE = 18.62$, $p < .0001$, $\eta p^2 = .41$, for condition, and $F < 1$ for the interaction. Confidence was higher for CC items than for CW items, but the dyadic interaction enhanced confidence for both types of items, from 83.24 to 86.78. As in Experiment 1, dyadic interaction improved accuracy and enhanced confidence for the CC items, whereas for CW items it impaired accuracy while enhancing confidence. For these items, a two-way ANOVA comparing accuracy and confidence for individual and joint decisions yielded $F(1, 39) = 21.44$, $MSE = 131.65$, $p < .0001$, $\eta p^2 = .35$, for the interaction.

As in Experiment 1, we examined whether the enhancement of confidence occurred even when the joint decision was wrong. For the CC items, there were 24 dyads that produced wrong joint decisions on some trials. Focusing on these trials, confidence in the individual and joint decisions averaged 77.48 and 74.83, respectively, $t(23) = 1.05$, $p < .32$, $d = 0.23$. The respective means for the CW items (using all 40 participants) were 79.83 and 83.86, $t(39) = 4.08$, $p < .0005$, $d = 0.34$. Thus, for the CW items, dyadic interaction enhanced the members' confidence in their joint decision even when that decision was wrong.

The relationship between confidence and accuracy. Why were two heads better than one for the CC items but worse for the CW items? As noted earlier, the joint decision tended to be the one ventured with stronger confidence by one of the two members of a dyad. To examine how the selection of the high confidence

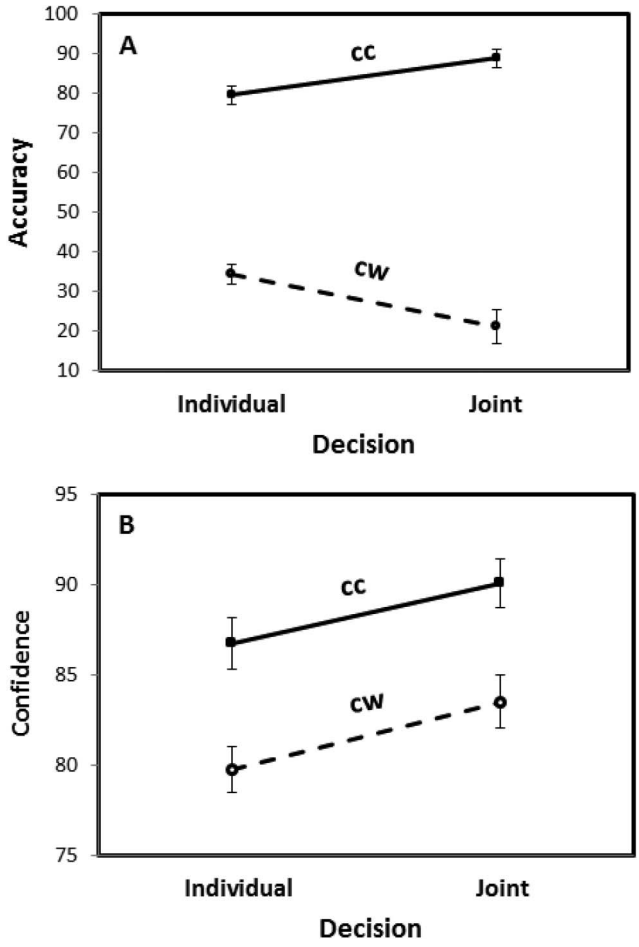

Figure 5. Panel A presents mean accuracy of individual decisions and joint decisions plotted separately for the consensually correct (CC) and consensually wrong (CW) items. Panel B presents the same results for mean confidence judgments (Experiment 2).

decision may have contributed to the accuracy of the joint decisions, we analyzed the confidence-accuracy relationship for the individual decisions. Figure 6A presents the pertinent results for CC and CW items for 64 participants who had data for all cells. A two-way ANOVA on these results yielded $F(1, 63) = 30.82$, $MSE = 107.14$, $p < .0001$, $\eta p^2 = .33$, for the interaction. For CC items, confidence was higher for correct decisions than for wrong decisions, $t(63) = 5.18$, $p < .0001$, whereas the opposite was true for CW items, $t(63) = 3.55$, $p < .001$. The interaction was also supported by the mean within-individual confidence/accuracy gamma correlation. This correlation was positive for the CC items, averaging +.27 for 67 participants with complete data, $t(66) = 3.78$, $p < .0005$. For the CW items, in contrast, it was significantly negative, averaging −.28 for 72 participants with complete data, $t(71) = 3.68$, $p < .0005$. Thus, assuming that the joint decision tended to be affected by the relative confidence of the two members, the pattern depicted in Figure 6A can explain in part the amplification pattern observed - why dyadic interaction was beneficial for CC items but detrimental for CW items.

Figure 6B presents the respective results for the joint decisions, which are based on 17 dyads that had data for all cells. The pattern is similar to that observed in Figure 6A. A similar ANOVA as before, yielded $F(1, 16) = 8.31$, $MSE = 111.25$, $p < .05$, $\eta p^2 = .34$, for the interaction. For CC items, confidence was higher for correct decisions than for wrong decisions, $t(16) = 3.88$, $p < .005$. For the CW items, confidence was higher for the wrong decisions than for correct decisions, but the difference was not significant, $t(16) = 1.24$, $p < .24$. The within-person gamma correlation was positive for the CC items, averaging +.51 for 24 dyads with complete data, $t(23) = 4.44$, $p < .0005$, whereas for CW items it was negative, averaging −.27 for 23 dyads with complete data, $t(22) = 1.96$, $p < .07$. Thus, as in Experiment 1, the consensuality principle was observed for the confidence of the members of a dyad in their joint decision.

