When Are Two Heads Better than One and Why?

Asher Koriat

Science **336**, 360 (2012);
DOI: 10.1126/science.1216549

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

When Are Two Heads Better than One and Why?

Asher Koriat[1]*

A recent study, using a perceptual task, indicated that two heads were better than one provided that the members could communicate freely, presumably sharing their confidence in their judgments. Capitalizing on recent work on subjective confidence, I replicated this effect in the absence of any dyadic interaction by selecting on each trial the decision of the more confident member of a virtual dyad. However, because subjective confidence monitors the consensuality rather than the accuracy of a decision, when most participants were in error, reliance on the more confident member yielded worse decisions than those of the better individual. Assuming that for each issue group decisions are dominated by the more confident member, these results help specify when groups will be more or less accurate than individuals.

Many everyday decisions are made jointly by two or more people. Studies that compared the accuracy of individual and group decisions have yielded somewhat inconsistent results. The groupthink phenomenon, which refers to a mode of decision-making that occurs within a cohesive group, has been claimed to underlie some of the disastrous decisions made in U.S. history (*1*–*3*). However, several studies have indicated that cooperative groups perform better than independent individuals on a wide range of tasks (*4*, *5*).

The immediate motivation for the present work comes from a recent study that delivers encouraging news. In the article *Optimally Interacting Minds* (*6*), Bahrami *et al*. compared individual and dyadic performance in a simple visual task. They asked, "[H]ow [can] signals from the same sensory modality (vision) in the brains of two different individuals ... be combined through social interaction?" In their experiments, participants judged which of two briefly presented stimuli contained an oddball target. Participants worked in dyads; they first made their decision individually, then shared their decisions, and if they disagreed, they discussed the matter until they reached a joint decision. The results led to the conclusion that "for two observers of nearly equal visual sensitivity, two heads were definitely better than one provided they were given the opportunity to communicate freely." In discussing the mechanism responsible for the two-heads-better-than-one (2HBT1) effect, the authors assumed that each individual can monitor the accuracy of his or her performance and can communicate his or her confidence accurately to the other member.

The present proposal, which capitalizes on recent work on the bases of subjective confidence (*7*–*9*), can account for the 2HBT1 effect in the absence of any communication between the members. On the one hand, this work, along with the results of (*6*), suggests a useful algorithm for combining judgments across two people who operate individually. In this algorithm—maximum-confidence slating (MCS)—for each trial, the decision that is made with higher confidence by one member of a virtual dyad is selected, circumventing dyadic interaction altogether. According to the self-consistency model (SCM) of subjective confidence (*7*), the MCS algorithm should yield a 2HBT1 effect, allowing a decision-maker to reach better decisions by combining information across a group of noninteracting participants. On the other hand, SCM implies boundary conditions to the group benefit so that under some conditions, two heads should be substantially worse than one.

SCM addressed the question of when confidence judgments are diagnostic of accuracy and why. In many two-alternative forced-choice (2AFC) tasks, a relatively high within-person confidence-accuracy (C/A) correlation is typically observed across items, suggesting that participants can monitor the accuracy of their choices. In attempting to clarify how people know that they know (*10*), it was noted (*11*) that in all previous studies of the C/A correlation, participants were more often correct than wrong, so that the correct answer was also the consensual or popular answer. When correctness and consensuality were dissociated by including a large number of items for which most participants chose the wrong answer, confidence was clearly correlated with the consensuality of the answer rather than with its correctness: For consensually correct (CC) items, the C/A correlation was positive, whereas for consensually wrong (CW) items, it was consistently negative. This consensuality principle has now been confirmed for word-matching, feeling-of-knowing judgments, general-information questions, memory of studied sentences, and perceptual judgments (*7*, *8*, *11*–*14*). Although in none of these studies were participants informed about others' choices, their confidence correlated strongly with the proportion of other participants who made that choice, not with the correctness of the choice.

SCM attempted to account for these and other findings. The model is described elsewhere (*7*). It assumes that people's responses to a 2AFC item are based on the random sampling of cues and representations associated with the item. An individual's subjective confidence, much like statistical level of confidence, is based on the consistency with which the decision reached is favored across the sampled representations. Assuming that the population of potential representations associated with each item is commonly shared across individuals, it was shown that a random sampling of representations is bound to yield higher confidence for consensual than for nonconsensual decisions (*7*–*9*).

