Supplementary Material for

When Are Two Heads Better Than One and Why?

Asher Koriat*

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

Published 20 April 2012, *Science* **336**, 360 (2012)
DOI: 10.1126/science.1216549

This PDF file includes:

Supporting Online Material

Materials, Methods and Analyses

Study 1

Method

Participants. 40 undergraduates at the University of Haifa (21 females and 19 males) participated in the experiment for pay.

Materials. An attempt was made to duplicate the materials and procedure of (*S*1) as closely as possible. Each trial included two displays, each consisting of six Gabor patches (standard deviation of the Gaussian envelope: 0.45 degrees; spatial frequency: 1.5 cycles/degree; contrast: 10%) organized around an imaginary circle (radius: 8 degrees) at equal distances from each other. One of the two displays contained an oddball target that was generated by elevating the contrast of one of the six Gabor patches.

As in (*S1*), the study included 256 trials in which the stimuli were randomly selected from all possible combinations of 3 parameters: oddball contrast (1.5%, 3.5%, 7.0% or 15%), target set (first or second display), and target location (one of 6 possible locations). However, the same stimuli were presented to all participants in the same random order. Presentation duration of the displays was set at 220 ms in order to obtain a reasonable degree of accuracy.

Apparatus and Procedure. The experiment was conducted on a personal computer and consisted of two successive sessions on the same day. Each session included 8 blocks of 16 trials each. In each trial, participants decided which of the two displays contained an

oddball target and indicated their confidence in their decision on a 50% -100% scale reflecting the assessed likelihood that the answer was correct. The participant initiated each trial by clicking the mouse. A central fixation cross (width: 0.75 degrees of visual angle) appeared on the screen for a variable period, drawn uniformly from the range 500-1000 ms. The two displays were separated by a blank display lasting 1000 ms. The fixation cross was replaced by a question mark after the second display to prompt the participants to respond. The response options (*1/2*) were added beneath the question mark. Participants clicked the chosen response and then marked their confidence by sliding a pointer on a slider using the mouse (a number in the range 50-100 corresponding to the location of the pointer on the slider was shown in a box). The instructions had indicated that 50% represents a chance level, and that participants should try to use the full range between 50% and 100%. After clicking a "confirm" box, the next trial began. The study began with a practice block of 16 trials. The two experimental sessions then followed, with a short break between them.

Analyses

The 40 participants were paired ad hoc on the basis of their percent correct to form 20 virtual dyads, matched roughly in terms of their percent correct. However, because one participant used a 50% confidence judgment throughout the experiment, that participant, as well as her matched participant, were eliminated from the analyses. The confidence judgments of the remaining 38 participants were standardized so that the mean and STD of each participant were the same as those of the raw scores across all participants.

Within each dyad, the member with higher percent correct was designated as High Performing (*HP*), and the other as Low Performing (*LP*), and in case of a tie, one was randomly designated as *HP*.

The within-person gamma correlation averaged .38, $t(37) = 14.98$, $p < .0001$, consistent with the assumption (see *S1)* that participants are able to monitor the accuracy of their performance in this task. A dyadic gamma correlation was calculated for each dyad as follows: The responses of both members of each dyad were collapsed to form a string of 512 confidence and accuracy values, and all items on which participants gave the same response were eliminated. A gamma correlation was then calculated across the remaining items. This gamma (based on an average of 198 items per dyad), averaged .16, $t(18) = 8.29$, $p < .0001$. Thus, when participants disagreed, the decision associated with higher confidence was correct. Also, across 254 items (for 2 items all participants gave the correct response), participants who were correct were significantly more confident (76.21%) than those who were wrong (70.02%), $t(253) = 11.92$, $p < .0001$.

The agreement between the members of each dyad was relatively high: The gamma correlation across trials between the choices of the two members averaged .42, $t(18) = 10.35$, $p < .0001$. The within-dyad gamma correlation between degree of agreement (scored as 1 for agreement and 0 for disagreement) and dyadic gain in accuracy (i.e., *D-HC* minus *HP*) was -.12, $t(18) = 5.62$, $p < .0001$, suggesting that the gain from the MCS algorithm increased with the degree of independence between the choices of the members of each dyad (*S2).*

To examine whether three heads are better than two, the original dyads were retained but one member (from another dyad) was added to each dyad to form 19 triplets so that the members of each triplet were matched as closely as possible on percent correct. The performance of the triplet *D-HC,* which was based on the response of the most confident

participant in the triplet, was better than that of the dyadic *D-HC*, which was based on the more confident member in the original dyad.