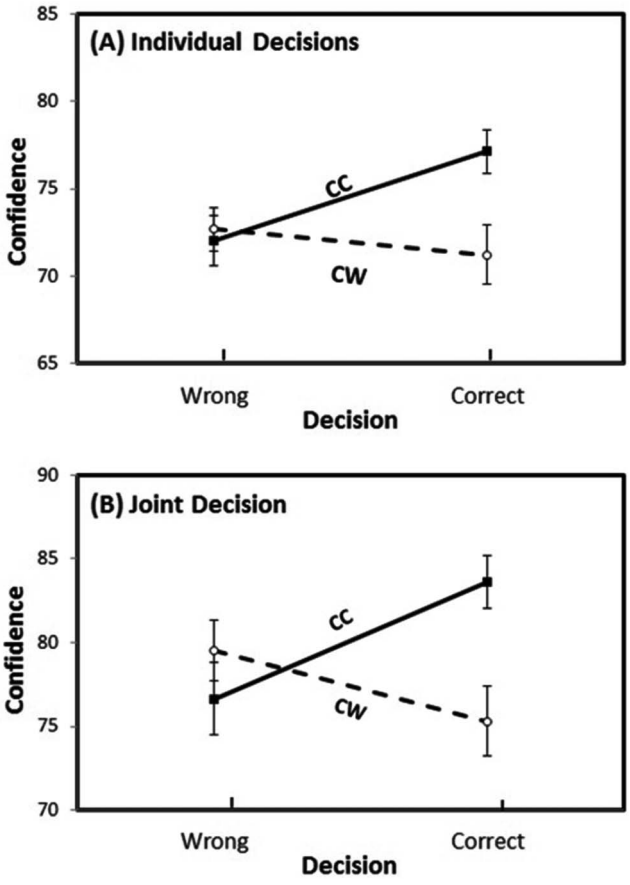

Figure 6. Panel A presents mean confidence for correct and wrong individual decisions, plotted separately for the items. Panel B presents the same results for mean confidence in the joint decisions (Experiment 2).

Block 3: Repeated individual decisions. The procedure for Block 1 was repeated in Block 3 in order to examine the effects of the dyadic interaction on subsequent individual decisions. For the CC items, accuracy increased from 79.46% in Block 1 to 83.04% in Block 3, $t(79) = 2.78$, $p < .01$, $d = 0.22$. In contrast, for the CW items, accuracy decreased from 34.11% in Block 1 to 25.71% in Block 3, $t(79) = 3.21$, $p < .005$, $d = .36$. Confidence increased for the CC items from 86.74 to 90.00, $t(79) = 5.10$, $p < .0001$, $d = 0.34$, and also for the CW items from 79.74 to 84.51, $t(79) = 5.82$, $p < .0001$, $d = 0.49$. Thus, the changes that occurred as a result of dyadic interaction persisted in part for the individual decisions in Block 3. It is interesting that for the CW items, dyadic interaction impaired the accuracy of the subsequent individual decisions.

We also compared accuracy in Block 3 with that in Block 2 using dyads as the unit of analysis. For the CC items, accuracy in Block 2 averaged 88.75%, whereas accuracy for dyads averaged 83.04% in Block 3, $t(39) = 4.61$, $p < .0001$, $d = 0.44$. For the CW items, in contrast, accuracy in Block 2 averaged 21.07%, whereas the accuracy for dyads in Block 3 averaged 25.71%, $t(39) = 2.16$, $p < .05$, $d = 0.19$. Thus, in Block 3, there was some regression toward the results obtained for the dyadic performance in Block 2.

In sum, the results of Experiment 2 were largely consistent with those of Experiment 1. Note that because in Experiment 2 the individual and dyadic decisions were made on two different blocks, some of the changes that occurred in the decisions reached in the two blocks may be due either to memory processes or to spontaneous changes that can occur independent of group interaction. Indeed, when participants were presented several times with the same set of 2AFC items, they were found sometimes to change their response from one presentation to another (Koriat, 2011, 2013; Koriat & Adiv, 2011, 2012; see Koriat, Adiv, & Schwarz, 2015). The results on the whole were consistent with the idea that participants construct their responses on the spot depending on the clues accessible when making the response. However, there were little systematic changes in accuracy across repetitions. Furthermore, the fact that the results obtained in Experiment 2 were very similar to those of Experiment 1 does much to mitigate against the possibility that the differences observed between Block 1 and Block 2 performance are due to memory processes or to spontaneous changes. Such, however might be true for the changes in accuracy that were observed between Block 3 and the previous two blocks.

Comparing Joint Decisions With the Application of the MCS Algorithm

In this section we compare the results obtained in the experiments with those that would be expected from the application of

the MCS algorithm. The two experiments yielded a very similar pattern of results, and therefore the analyses were carried out across the results of both experiments.

As in Koriat (2012b), for each dyad, the member with higher percentage correct was designated as high performing (*HP*), and the other as low performing (*LP*). In case of a tie (which occurred for 15 dyads out of the 80 dyads), one was randomly designated as *HP* and the other as *LP*. In addition, two dummy participants were formed based on the confidence in the individual decisions: For each item, the response of the member with higher confidence was slated to the dummy high-confidence (*D-HC*) participant, and the other to the dummy low-confidence (*D-LC*) participant. In case of a tie (same confidence score for both dummy participants), the members were divided randomly; for half of them the first member was considered as *D-HC* and the second member as *D-LC*, with the assignment counterbalanced across dyads with a tie. When the results were analyzed by item type (CC vs. CW) the same procedure was applied separately for each item type. Percent accuracy was then calculated for the four "participants." This was done first across all items, and then separately for the CC and CW items. Note that the *D-HC* participant represents the MCS algorithm.

Figure 7 presents the results that were obtained when the analysis was carried out separately for the CC and CW items. The figure presents the means for *D-LC*, and *D-HC*, and also includes the mean actual accuracy of the individual decisions and joint decisions (the accuracy of individual decisions was plotted midway between *D-LC* and *D-HC* accuracy because it represents their average).