The implications for the 2HBT1 effect are twofold. First, in many situations the knowledge that is shared by all participants corresponds by and large to the truth, so that the MCS algorithm as well as social interaction are expected to yield decisions that are more accurate than those of each individual alone. Thus, MCS is expected to yield a 2HBT1 effect for many perceptual and general-knowledge tasks in which the items are representative of their domain. Indeed, the wisdom-of-crowds phenomenon suggests that information that is aggregated across participants is generally closer to the truth than is the information provided by each individual participant (*15*–*18*).

The second implication, however, is that if confidence is tuned to the "common knowledge" rather than to the truth, reliance on confidence can be misleading when the shared knowledge departs from the truth. The psychological literature is replete with documentations of situations in which participants' perceptions, judgments, and beliefs deviate consistently from the truth (*19*, *20*). Examples include perceptual and memory illusions, deceptive general-knowledge questions, reconstructive memory errors, illusory truth judgments, and various judgmental biases. Assuming that collec-

Table 1. Percentage of correct decisions. In study 1, participants decided which of two displays contained an oddball target. In study 3, participants decided which of two lines was longer (Lines) or which of two shapes had a larger area (Area).

		HP (%)	*LP* (%)	*D-HC* (%)	*D-LC* (%)
		Study 1			
Oddball target		67.82	66.98	69.88	64.93
		Study 3			
Lines	CC	81.58	80.59	85.03	77.14
	CW	25.00	26.31	17.10	34.21
Shapes	CC	83.33	84.58	86.67	81.25
	CW	28.13	24.06	22.50	29.69

[1]Department of Psychology and Institute of Information Processing and Decision Making, University of Haifa, Haifa 31905, Israel.

*To whom correspondence should be addressed. E-mail: akoriat@research.haifa.ac.il

tive decisions are dominated by the most confident members (*6*, *21*, *22*), these decisions (and the MCS algorithm) should yield performance in these cases inferior to the performance of the individual members. For example, the results of (*11*) suggest that when two participants decide whether the capital of Norway is Copenhagen or Oslo, the more confident participant is the more likely to be correct, but when they decide whether the capital of Ausralia is Canberra or Sydney, the more confident member is more likely to be wrong.

In studies 1 to 4, participants responded individually to 2AFC questions and indicated their confidence in each response. To nullify chronic individual differences in confidence (*23*), the confidence judgments were first standardized so that the mean and SD of all participants were equal to the mean and SD of the raw scores. Participants were then paired ad hoc to form virtual dyads. The member with higher percent correct was designated as high performing (*HP*), and the other as low performing (*LP*). In addition, two dummy participants were created as follows: 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. Percent accuracy was then calculated for the four "participants." The performance of the *D-HC* participant represents the MCS algorithm (*24*).

In studies 1 and 2, the 2AFC items were selected to be representative of their domains. Hence, participants' decisions are more likely to be correct than wrong, and the MCS algorithm is expected to yield better decisions than those of the better individual.

Study 1 used the same task as in (*6*), but participants performed the entire experiment individually and indicated their confidence in each response on a 50-to-100% scale. The stimuli were selected randomly from all possible combinations, but the same pairs were presented to all participants in the same random order. 38 participants were paired ad hoc on the basis of their percent correct to form 19 virtual dyads. Mean percentage of correct answers was calculated for each of the four "participants" (Table 1).

First, accuracy was higher for *D-HC* than for *D-LC* [$t(18) = 7.52$, $P < 0.0001$], indicating a positive C/A correlation in a between-individual comparison. Second, performance was more accurate for *D-HC* than for *HP* [$t(18) = 6.69$, $P < 0.0001$], supporting the 2HBT1 effect. The superiority of *D-HC* performance was observed for 18 out of the 19 dyads ($P < 0.0001$) by a binomial test. Last, percent correct was significantly lower for *D-LC* than for *LP* [$t(18) = 6.69$, $P < 0.0001$]. The same pattern was observed when the raw (rather than standardized) confidence judgments were used.

I also examined whether three heads are better than two by adding one participant to each dyad and selecting the response of the most confident participant in the triplet. The performance of the triplet *D-HC* (71.88%) was better than that of the dyadic *D-HC* (69.88%) [$t(18) = 5.05$, $P < 0.0001$].

Study 2 (*24*) replicated the results of study 1 by using a general-knowledge task in which participants judged which of two European countries has a larger area or a larger population. The stimuli were constructed to be representative of their domain (*25*). The results indicated that performance was more accurate for *D-HC* than for *HP* for both the area task and the population task. For both tasks, the results also indicated that three heads were better than two.