Study 2

Method

Participants. 60 Hebrew-speaking University of Haifa undergraduates (34 females and 26 males) participated in the experiment either for course credit or for payment.

Materials. A general-knowledge task was used, with the stimuli constructed to be representative of their domain. A list of all 45 names of all European countries taken from the Complete Atlas of the World (London: Dorling Kindersley, c2007; The Vatican City was not included) was used. Two lists of pairs were formed for each of the two tasks, Area and Population. Each list was created by pairing randomly each of the countries with a different country, so that each country appeared once in the first position and once in the second position, avoiding pair repetitions within list or across lists (e.g., Spain-France, and France-Spain). The pairing was otherwise random. In this manner, the 90 pairs for each task could be treated as independent items. The same pairing was retained across participants except that the order of the pairs as well as the order of the countries within a pair was determined randomly for each participant. The actual area and population of each country was determined on the basis of The Europa year book (London: Europa Publications, JN1.E85 000196018 www.europaworld.com).

Apparatus and Procedure. The experiment was conducted on an IBM-compatible personal computer and included two sessions, one involving the Area task, and the other involving the Population task, in counterbalanced orders across participants. The two lists

were presented in sequence to each participant, each list preceded by 5 practice pairs (involving non-European countries).

On each trial, the names of the two countries (in Hebrew) appeared side-by-side on the screen, with a button beneath each country. In the Area task, participants were asked to decide which of the two countries has a larger area, whereas in the Population task they were asked to decide which country has a larger population. For both tasks, participants indicated their answer by clicking the button beneath the country that corresponded to their answer. Immediately after responding, a confidence scale (50%–100%) appeared beneath the two buttons, and participants marked their confidence by sliding a pointer on the scale using the mouse, and then clicking a "confirmation" button. The instructions had indicated that 50% represents a chance level and that participants should try to use the full range between 50% and 100%. Participants initiated the next trial by pressing the space bar.

Analyses

The participants were paired ad hoc as in Study 1. The results for each of the two tasks (Table S1) replicated the pattern obtained in Study 1. First, accuracy was better for *D-HC* than for *D-LC* for the Area task, $t(29) = 6.54$, $p < .0001$, and for the Population task, $t(29) = 5.34$, $p < .0001$.

Second, performance was more accurate for *D-HC* than for *HP* for both the Area task, $t(29) = 5.93$, $p < .0001$, and the Population task, $t(29) = 4.95$, $p < .0001$. For the Area task, the superiority of *D-HC* was observed for 24 out of 27 dyads (3 dyads yielded equal performance), $p < .0001$, by a binomial test. For the Population task this superiority was found for 21 out of 26 dyads (4 dyads yielded equal performance), $p < .0001$, by a binomial test. This pattern supports the 2HBT1 effect.

Finally, as in Study 1, percent correct was significantly lower for *D-LC* than for *LP*, $t(29) = 5.93$, $p < .0001$, for the Area task, and $t(29) = 4.95$, $p < .0001$, for the Population task. The same pattern of results as in Table S1 was observed for the raw (rather than standardized) confidence judgments.

The within-person gamma correlation was .43 for the Area task and .45 for the Population task. A Dyadic Gamma correlation, calculated as in Study 1, averaged .24, $t(29) = 7.89$, $p < .0001$, for the Area task (based on an average of 46 items per dyad), and .20, $t(29) = 6.21$, $p < .0001$, for the Population task (based on an average of 42 items). Thus, when participants disagreed, the more confident participant was the more likely to be correct. Also, for each item, participants who chose the correct answer were more confident than those who chose the wrong answer. For the Area task, this was true for 78 items out of the 87 items, compared with 9 items in which the pattern was reversed (for 3 items all participants chose the correct answer), $p < .0001$, by a binomial test. The respective figures for the Population task were 72 and 8 (for 10 items all participants chose the correct answer), also significant ($p < .0001$) by a binomial test.