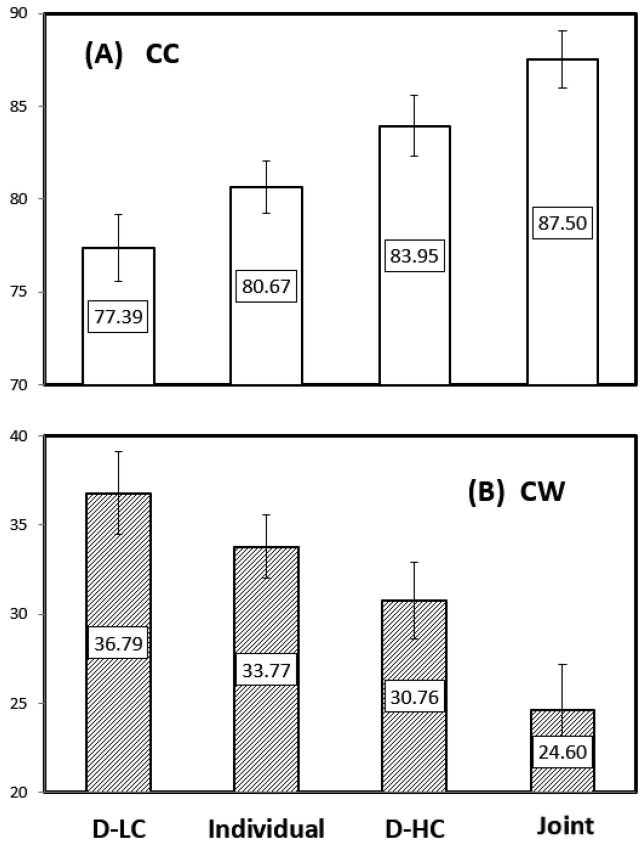

Figure 7. Mean accuracy for dummy-low confidence (*D-LC),* dummy-high confidence (*D-HC*), individual decisions and joint decisions across Experiments 1 and 2. The upper panel presents the results for consensually correct (CC) items whereas the lower panel presents the results for consensually wrong (CW) items.

Let us consider the results for the application of the MCS algorithm. First, it should be noted that in the analysis across all items, there was little difference between *D-LC* and *D-HC*: Their means were 60.20 and 61.80, respectively, $t(79) = 1.02$, $p < .32$, $d = 0.15$, so that confidence judgments were of little benefit as predictors of accuracy. This result parallels the finding of little difference in accuracy between individual and group decisions (which averaged 61.00 and 61.69, respectively, across the two experiments).

We turn next to the CC-CW comparison. Consistent with the simulation of Koriat (2012b), a two-way ANOVA comparing *D-HC* and *D-LC* for CC and CW items yielded a significant interaction, $F(1, 39) = 13.62$, $MSE = 232.80$, $p < .0005$, $\eta p^2 = .15$. For CC items, *D-HC* accuracy was higher than *D-LC* accuracy, $t(39) = 3.41$, $p < .001$. For CW items, in contrast, *D-HC* accuracy was worse than *D-LC* accuracy, $t(39) = 2.27$, $p < .05$. Thus, for CW items, better accuracy is achieved by selecting for each item the decision of the *less* confident member.

In comparing the simulation results to the actual accuracy of individual decisions, it can be seen that the selection of high-confidence judgments enhanced the mean accuracy of these judgments for CC items, $t(39) = 3.41$, $p < .05$, $d = .43$, but impaired the mean accuracy of these judgments for CW items (30.76 vs. 33.77), $t(39) = 2.27$, $p < .05$, $d = 0.17$. These results are also in line with Koriat's (2012b) application of MCS to virtual (rather than empirical) dyads, supporting the amplification pattern.

However, although *D-HC* outperformed the average accuracy of individual decisions, it did not outperform the accuracy of the better (*HP*) member for CC items as was the case in Koriat (2012b): *HP* averaged 86.50, which was significantly higher than *D-HC,* $t(79) = 2.31$, $p < .05$, $d = .19$. This difference may stem from the fact that the members of the virtual dyads were matched on accuracy and confidence in Koriat (2012b), which was not the case for the empirical dyads in this study (see General Discussion).

We turn finally to a comparison between the performance of the dummy participants and the actual joint decisions. For the CC items, actual dyadic decisions outperformed the accuracy predicted by MCS, $t(39) = 2.53$, $p < .05$, $d = 0.25$, and in fact, was practically the same as the accuracy of the member with the highest accuracy. In turn, for the CW items, the accuracy of the joint decision was significantly worse than what was predicted by MCS, $t(39) = 3.14$, $p < .005$, $d = 0.29$, and was almost the same as that demonstrated by the worse of the two members (*LP* = 27.75). Both of these observations suggest that the joint decision was influenced by other factors beyond degree of confidence, possibly factors that have to do with the content of the arguments that were raised within the group (see Trouche et al., 2014). The influence of these added factors, however, was not uniformly beneficial: It enhanced accuracy for CC items but impaired accuracy for CW items. Like the effect of the application of MCS, the added effect of dyadic interaction yielded an amplification pattern.

General Discussion

In this study, we compared accuracy and confidence for individual and joint decisions in two different tasks. The first task involved perceptual judgments for which all the information needed to reach a correct decision was available to both members of a dyad. The second task, in contrast, involved general knowledge. For this task, the information is possibly not equally available to both members of a dyad. In view of these differences, the overall similarity of the results across the two tasks is quite impressive. We will review and discuss the main findings across the two experiments.

The Distinction Between CC and CW Items

Research has largely confirmed the benefits that ensue from the use of statisticized groups: When judgments are averaged across several individuals, the results generally support the wisdom-of-crowds phenomenon, yielding more accurate judgments than those of the average individual. However, there has been some ambivalence regarding the relative benefits of group decisions versus individual decisions. Whereas judgments made by a group under the requirement to reach a consensual decision are generally better than those made by individuals, several results and discussions suggest that group interaction can be harmful to decision accuracy under some conditions (Janis, 1982; Larrick et al., 2012; Lorenz et al., 2011; Sunstein & Hastie, 2008, 2015). The present study, as well as Koriat's (2012b) simulation, brings to the fore a general moderator variable: The baseline accuracy achieved by individuals for a particular item. When individuals' decisions for a 2AFC nonintellective item are better than chance, as is generally the case for representative items, group deliberation is likely to be beneficial. In contrast, for the relatively rare CW items, for which individuals' decisions are below chance level, group discussion is liable to be detrimental to decision accuracy. The principle is consensual amplification (see Sunstein & Hastie, 2015, and see further below): Group discussion amplifies the trend that exists for the accuracy of individual performance. Whereas Koriat (2012b) demonstrated this principle for statisctized groups, when the MCS algorithm was applied to virtual dyads, the present study confirmed this principle for interacting groups. In both studies, two heads were better than one for CC items but worse than one for CW items. Note that in both experiments the analysis failed to yield an overall main effect for group decision. The interaction, however, was significant, revealing the differential effects of group discussion for the two types of items.