The results of studies 1 and 2 yielded a 2HBT1 effect in the absence of interaction between the members of the dyad. This effect derived from confidence being diagnostic of accuracy even in a between-individual analysis: When the members of a dyad disagreed, the decision associated with higher confidence was the more likely to be correct.

I then turned to tasks for which "common knowledge" did not always correspond to the truth. SCM predicts that when participants' decisions are incorrect by and large, the selection of the high-confidence decisions should yield inferior accuracy than that of each individual alone.

Study 3 was based on a reanalysis of the data from two experiments that tested basic predictions of SCM for perceptual judgments (*8*). In experiment 1, participants decided which of two irregular lines was longer (Lines), and in experiment 2, participants decided which of two shapes (Shapes) had a larger area. A deliberate attempt was made to include a sufficiently large number of CW pairs—pairs for which the majority of participants are likely to choose the wrong answer.

The items were classified as CC or CW according to whether most participants chose the correct or the wrong answer (Fig. 1). Nineteen and 20 virtual dyads were formed for the Lines and Shapes tasks, respectively. The results (Table 1) indicated a different pattern for the CC and CW items. For the CC items, *D-HC* exhibited the best performance, which is consistent with the 2HBT1 effect. For the CW items, in contrast, *D-HC* exhibited the worst performance of all "participants." This pattern was observed for each of the two tasks.

A two-way analysis of variance (ANOVA) across the two tasks, which compared *D-HC* with *HP* for CC and CW items, yielded $F_{1,38} = 10.49$, mean squared error (MSE) = 95.26, $P < 0.005$ for the interaction. For the CC items, performance was better for *D-HC* (85.87%) than for *HP* (82.48%) [$t(38) = 2.82$, $P < 0.01$]. In contrast, for the CW items percent correct was lower for *D-HC* (19.87%) than for *HP* (26.60%) [$t(38) = 2.86$, $P < 0.01$].

For the CW items, *D-HC* performance was worse than that of the worst participant (*LP*) (25.16%) [$t(38) = 2.20$, $P < 0.05$]. For these items, the best accuracy was achieved by *D-LC* (31.89%), so that a 2HBT1 effect can be obtained if the responses of the participant with lower confidence are selected.

Study 4 was based on a reanalysis of the results of (*11*), which included CC and CW 2AFC general-knowledge items. For the CC items, *D-HC* exhibited the best performance, whereas for the CW items, it exhibited the worst performance (*24*).

For an external observer who cannot tell CW from CC items (cannot tell whether the consensual or high-confidence answer is right or wrong), applying the same heuristic across the board (for example, "take the high-confidence response"), or relying on dyadic decisions, can generally be beneficial. However, if the "crowd" is in error, reliance on confidence is bound to be misleading.

Study 5 extended the MCS algorithm to a within-individual design. This extension can allow

Table 2. Percentage of correct decisions in study 5. The study was similar to study 3, but participants performed the tasks twice with a 1-week interval.

		AP (%)	*D-HC* (%)	*D-LC* (%)
Lines	CC	81.16	82.75	79.56
	CW	22.63	21.00	24.25
Shapes	CC	81.33	82.33	80.33
	CW	27.38	27.13	27.63

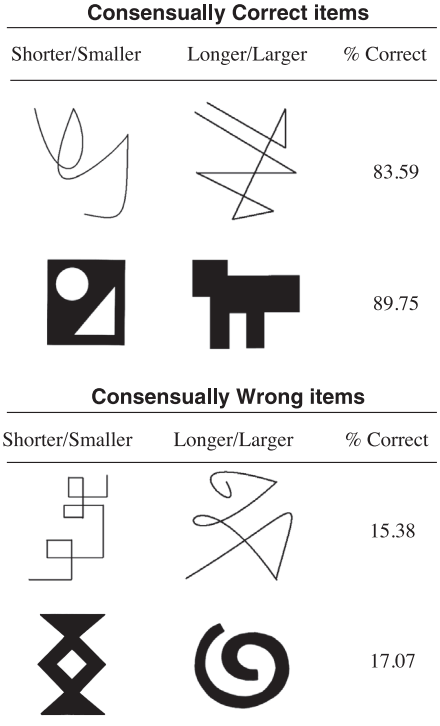

Fig. 1. Examples of the stimuli used in studies 3 and 5, divided into those for which the consensual answer was the correct answer (consensually correct) and those for which the consensual answer was the wrong answer (consensually wrong). The percentage of correct responses in study 3 is indicated.

generalizing the findings beyond the theoretical context of (*6*), which focused on the benefits of social interaction. Indeed, predictions from SCM were confirmed for a within-individual design: When participants responded to 2AFC items on several occasions, they were more confident when they made their more frequent decision than when they made their less frequent decision (*7*–*9*). The tasks used in study 3 were presented twice with a 1-week interval. The hypothesis tested is that a compilation of the high-confidence choices across the two presentations should yield the same pattern of results as that observed for a between-person compilation.