The gamma correlation between the within-dyad agreement and the accuracy gain averaged -.24, $t(29) = 6.35$, $p < .0001$, for the Area task, and -.17, $t(29) = 4.09$, $p < .001$, for the Population task, suggesting that the gain from the MCS algorithm increases with the degree of independence between the answers of the members of a dyad.

I also examined whether three heads are better than two. For the Area task, the performance of the triplet *D-HC* (83.41%) was better than that of the dyadic *D-HC* (81.44%), $t(29) = 4.10$, $p < .0005$. The respective means for the Population task were 82.96% and 81.96%, $t(29) = 2.27$, $p < .05$.

The results of Study 2 suggest that for a representative set of general-knowledge items, confidence-based selection of responses can improve decision accuracy.

Study 3

Method

Participants. 80 Psychology undergraduates (58 females and 22 males) participated in the experiments either for pay or for course credit, 39 in Experiment 1 (Lines), and 41 in Experiment 2 (Shapes).

Stimulus Materials. The experimental materials in Experiment 1 consisted of 40 different line drawings. These were paired to form 40 pairs so that each drawing appeared twice but each time it was paired with a different drawing. 10 additional line drawings were used to create the practice trials. The experimental materials for Experiment 2 consisted of 40 geometric shapes. They were paired to form 40 pairs so that each geometric shape appeared twice but each time it was paired with a different shape. Ten additional shapes were used to create the practice trials. Each of the stimuli subtended a visual angle of approximately 5.80 degrees.

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

Apparatus and Procedure. The experiments were conducted on an IBM-compatible personal computer. Each experiment consisted of 5 blocks in which the entire set of 40 pairs was presented. In Experiment 1, participants judged which of the two lines was

longer. The two lines appeared side by side, and remained on the screen until the participants indicated their response. After clicking a "confirm" box, participants indicated their confidence on a 0–100 scale. In Experiment 2, new participants judged which of the two shapes had a larger area. The procedure was the same as that of Experiment 1, with the exception that participants reported their confidence in the form of assessed probability in the range 50%–100%.

Further details about the methods of the two experiments are found in (*S*3).

Analyses

For the Lines task, 32 items with more than 50% correct answers were classified as CC items, and 8 items with less than 50% correct answers were classified as CW items. For the Shapes task, there were 24 CC items and 16 CW items. Both experiments included 5 blocks in which the same task was presented, but here we will focus only on the results from the first block (but the division into CC and CW was based on all blocks combined).

One participant (with the highest percent correct) was deleted from the results of each task in order to form 19 virtual dyads for the Lines task and 20 dyads for the Shapes task. The members of each dyad were matched as closely as possible on percent correct. The analyses were carried out separately for the CC and CW items for each of the two experiments.

The within-person gamma correlation averaged .35 across the CC items, $t(67) = 9.56$, $p < .0001$, but -.27 across the CW items, $t(67) = 5.04$, $p < .0001$. Importantly, this difference was observed also in a between-individual analysis: For 40 CC items, participants who chose the correct answer were more confident than those who chose the wrong answer, in comparison with 10 items that displayed the opposite trend, $p < .0001$, by

a binomial test (for 6 items all participants were correct). For the CW items, in contrast, those who gave the wrong answer tended to be the more confident. This was true for 21 items whereas 3 items exhibited the opposite trend, $p < .0005$, by a binomial test.

A dyadic gamma correlation was calculated as follows: Because confidence was assessed on different scales in the two experiments, the confidence judgments were first standardized so that the mean and STD of each participant in the Lines task were set as those in the Shapes task. After eliminating all items on which participants gave the same response, the gamma correlation calculated across the remaining items averaged .18, $t(38) = 3.29$, $p < .005$ for the CC items (based on an average of 14 responses per dyad). The respective correlation for the CW items, in contrast, averaged -.28 (based on an average of 9 responses per dyad), $t(38) = 3.05$, $p < .005$. Thus, for the CC items, confidence-based dyadic selection of responses improved performance beyond what was achieved by each member alone. For the CW items, it impaired performance.

Study 4

Study 4 examined the same ideas as Study 3 using a general-knowledge task. The study was based on a reanalysis of the results of (*S4*). In that study, participants answered 2AFC general-knowledge questions and indicated their confidence.