The distinction between CC and CW items emerged originally in the context of the confidence-accuracy relationship. Let us examine the results obtained in that context because their explanation provides some clues to the benefits and costs of group versus individual decisions. The question addressed by Koriat (1976, 2008b, 2011, 2012a) was: Why does subjective confidence in decisions predict the accuracy of these decisions? In attempting to answer that question, it turned out that the confidence/accuracy correlation is generally positive only because in many domains (e.g., general knowledge, psychophysical judgments, recognition memory), people's object-level accuracy is better than chance: For 2AFC questions, people's choices are more likely to be correct than wrong. However, when CW items were used, for which people's answers were more likely to be wrong than correct, the confidence/accuracy correlation was found to be negative. This pattern of results, which has been replicated across a number of domains, follows the consensuality principle (Koriat, 2008b): Confidence judgments are correlated with the consensuality of the answer—the likelihood that it would be chosen by most people, rather than with its accuracy.

How do these results bear on the question whether group decisions are more accurate or less accurate than individual decisions? This question can be addressed at two levels—at a surface, empirical level, and at a deeper, theoretical level. The surface level is exemplified by the MCS algorithm (Koriat, 2012b), which was based on two empirical findings. First, group decisions for a given issue are more strongly influenced by the members with higher confidence in their view on that issue. Second, as just noted, confidence is correlated positively with accuracy for CC items but negatively for CW items. The application of the MCS algorithm to virtual dyads (whose confidence judgments had been standardized) yielded very clear results. First, when items were drawn randomly from their domain so that they were representative of that domain, group decisions were more accurate than the best member of a dyad. This was also true for CC items (for which most participants chose the correct answer). In contrast, for CW items, group decisions were worse even than the worse member of a dyad. These results are quite impressive given that they were based on virtual dyads.

Note that the MCS algorithm exploits the observation (see Koriat, 2012a) that the confidence-consensuality correlation is obtained both within individuals across items, but also across individuals for a given item (Koriat et al., 2015). It also takes advantage of the observation that this correlation is also obtained across the decisions and confidence of the two members of a virtual dyad (dyadic gamma correlation, Koriat, 2012b): When the two members disagree, the member with higher confidence is more likely to be correct for CC items but wrong for the CW items.

At a deeper, theoretical level, the self-consistency model of subjective confidence provides a conceptualization that permits predictions regarding the accuracy of individual versus group decisions. According to SCM (Koriat, 2012a; Koriat & Adiv, 2011), in attempting to respond to a 2AFC item, participants draw a small sample of clues from a commonly shared population of clues associated with that item. They base their decision on the balance of evidence in favor of the two responses, and their confidence reflects the consistency with which the sampled clues support the chosen response.

Consider a set of representative (or CC) items. For these items, the majority of the clues in each item-related population is assumed to favor the correct answer by virtue of people's adaptation to the environment (Dhami et al., 2004; Herzog & Hertwig, 2013) and therefore, most participants will be likely to choose the correct answer. Importantly, however, because the size of the sample of clues underlying each decision is assumed to be quite small (see Koriat, 2012a), different people may reach the correct decision for different "reasons". What people with the same experience have in common is not necessarily the specific clues on which they base their answer, but the population of clues from which they draw the ingredients for the construction of their answer or decision in each occasion. It is this population of clues that embodies the distributed wisdom of crowds (Koriat & Sorka, 2015). Although this shared wisdom may be largely redundant across a group of participants

with the same experience, the sample of clues underlying different people's decisions may be nonredundant, so that the accessible information favoring the correct answer may be dispersed among different individuals.

In an interacting group of individuals, to the extent that different members bring in different considerations and clues in favor of the correct answers, the aggregated information may increase the accuracy of the group decisions beyond what would be predicted by the relative confidence of the individual members. Indeed, discussions of the wisdom-of-crowds phenomenon have emphasized the importance of independence and diversity between the members for the benefit that ensues from the aggregation of judgments across individuals (see Larrick et al., 2012; Lorenz et al., 2011; Surowiecki, 2005). Likewise, discussions of the wisdom of the inner crowd (Herzog & Hertwig, 2014; Hourihan & Benjamin, 2010; Steegen et al., 2014; Vul & Pashler, 2008) suggest that the benefits that ensue from averaging multiple estimates provided by the same person increase with the independence between these estimates. Therefore, although the application of the MCS algorithm to statisticized groups may produce greater accuracy than that achieved by individual members for CC items, interacting groups may yield greater accuracy still, to the extent that members bring in nonoverlapping sets of clues.

For the same reasons, CW items might be expected to yield worse performance for interacting groups than for confidence-based judgments under the MCS algorithm. CW items are assumed to be associated with populations of clues whose distribution is skewed in favor of the wrong answers. Therefore, it is the minority answers that are expected to be correct. To the extent that participants base their answer on nonoverlapping samples of clues, group discussion should be likely to result in the adoption of the wrong answer even more so than would be expected by the relative confidence of the group members. Thus, in the case of CW items, the pooling of information that members of a group collectively bring in is liable not only to perpetuate errors but also to amplify them.

It should be stressed that reliance on the distributed wisdom of crowds may occasionally yield incorrect answers for CC items and correct answers for CW items (e.g., Koriat, 2008b, 2011; see Koriat, 2012b). This may occur even for the same person when that person is presented several times with the same question (Koriat, 2011, 2012b). However, in many domains the crowd is "wise" in the sense that reliance on the distributed wisdom is more likely to yield correct answers than wrong answers to the majority of 2AFC questions.

Confidence-Based and Argument-Based Contributions to Accuracy

The foregoing discussion helps distinguish between the contributions of confidence-based factors and argument-based factors to the accuracy of group decisions. These factors were pitted against each other in the study of Trouche et al. (2014). Their study, however, focused on intellective tasks for which there exist demonstrably correct answers. Although such was not the case for the tasks used in the present study, the results suggest that the exchange of arguments between the members of a group did affect the accuracy of the joint decisions beyond what was predicted by the prediscussion confidence of the members.