The extension of the wisdom-of-crowds idea to a within-person context (*26*–*28*) indicated that when participants estimated a quantity on two occasions, their average estimate was closer to the truth than their individual estimates. Therefore, in Study 5 *D-HC* performance was compared with average performance (*AP*) across the two sessions. The items were classified as CC or CW by using the same classification as in (*8*) and in study 3. Between-session differences in confidence were first nullified by setting the mean and SD of confidence judgments in session 2 as those of session 1 for each task and participant. For each item, the response associated with higher confidence across the two sessions was slated to *D-HC* and the other to *D-LC*. Mean percent correct for *D-HC*, *D-LC*, and *AP* is presented in Table 2 for the CC and CW items in each task.

A three-way ANOVA, task (Shapes versus Lines) × measure (*D-HC* versus *AP*) × item type (CC versus CW) yielded $F_{1,49} = 5.03$, MSE = 24.83, $P < 0.05$ for the measure × item type interaction. For the CC items, *D-HC* accuracy (82.54%) was higher than *AP* accuracy (81.24%) [$t(49) = 2.67$, $P < 0.05$]. For the CW items, *D-HC* accuracy (24.06%) tended to be somewhat lower than *AP* accuracy (25.00%) [$t(49) = 1.05$, $P < 0.31$]. In addition, across the CC items, confidence was higher when the correct choice was made than when the wrong choice was made, whereas the opposite was true for the CW items.

A comparison of *D-HC* with *D-LC* suggests that the benefit from the MCS algorithm in the case of CC items was more limited for the within-person confidence-based slating (study 5; 2.68 percentage points) than for the cross-person slating (study 3; 6.62 percentage points). The reason derives from the greater independence between decisions of two members of a dyad (study 3; a correlation of 0.02) than between the two decisions of the same person (study 5; a correlation of 0.63) (*28*).

The present work delivers three messages. First, under many conditions in which participants' decisions are correct by and large, a 2HBT1 effect should be observed. The results of the present study are consistent with Bahrami *et al.*'s (*6*) proposition that the benefit from dyadic interaction derives from individuals communicating their level of confidence accurately to each other. Here, however, a 2HBT1 effect was observed (studies 1 and 2) in the absence of social interaction. The selection of responses on the basis of confidence improved accuracy beyond the improvement achieved by the aggregation of responses across individuals (*15*).

Second, however, in situations in which most participants tend to make the wrong decisions, the MCS algorithm, as well as social interaction, is expected to yield group decisions that are even less accurate than those of each individual alone. In such cases, it is the low-confidence individuals who are more likely to be correct, and reliance on the more confident members should lead the group astray.

Last, the within-individual results (study 5) highlight a general perspective for the analysis of decision accuracy that goes beyond the effects of social interaction (*6*). This perspective, as captured by SCM, involves the variations in confidence that occur both within individuals and between individuals when choice and confidence are based on the sampling of clues from a common database (*7*, *27*).