Method

Participants. 41 Hebrew speaking University of Haifa psychology undergraduates (33 females and 8 males) participated in the experiment for course credit.

Stimulus Materials. The experiment included 105 2AFC general-knowledge items (in Hebrew), with questions covering a broad range of topics. All answers were one- or two-word long, either a concept or a name of a person or a place [e.g., "What actress played

Dorothy in the original version of the movie The Wizard of Oz? (a) Judy Garland, (b) Greta Garbo]. The questions were chosen deliberately to yield a large representation of "deceptive" or CW items.

Apparatus and Procedure. The experiment was conducted on an IBM-compatible personal computer. Each question remained on the screen until the participant pressed the space bar to indicate that he or she had finished reading it. Immediately after, the two answers, labeled *a* and *b*, were presented beneath the question, and the participant indicated his or her answer by pressing one of two keys. The statement “confidence (50%–100%)” appeared on the screen immediately after the choice of an answer. Participants typed in a number at that range, which expressed their confidence in the correctness of the answer. The order of the alternative answers was counterbalanced across participants, and the order of the questions was random for each participant. The experiment included a second session in which the entire task was repeated but here I focus only on the results from the first session. For further methodological details, see (*S4*).

Analyses

The results indicated that for 48 items, participants' choices differed significantly from 50%. Of these, there were 35 CC items and 13 CW items (with percent correct averaging 80.63% and 22.89%, respectively). The results were analyzed as in Study 3. One participant, with the highest percent correct, was deleted. The remaining 40 participants formed 20 dyads by matching the members of each dyad as closely as possible on percent correct.

The results were analyzed as in Study 3 (see Table S1). Although the differences were not strong, the overall pattern was qualitatively similar to that observed in Study 3.

For the CC items, *D-HC* exhibited the best performance, whereas for the CW items it exhibited the worst performance. A two-way ANOVA comparing *D-HC* with *HP* for CC and CW items yielded $F(1, 19) = 6.22$, $MSE = 62.86$, $p < .05$, for the interaction. *D-HC* performed better than *HP* for the CC items, $t(19) = 4.27$, $p < .001$, but somewhat worse than *HP* for the CW items, $t(19) = 1.34$, $p < .20$.

The average within-person C/A correlation was positive for the CC items, but negative for the CW items (see *S4*). This difference was observed also in a between-individual analysis (based on 39 participants because one participant gave only wrong responses to all CW items): For the CC items, participants who chose the correct answer tended to be more confident (80.53%) than those who chose the wrong answer (66.76%), $t(38) = 11.62$, $p < .0001$, whereas for the CW items, confidence was lower for those who were correct (65.88%) than for those who were wrong (70.59%), $t(38) = 2.23$, $p < .05$.

Study 5

Method

Participants. 50 University of Haifa psychology undergraduates (43 females and 7 males) participated in the experiment for pay or for course credit.

Stimulus Materials. The stimuli were the same as those used in (*S3*) (see Study 3).

Apparatus and Procedure. The experiment was conducted on an IBM-compatible personal computer. It consisted of two sessions with a one-week interval between them. Each of the sessions included 2 blocks, the first involving the Shapes task, and the second the Lines task. The procedure was the same as in study 3 except that in both tasks participants reported their confidence in the form of assessed probability in the range 50% –

100%. The order of the pairs was determined randomly for each participant and for each block.

Table S1. Percentage of correct decisions. In Study 2, participants decided which of two European countries had (a) a larger area or (b) a larger population. In Study 4, participants chose the answer to 2AFC general-knowledge questions.

	HP	*LP*	*D-HC*	*D-LC*
	Study 2			
Area	78.44%	77.93%	81.44%	74.93%
Population	79.67%	79.41%	81.96%	77.11 %
	Study 4			
CC	80.57%	79.71%	85.57%	74.71%
CW	23.08%	22.69%	19.23%	26.53%

References

S1. B. Bahrami *et al*., *Science* **329**, 1081 (2010).

S2. D. Ariely *et al.*, *J. Exp. Psychol. Appl.* **6**, 130 (2000).

S3. A. Koriat, *J. Exp. Psychol. Gen.* **140**, 117 (2011).

S4. A. Koriat, *J Exp. Psy: Learn. Mem. Cog*. **34**, 945 (2008).