Dyadic Cooperation Enhances Retrieval and Recall of Crossword Solutions

Janelle Szary (jszary@ucmerced.edu)
Rick Dale (rdale@ucmerced.edu)
Cognitive and Information Sciences, University of California, Merced, 5200 North Lake Road
Merced, CA 95343 USA

Abstract

The benefits of collaborative activities have been demonstrated in many domains, but there remain mixed results across several others as to whether collaborative groups can achieve greater performance than individuals, and can achieve greater performance than nominal group comparisons. Here we develop a task that is especially suited to testing collaborative gains. In a collaborative crossword game, two individuals solved puzzle questions either alone or collaboratively through discussion. When talking, participants solved more puzzle questions, solved them more quickly and accurately, and in general seemed to recall the words from collaborative contexts better than from matched independent contexts. By extracting the audio of their interaction, we also demonstrate interesting relationships between spoken interaction and performance on the collaborative tasks. This task environment further substantiates the notion that, in the context of knowledge retrieval, two heads are better than one.

Keywords: Dyadic cooperation; collaborative recall.

Introduction

Knowledge can be thought of as a probabilistic distribution. As samples are repeatedly taken from this distribution, a more complete picture emerges of the underlying knowledge. Often, as is implied by the phrase "the wisdom of crowds", the probability distribution is quite accurate with respect to its information representation—so that as samples are collected, an increasingly accurate picture emerges. For example, when eight-hundred attendees of a stock and poultry exhibition were asked to estimate the weight of a large ox, the mean of their estimates was very accurate (Galton, 1907). The error of the mean estimate was in fact much lower than the mean error of each individual's estimate. This "wisdom of crowds" effect has continued to be demonstrated in a number of domains: aggregate financial forecasts tend to be better than individual forecasts (Clemen, 1989), polls of the audience in game shows tend to reveal correct answers (Surowiecki, 2004).

The probabilistic nature of knowledge is also apparent when an individual accesses his or her own knowledge over time. When individuals were asked to make quantitative estimates of worldly information on two separate instances, the average of their estimates tended to be more accurate than either individual estimate (Vul & Pashler, 2008). When multiple individuals work interactively on a joint decision, however, the "two heads are better than one" intuition does not always hold. In general, the literature on group performance shows that groups rarely outperform their best members—the whole is rarely greater than the sum of its parts (Bahrami et al., 2010; Hastie & Kameda, 2005; Kerr & Tindale, 2004). In fact, across a large number of contexts, individuals tend to remember *less* when they're working with others (Rajaram & Pereira-Pasarin, 2010).

In these studies, subjects are usually presented with a list of items and must study and reproduce the items either individually or as a group. On average, groups tend to recall more items than individuals, but recall fewer items than nominal groups (consisting of the pooled, non-overlapping items recalled by individuals working alone; Barnier, Sutton, Harris & Wilson, 2008). That is, individuals working in a group context don't perform at their full potential. The leading explanation for this observation is the *retrieval disruption hypothesis* (Basden, Basden, Bryner & Thomas, 1997). According to this hypothesis, individuals use their own, idiosyncratic, strategies to organize and encode information. When recall takes place in an interactive context, the output of one member disrupts the retrieval strategies of the other(s), inhibiting recall performance.

The large body of empirical work providing evidence for the detrimental effects of collaboration on memory is unified by the term *social contagion* research (Barnier, Sutton, Harris & Wilson, 2008; and see Rajaram & Pereira-Pasarin, 2010, for a review). In addition to disrupting the recall of correct items, collaborators can even introduce the recall of incorrect items. When a confederate collaborator misleadingly recalled an incorrect item, participants later recalled that item themselves, as if it had been in the original recall list (Roediger, Meade & Bergman, 2001). This effect extends beyond laboratory recall studies, as individuals can often misremember important life events. Loftus has worked extensively on issues surrounding the fallibility of memory, especially as it applies to false memories and eyewitnesses, showing that social context can significantly impact the accuracy of memory (Loftus, 1996).

A related example of the negative consequences of social context is groupthink—a phenomenon where groups of people may end up making poor decisions, generally because of a motivation to reduce conflict and reach consensus, therefore failing to continue the search for an optimal solution (see Esser, 1998). This collaborative inhibition may be related to both retrieval disruption or *social loafing* (reduced effort or motivation when in a group context; Weldon, Blair & Huebesch, 2000).