References and Notes

1. R. S. Baron, in *Advances in Experimental Social Psychology*, M. P. Zanna, Ed. (Elsevier Academic Press, San Diego, CA, 2005), pp. 219–253.
2. J. K. Esser, *Organ. Behav. Hum. Decis. Process.* **73**, 116 (1998).
3. I. Janis, Ed., *Groupthink: A Psychological Study of Policy Decisions and Fiascos* (Houghton Mifflin, Boston, 1982).
4. G. W. Hill, *Psychol. Bull.* **91**, 517 (1982).
5. P. R. Laughlin, E. C. Hatch, J. S. Silver, L. Boh, *J. Pers. Soc. Psychol.* **90**, 644 (2006).
6. B. Bahrami *et al.*, *Science* **329**, 1081 (2010).
7. A. Koriat, *Psychol. Rev.* **119**, 80 (2012).
8. A. Koriat, *J. Exp. Psychol. Gen.* **140**, 117 (2011).
9. A. Koriat, S. Adiv, *Soc. Cogn.* **29**, 577 (2011).
10. J. Dunlosky, J. Metcalfe, *Metacognition* (Sage, Thousand Oaks, CA, 2009)
11. A. Koriat, *J. Exp. Psychol. Learn. Mem. Cogn.* **34**, 945 (2008).
12. W. F. Brewer, C. Sampaio, *Memory* **14**, 540 (2006).
13. A. Koriat, *Mem. Cognit.* **4**, 244 (1976).
14. A. Koriat, *J. Exp. Psychol. Gen.* **124**, 311 (1995).
15. D. Ariely *et al.*, *J. Exp. Psychol. Appl.* **6**, 130 (2000).
16. R. Clemen, *Int. J. Forecast.* **5**, 559 (1989).
17. J. Surowiecki, *The Wisdom of Crowds* (Anchor Books, New York, 2005).
18. T. S. Wallsten, A. Diederich, *Math. Soc. Sci.* **41**, 1 (2001).
19. For a popular presentation, see (*29*).
20. D. Kahneman, *Thinking Fast and Slow* (Farrar, Straus & Giroux, New York, 2011).
21. B. L. Cutler, S. D. Penrod, T. E. Stuve, *Law Hum. Behav.* **12**, 41 (1988).
22. Z. L. Tormala, D. D. Rucker, *Soc. Personal. Psychol. Compa.* **1**, 469 (2007).
23. K. Kleitman, L. Stankov, *Appl. Cogn. Psychol.* **15**, 321 (2001).
24. Materials and methods, and additional analyses, are available as supplementary materials on *Science* Online.
25. M. K. Dhami, R. Hertwig, U. Hoffrage, *Psychol. Bull.* **130**, 959 (2004).
26. S. M. Herzog, R. Hertwig, *Psychol. Sci.* **20**, 231 (2009).
27. K. L. Hourihan, A. S. Benjamin, *J. Exp. Psychol. Learn. Mem. Cogn.* **36**, 1068 (2010).
28. E. Vul, H. Pashler, *Psychol. Sci.* **19**, 645 (2008).
29. R. Burton, *On Being Certain: Believing You Are Right Even When You're Not* (St. Martin's Press, New York, 2008).

Acknowledgments: I am grateful to R. Gil and D. Klein for their help in the analyses of the results, and to S. Adiv for her assistance in the preparation of the article. Support for this project was received from the Max Wertheimer Minerva Center for Cognitive Processes and Human Performance, University of Haifa.

Suppplementary Materials

www.sciencemag.org/cgi/content/full/336/6079/360/DC1
Materials and Methods
SOM Text
Table S1
References

14 November 2011; accepted 29 February 2012
10.1126/science.1216549

Structure of an Intermediate State in Protein Folding and Aggregation

Philipp Neudecker,[1,2,3,4,5] Paul Robustelli,[6] Andrea Cavalli,[6] Patrick Walsh,[1,7] Patrik Lundström,[1,2,3] Arash Zarrine-Afsar,[1,2] Simon Sharpe,[1,7] Michele Vendruscolo,[6] Lewis E. Kay[1,2,3,7]*

Protein-folding intermediates have been implicated in amyloid fibril formation involved in neurodegenerative disorders. However, the structural mechanisms by which intermediates initiate fibrillar aggregation have remained largely elusive. To gain insight, we used relaxation dispersion nuclear magnetic resonance spectroscopy to determine the structure of a low-populated, on-pathway folding intermediate of the A39V/N53P/V55L (A, Ala; V, Val; N, Asn; P, Pro; L, Leu) Fyn SH3 domain. The carboxyl terminus remains disordered in this intermediate, thereby exposing the aggregation-prone amino-terminal β strand. Accordingly, mutants lacking the carboxyl terminus and thus mimicking the intermediate fail to safeguard the folding route and spontaneously form fibrillar aggregates. The structure provides a detailed characterization of the non-native interactions stabilizing an aggregation-prone intermediate under native conditions and insight into how such an intermediate can derail folding and initiate fibrillation.

Insoluble β sheet–rich fibrillar aggregates, called amyloid fibrils, form conspicuous deposits in tissue associated with a wide range of human pathologies, including Alzheimer's and Parkinson's diseases and type 2 diabetes (*1*–*4*). Fibril formation has been reported for many