Clearly, the contribution of confidence-based decisions cannot be separated from that of argument-based decisions because subjective confidence is based in part on arguments and considerations (Brewer & Sampaio, 2012; Trouche et al., 2014; Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980). However, we can assess the potential contribution of preinteraction confidence to decision accuracy. The application of the MCS algorithm to the preinteraction individual judgments yielded a similar pattern of results to what was demonstrated by Koriat (2012b): For CC items, pooling high-confidence decisions across the members of a dyad improved accuracy beyond the average accuracy of the two members, whereas for CW items it resulted in lower accuracy than the average accuracy of the two members.

Two observations suggest that confidence-based selection of decisions may have contributed in part to the differential effects of group discussion on decision accuracy for CC and CW items. The first is that the individual responses yielded a pattern consistent with the consensuality principle: For CC items, confidence was higher for correct answers than for wrong answers, but the opposite was true for CW items. The second observation is that when the individual decisions of the two members of a dyad differed for a particular item, the joint decision was more likely to follow the decision of the member with higher confidence on that item.

Several results, however, suggest an added contribution of group deliberation to performance. First, whereas the MCS simulation assumed that the joint decision is dominated entirely by the more confident member, only 58% (in Experiment 1) and 65% (in Experiment 2) of the joint decisions followed the decision of the high-confidence participant when the two members disagreed. Second, in 10% of the trials in Experiment 2, the joint decision differed from *both* of the individual decisions, suggesting that group discussion led both members to change their mind. Finally, as the results presented in Figure 7 suggest, the joint discussion affected accuracy beyond what would be expected from the confidence of the members in their individual decisions.

These results accord with the claim of Trouche et al. (2014) that the exchange of arguments within a group contributes to performance beyond the confidence of the group members. This claim was made specifically with regard to intellective tasks. For these tasks, it was argued that "the scheme that best describes group performance is "truth wins"" (p. 1958). Such is clearly untrue of the tasks used in the present study. Although group interaction enhanced further the accuracy of the joint decision beyond confidence for CC items, it also impaired further the accuracy of joint decisions for CW items. Possibly faulty arguments in favor of the wrong answer for these items can be quite convincing. As was argued by Dunning, Johnson, Ehrlinger, and Kruger (2003), the skills needed to produce correct responses are virtually identical to those needed to evaluate the accuracy of one's responses. Thus, for CW items, group discussion not only failed to mitigate the errors exhibited by individuals, but actually amplified them. The results on the whole suggest that confidence-based selection of responses (the MCS algorithm) and group discussion affect decision accuracy in the same direction, amplifying the trend that characterizes individual decisions.

Examination of the Results for the Application of the MCS Algorithm

Several recent studies compared the success of the MCS algorithm with that of collaborative interaction. The results suggested two moderator variables that may affect the success of MCS relative to actual interaction: the similarity between the two members of a dyad in their overall performance, and the effectiveness of the communication between them. Let us examine our results in light of these findings.

In the task used by Bahrami et al. (2010), participants decided which of two briefly presented visual stimuli contained an oddball target. The items were presumably representative of their domain. Dyadic decisions yielded a 2HBT1 effect, but the benefit from dyadic interaction varied with the similarity between the members of a dyad in the accuracy of their perceptual decisions: For similar dyad members, two heads were better than one, whereas for dissimilar dyad members, two heads were worse than the better of the two members.

Using the same visual discrimination task, Bang et al. (2014) compared the accuracy that followed from the MCS algorithm (and from a similar, Minimum Reaction Time Slating algorithm) with that of dyadic joint decisions. For dyads in which the two members exhibited similar accuracy, actual interaction had no added benefit over the benefit that ensued from the MCS algorithm, whereas for dissimilar dyad members, actual interaction had an added benefit. It was proposed that, actual interaction is particularly important in the case of dissimilar members because the members can overrule the decision of the more confident but less competent member (see also Trouche et al., 2014). Similar differences were obtained for a numerosity task by Massoni and Roux (2012) who divided dyads in terms of their similarity in calibration (the discrepancy between confidence and accuracy) rather than in terms of their overall accuracy.

These results may explain why the success of the MCS algorithm was more limited in the present study than in Koriat (2012b). In the present study, as noted earlier, for CC items, *D-HC* was found to outperform the average accuracy of the two members but not the accuracy of the better of them as had been found in Koriat (2012b). The difference between the results of the two studied may be due to the fact that in the latter study, the two members of virtual dyads were matched on accuracy. To examine this possibility, we formed *virtual* dyads (regardless of their empirical pairing), for which the members of each dyad were matched in terms of their accuracy across all items in each experiment. Their confidence judgments were also standardized to nullify chronic differences between them in confidence. The application of the MCS algorithm to these dyads, however, yielded results very similar to those obtained for the empirical dyads (see Figure 7). In particular, for CC items, mean accuracy for *D-HC* was 84.02 for the virtual dyads, which was still lower than the accuracy of the empirical joint decisions (87.50), but similar to what was found for *D-HC* when computed for the empirical dyads (83.95, see Figure 7).

We also conducted several analyses to examine possible contributions of the similarity between the members (in confidence and/or accuracy) to the accuracy of the joint decisions. The results, which will not be reported here, were not conclusive. We suspect that detailed analyses of this contribution requires a larger number of observations per dyad than was available in the present study (in the present study, there were 37 trials across the two experiments, whereas the analyses of Bang et al., 2014, were based on 256 trials, and those of Massoni & Roux, 2012, were based on 150 trials). It should be noted that in a recent study, Hautz et al. (2015), who presented medical students with medical data on 6 clinical cases, found higher diagnostic accuracy for those who worked in pairs than for those who worked alone, although the application of the MCS algorithm to their individual confidence judgments did not yield better accuracy than individual performance. Perhaps the more limited success of MCS in the present study than in Koriat (2012b) stems from the small number of items used in the present study.

Another contributing factor to the accuracy of dyadic decisions is the quality of the communication between the two members. A linguistic analysis (Fusaroli et al., 2012) of the conversations in Bahrami et al.'s (2010) study indicated that the more the dyad members converged on a shared set of expressions of confidence, the higher was the benefit they achieved from cooperation. Our intention in recording the conversations between the members was to examine the possibility that the quality of the communication within a dyad will be correlated with improved accuracy of the joint decisions only for CC item, whereas for CW items it will be correlated with inferior accuracy. The conversations, however, were quite rich, possibly much richer than those in Fusaroli et al. (2012), and their analyses turned out to be more complex than we had anticipated. The results might be reported at some later time. Clearly, the analysis of the dynamics of the interaction within a dyad can provide useful clues to the processes underlying consensual amplification.

As has been noted by several authors (e.g., Minson & Mueller, 2012), group deliberation is time consuming and expensive. Therefore it is important to seek methods for exploiting statisticized groups to improve the accuracy of collective decisions (see Mannes, Soll, & Larrick, 2014). MCS incorporates a frugal heuristic that proved effective in improving decision accuracy for typical, CC items (Hertwig, 2012). It can be easily applied to virtual groups of different sizes, and like other methods that rely on statisticized groups, is not susceptible to some of the problems involved in social interaction, such as conformity pressures, herding, informational cascade, or social loafing (see Karau & Williams, 1993; Sunstein & Hastie, 2015; Surowiecki, 2005). Although it was found to be less effective for CC items than collaborative decisions, it is important to study closely the factors that affect its success.

The Effects of Dyadic Interaction on Confidence Judgments

Our results indicated that dyadic interaction also enhanced confidence in the joint decision. These results are consistent with previous findings. Three observations, however, are of particular interest. First, dyadic interaction enhanced confidence in the joint decision not only for CC items but also for CW items. The enhanced confidence for CW items is surprising; it might have been expected that for these items group interaction should at least raise some doubts about the chosen answer, but this is not what happened. Rather, participants felt even more confident after converging on the wrong decision than they had been prior to group interaction. In fact, for CC items, dyadic interaction improved accuracy and also enhanced confidence. For CW items, in contrast,

it impaired accuracy while enhancing confidence, thus strengthening further the illusion of validity for these items (see Kahneman, 2011). This pattern was observed in both experiments. Heath and Gonzalez (1995), who found the interaction with others to increase individuals' confidence even when it did not enhance decision accuracy, proposed that group interaction forces people to explain their choices to others, and it is explanation generation that results in enhanced confidence.

Second, in Experiment 1, dyadic interaction enhanced confidence in the joint decision even when the two members initially disagreed. It also enhanced the members' confidence in the joint decision even when that decision was wrong. This was true for both the CC and CW items in Experiment 1, and for the CW items in Experiment 2.

Finally, it is interesting that the consensuality principle was observed not only for individual decisions but also for the joint decisions. For CC items, the confidence of the members in the joint decision was higher for correct decisions than for wrong decisions, whereas for CW items the opposite pattern was observed.

The results obtained for CW items may have deplorable consequences given people's tendency to rely on confidence in translating their beliefs into action (Koriat & Goldsmith, 1996). Because for CW items, group discussion resulted in reduced accuracy coupled with enhanced confidence in the joint decision, this may increase the likelihood of groups acting on the wrong decisions. This should be even more so given that for these items it was the wrong decision that was associated with higher confidence.

Lorenz et al. (2011), who obtained repeated estimates from groups of participants, found social influence to trigger the convergence of individual estimates and to boost participants' confidence after convergence despite lack of improved accuracy. They argued that the boost in confidence subverts the wisdom of crowd effect psychologically, leading to the false belief of collective accuracy as a result of convergence. Yaniv, Choshen-Hillel, and Milyavsky (2009) also observed that the revision of one's opinion on the basis of opinions obtained from others may lead decision makers to experience greater confidence in their less accurate judgments. These results suggest that confidence is influenced by the consistency of the information sampled from the outside world (others' opinions) in the same way that it is influenced by the consistency of the clues sampled from one's own memory.

Why Can Two Heads Be Worse Than One? The Case of CW Items

CW items illustrate one condition in which group interaction may be detrimental. Sunstein and Hastie (2015) reviewed several observations suggesting that group deliberations can reinforce and exacerbate individuals' biases and errors. They proposed that if most group members fall prey to some of the well-known cognitive biases (see Kahneman, 2011), others in the group may tend to make the same errors either because of the informational signals that they receive from them or because of reputational (conformity) pressures. The dynamics of social interaction may also aggravate biases and errors through such processes as herding—the alignment of thoughts or behaviors of individuals in a group through local interactions (Raafat, Chater, & Frith, 2009) or through informational cascade, as when group members follow the views of those who spoke first.

Indeed, unlike the results that have been typically obtained for some of the intellective problems (see Trouche et al., 2014), several studies that focused on specific judgmental biases indicated that group deliberation is liable to amplify the error exhibited by individuals. For example, group deliberation was found to aggravate the planning fallacy (Buehler, Griffin, & Peetz, 2010), leading groups to make even less realistic predictions than individuals. Groups also evidenced a stronger sunk-cost effect (Whyte, 1993), and dyads were more reluctant than individuals working alone to revise their judgments (Minson & Mueller, 2012).

The idea of group amplification of individual tendencies has been invoked primarily in connection with erroneous responses. The view endorsed in the present article, in contrast, assumes that consensual amplification underlies the benefits as well as the costs of group deliberation. The processes that underlie the choice of correct responses for CC items and confidence in these responses are the same as those that underlie the choice of wrong responses for CW items and confidence in these responses (Koriat, 2012a). This view is consistent with the core assumption of the metamemory theory 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 nondeceptive items" (Brewer & Sampaio, 2012, p. 68). For both CC and CW items, the confidence-based selection of responses (the MCS algorithm) and group interaction amplify the trend that is exhibited by individual decisions. Possibly, the amplification pattern would be expected to be stronger still for larger groups than for dyads, because of the stronger operation of social pressures (see Koriat et al., 2015). Note that the amplification pattern observed in this study is consistent with the phenomenon of group polarization: Several studies indicated that group deliberation can lead the members of a group to adopt a more extreme version of their predeliberation positon (e.g., Moscovici & Zavalloni, 1969; see Sunstein & Hastie, 2015).

In discussing the processes that may lead to error amplification, researchers mentioned the idea that groups often focus on "what everybody knows", emphasizing broadly shared information while neglecting information that is held by one or a few members (Lightle et al., 2009; Stasser & Titus, 1985; Sunstein & Hastie, 2015). Actually, we assume that this process operates even within a dyad. According to SCM, people base their choices and confidence on clues and considerations that are retrieved from a commonly shared pool. Group deliberation amplifies the contribution of the shared clues, whether these clues favor the correct decision or the wrong decision. Because subjective confidence is assumed to correlate with the consensuality of the decision, a similar pattern of consensual amplification should emerge when high-confidence responses are pooled across the members of virtual dyads, or when responses are simply aggregated across a group of participants.

A question that suggests itself, however, is whether the clues and considerations that emerge in group discussion might not be different from those underlying individual decisions, and hence underlying the effectiveness of the MCS algorithm or the wisdom-of-crowds phenomenon. We might expect the arguments that are raised in group deliberations to be more concrete and analytic than the type of mnemonic cues that have been assumed to underlie metacognitive judgments (Benjamin & Bjork, 1996; Kelley & Lindsay, 1993; Koriat, 1997). In their study on gambling behavior in sports, Simmons, Nelson, Galak, and Frederick (2011) argued that reliance on emotional, intuitive responses as against more

rational, deliberative responses (see Denes-Raj & Epstein, 1994; Kahneman & Frederick, 2002) sometimes leads the crowd to be systematically biased and ultimately unwise. Koriat (2008a), in turn, suggested that group discussion can produce a shift in mode of reasoning, causing members to engage in an analytic process that helps members overcome the biases entailed in experience-based judgments. His study, which examined several methods for alleviating the strong tendency to overestimate conditional predictions (the assessed probability that a certain outcome will occur given a certain condition, see Koriat, Fiedler, & Bjork, 2006), found dyadic interaction to be the most effective method. He proposed that the attempt to convince each other activates a mode of reasoning in which an appeal is made to rational, verbalizable considerations. This idea invites investigations that link the work on group decisions to dual-process theories (Kahneman, 2011). If indeed group discussion induces a change in mode of reasoning, this would imply that the effectiveness of group deliberation over MCS may derive in part from a shift in the quality of the clues and arguments that underlie joint decisions in comparison with those that underlie individual decisions. Note, however, that this shift, if it does occur, does not guarantee increased accuracy, but only results in stronger consensual amplification.

A final comment concerns the practical implications of this study. A great deal of work has been carried out on the wisdom-of-crowds phenomenon and on the value of group decisions in comparison with individual decisions. Much of that work was motivated by practical concerns (Sunstein & Hastie, 2015). So what practical conclusions might a reader draw from the results presented in this article?

Clearly, in many domains the wisdom of crowds converges on the correct decisions for the majority of 2AFC questions. Therefore, MCS, as well as collective decisions based on statisticized or interacting groups are likely to outperform individual decisions. The problem exists for domains and issues for which there is insufficient knowledge and a great deal of uncertainty. However, even for well-treaded domains, the knowledge, skills, and heuristics that have been adapted to converge largely on the correct decisions, may lead people astray for some unrepresentative questions. Can groups spot questions that are "tricky" or "deceptive"? If so, would they then be able to bring themselves to adopt the minority opinion, perhaps even taking the advice of the *least* confident member? Unfortunately, it seems that groups will not have to face that challenge. Our ongoing research suggests that people fail to discriminate between CC-type and CW-type items even when they are warned that some of the items lead most people to choose the wrong answer (see Brewer & Sampaio, 2012).

References

Aramovich, N. P., & Larson, J. R. (2013). Strategic demonstration of problem solutions by groups: The effects of member preferences, confidence, and learning goals. *Organizational Behavior and Human Decision Processes, 122,* 36–52. http://dx.doi.org/10.1016/j.obhdp.2013.04.001

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., . . . Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology, 6,* 130–147. http://dx.doi.org/10.1037/1076-898X.6.2.130

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). New York: Kluwer. http://dx.doi.org/10.1007/978-0-306-47630-3_19

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329,* 1081–1085. http://dx.doi.org/10.1126/science.1185718

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y., . . . Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition, 26,* 13–23. http://dx.doi.org/10.1016/j.concog.2014.02.002

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.

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: Lawrence Erlbaum.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics, 54,* 75–81. http://dx.doi.org/10.3758/BF03206939

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory (Hove, England), 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,* 59–67. http://dx.doi.org/10.1016/j.jml.2012.04.002

Buehler, R., Griffin, D., & Peetz, J. (2010). Chap. one-the planning fallacy: Cognitive, motivational, and social origins. *Advances in Experimental Social Psychology, 43,* 1–62.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5,* 559–583. http://dx.doi.org/10.1016/0169-2070(89)90012-5

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior, 12,* 41–55. http://dx.doi.org/10.1007/BF01064273

Dalkey, N. C. (1969). An experimental study of group opinion. *Futures, 1,* 408–426. http://dx.doi.org/10.1016/S0016-3287(69)80025-X

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology, 66,* 819–829. http://dx.doi.org/10.1037/0022-3514.66.5.819

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.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83–87. http://dx.doi.org/10.1111/1467-8721.01235

Esser, J. K. (1998). Alive and well after 25 years: A review of groupthink research. *Organizational Behavior and Human Decision Processes, 73,* 116–141. http://dx.doi.org/10.1006/obhd.1998.2758

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

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological Science, 23,* 931–939. http://dx.doi.org/10.1177/0956797612436816

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3,* 20–29. http://dx.doi.org/10.1111/j.1745-6916.2008.00058.x

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

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

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

Hautz, W. E., Kämmer, J. E., Schauber, S. K., Spies, C. D., & Gaissmaier, W. (2015). Diagnostic performance by medical students working individually or in teams. *Journal of the American Medical Association, 313,* 303–304. http://dx.doi.org/10.1001/jama.2014.15770

Heath, C., & Gonzalez, R. (1995). Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organizational Behavior and Human Decision Processes, 61,* 305–326. http://dx.doi.org/10.1006/obhd.1995.1024

Hertwig, R. (2012). Psychology. Tapping into the wisdom of the crowd—With confidence. *Science, 336,* 303–304. http://dx.doi.org/10.1126/science.1221403

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20,* 231–237. http://dx.doi.org/10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2013). The ecological validity of fluency. In C. Unkelbach & R. Greifeneder (Eds.), *The experience of thinking: How the fluency of mental processes influences cognition and behavior* (pp. 190–219). New York: Psychology Press.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences, 18,* 504–506. http://dx.doi.org/10.1016/j.tics.2014.06.009

Hill, G. W. (1982). Group versus individual performance: Are N+1 heads better than one? *Psychological Bulletin, 91,* 517–539. http://dx.doi.org/10.1037/0033-2909.91.3.517

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1068–1074. http://dx.doi.org/10.1037/a0019694

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

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

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

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgments* (pp. 49–81). New York: Cambridge University Press.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65,* 681–706. http://dx.doi.org/10.1037/0022-3514.65.4.681

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

Kerr, N. L., & Tindale, R. S. (2011). Group-based forecasting? A social psychological analysis. *International Journal of Forecasting, 27,* 14–40. http://dx.doi.org/10.1016/j.ijforecast.2010.02.001

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. (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. (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. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370. http://dx.doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2008a). Alleviating inflation of conditional predictions. *Organizational Behavior and Human Decision Processes, 106,* 61–76. http://dx.doi.org/10.1016/j.obhdp.2007.08.007

Koriat, A. (2008b). 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). 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., & Adiv, S. (2011). The construction of attitudinal judgments: Evidence from attitude certainty and response latency. *Social Cognition, 29,* 577–611. http://dx.doi.org/10.1521/soco.2011.29.5.577

Koriat, A., Adiv, S., & Schwarz, N. (2015). Views that are shared with others are expressed with greater confidence and greater fluency independent of any social influence. *Personality and Social Psychology Review*. Advance online publication. http://dx.doi.org/10.1177/1088868315585269

Koriat, A., Fiedler, K., & Bjork, R. A. (2006). Inflation of conditional predictions. *Journal of Experimental Psychology: General, 135,* 429–447. http://dx.doi.org/10.1037/0096-3445.135.3.429

Koriat, A., & Goldsmith, M. (1996). 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., 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., & Sorka, H. (2015). The construction of categorization judgments: Using subjective confidence and response latency to test a distributed model. *Cognition, 134,* 21–38. http://dx.doi.org/10.1016/j.cognition.2014.09.009

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

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology

of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers of social psychology, social judgment and decision making* (pp. 227–242). Philadelphia: Psychology Press.

Laughlin, P. R. (2011). *Group problem solving*. Princeton, NJ: Princeton University Press. http://dx.doi.org/10.1515/9781400836673

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22,* 177–189. http://dx.doi.org/10.1016/0022-1031(86)90022-3

Lightle, J. P., Kagel, J. H., & Arkes, H. R. (2009). Information exchange in group decision making: The hidden profile problem reconsidered. *Management Science, 55,* 568–581. http://dx.doi.org/10.1287/mnsc.1080.0975

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, USA of the United States of America, 108,* 9020–9025. http://dx.doi.org/10.1073/pnas.1008636108

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107,* 276–299. http://dx.doi.org/10.1037/a0036677

Massoni, S., & Roux, N. (2012). *Optimal group decision: A matter of confidence calibration*. Retrieved from. https://samm.univ-paris1.fr/IMG/pdf/MR_Group.pdf

Mata, A., & Almeida, T. (2014). Using metacognitive cues to infer others' thinking. *Judgment and Decision Making, 9,* 349–359.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34,* 57–74. http://dx.doi.org/10.1017/S0140525X10000968

Mercier, H., & Sperber, D. (2012). "Two heads are better" stands to reason. *Science, 336,* 979. http://dx.doi.org/10.1126/science.336.6084.979-a

Minson, J. A., & Mueller, J. S. (2012). The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psychological Science, 23,* 219–224. http://dx.doi.org/10.1177/0956797611429132

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology, 12,* 125–135. http://dx.doi.org/10.1037/h0027568

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning, 4,* 231–248. http://dx.doi.org/10.1080/135467898394148

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the collective dynamics of opinion formation. *PLoS ONE, 8,* e78433. http://dx.doi.org/10.1371/journal.pone.0078433

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

Nussbaum, E. M. (2008). Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology, 33,* 345–359. http://dx.doi.org/10.1016/j.cedpsych.2008.06.001

Pansky, A., & Goldsmith, M. (2014). Metacognitive effects of initial question difficulty on subsequent memory performance. *Psychonomic Bulletin & Review, 21,* 1255–1262. http://dx.doi.org/10.3758/s13423-014-0597-2

Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences, 13,* 420–428. http://dx.doi.org/10.1016/j.tics.2009.08.002

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

Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. (2011). Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *The Journal of Consumer Research, 38,* 1–15. http://dx.doi.org/10.1086/658070

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

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology, 48,* 1467–1478. http://dx.doi.org/10.1037/0022-3514.48.6.1467

Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology, 5,* 786. http://dx.doi.org/10.3389/fpsyg.2014.00786

Sunstein, C. R., & Hastie, R. (2008). Four failures of deliberating groups. *University of Chicago Law & Economics, Olin Working Paper,* (401).

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

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor Books.

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

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General, 143,* 1958–1971. http://dx.doi.org/10.1037/a0037099

Tversky, B. (1981). Distortions in memory for maps. *Cognitive Psychology, 13,* 407–433. http://dx.doi.org/10.1016/0010-0285(81)90016-5

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19,* 645–647. http://dx.doi.org/10.1111/j.1467-9280.2008.02136.x

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10,* 243–268. http://dx.doi.org/10.1002/(SICI)1099-0771(199709)10:3<243::AID-BDM268>3.0.CO;2-M

Whyte, G. (1993). Escalating commitment in individual and group decision making: A prospect theory approach. *Organizational Behavior and Human Decision Processes, 54,* 430–455. http://dx.doi.org/10.1006/obhd.1993.1018

Williams, E. F., Dunning, D., & Kruger, J. (2013). The hobgoblin of consistency: Algorithmic judgment strategies underlie inflated self-assessments of performance. *Journal of Personality and Social Psychology, 104,* 976–994. http://dx.doi.org/10.1037/a0032416

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

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93,* 1–13. http://dx.doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I. (2011). Group diversity and decision quality: Amplification and attenuation of the framing effect. *International Journal of Forecasting, 27,* 41–49. http://dx.doi.org/10.1016/j.ijforecast.2010.05.009

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 558–563. http://dx.doi.org/10.1037/a0014589

Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology, 33,* 345–366. http://dx.doi.org/10.1006/jesp.1997.1326

Received February 26, 2015
Revision received June 1, 2015
Accepted June 3, 2015 ■