Despite the abundance of theories and supporting evidence for social contagion, there exists an intuitive feeling that we should benefit from working with others. In addition to social contagion research, Barnier and colleagues (2008) define two other approaches to

investigating the effects of social context on memory: collaborative recall, and transactive memory. These approaches tend to seek out the beneficial effects of social context. In c*ollaborative recall* research, the social context is conceptualized as part of a broader environmental and situational context which can facilitate an individual's recall through priming. This priming could be detrimental, such as in retrieval disruption, or could be beneficial through cueing or triggering of correct information.

Bahrami and colleagues (2010) found that group performance interacted dynamically with social context. They designed a low-level perceptual decision-making task where members of a dyad reported their own decisions then agreed on a joint decision to report. When members of a dyad had unequal performance levels, the dyad tended to do worse overall than the better-performing member. However, performance exceeded aggregate individual performance when members had equal visual sensitivities and could communicate openly to discuss their observations (Bahrami et al., 2010), and when they used similar task-relevant linguistic forms (Fusaroli et al., 2012). In order to come to an agreement regarding an ambiguous low-level stimulus, members of a dyad must attempt to communicate subjective and graded confidence levels. The combination of information for higher-level decision-making tasks, such as those involving knowledge and memory, may be very different. For example, if two friends are attempting to recall the Spanish word for "countryside" from a long-ago language course, one may offer: "I think it was something like *camping*", which may trigger the other to remember the correct "campo." In this sense, members of a dyad can prime each other and iteratively build greater information.

Finally, in *transactive memory* research, the group is conceptualized as the unit of analysis: individuals are components of a coupled, distributed memory system (Wegner, 1987). In these transactive memory systems, group members may share the tasks of encoding, storing, or retrieving information in any combination. Wegner (1987) notes that memories are connected concepts—such as the concept "tomato" with the concept "red"—and these connections are formed through encoding, which can be done at the group level. As an example, consider a couple discussing the odd behavior of a mutual friend. The male partner mentions that their mutual friend seemed quiet at a recent party, while the female partner instead thought he seemed overly friendly. This reminds the man that their mutual friend had been thinking about splitting from his wife, which leads the couple to conclude that their mutual friend had been flirting with the female partner, and subsequently acted awkwardly around the male partner (from Wegner, Giuliano & Hertel, 1985). Through collaboration (discussion), the couple in this example was able to bind information and encode a quantitatively and qualitatively different memory than either would have achieved individually. Conceptualizing the distributed storage of memories is more intuitive: We already store much of our information externally (books, to-do lists, smart phones), and in much the same way we could rely on a partner to remember something for us (essentially 'outsourcing' the storage of that information to another person).

From the perspectives of both the collaborative recall and the transactional memory traditions, the performance of a group can come to be greater than the performance of its members. In this paper, we work from these intersecting perspectives to investigate the potential benefit of working with two minds instead of one on a knowledge-based trivia task. Individuals are randomly assigned to dyads and given trivia questions, which they solve either independently or collaboratively. These general knowledge trivia questions provided a set of stimuli on which subjects' knowledge varied widely, and allowed for rich discussions during collaborative sessions. Following four rounds of ten trivia questions, subjects were given individual recall tests for the answers to the preceding trivia questions.

As described by Hare (1976), research on social influence can be characterized by two factors: the "social climate", which could be either individuals collaborating or individuals working independently; and the "task completion", which is a measure of either the group product or the individual product. Consistent with previous work on joint performance measures (i.e., Hill, 1982), the current study design allowed us to first compare [1] the group product of collaborating individuals (group performance on collaborative trivia rounds) to [2] the individual product of individuals working alone (individual performances on independent trivia rounds). The recall task allowed us to compare [1] the individual product of collaborating individuals (individual recall of trivia items from collaborative rounds) to [2] the individual product of individuals working alone (individual recall of items from independent rounds).

By analyzing task performance and efficiency at the group and individual levels, and resultant memory at the individual level, we substantiate the beneficial gain of collaborative cognitive performance. Our results suggest that in knowledge-based tasks, two heads are indeed better than one.

Methods

Sixty two participants were recruited from a subject pool of University of California, Merced, undergraduate students who participated for course credit. The participants had an average age of 19.6 (SD = 1.7) and were mostly female (16 male; 46 female). The participants were grouped into thirty-one dyads. Each dyad participated in four rounds of a trivia game, where each round of ten questions was to be solved individually or collaboratively, followed by a surprise recall task after all four rounds.

Participants were seated directly across from each other at a small table with IBM ThinkPad laptop computers. This allowed each participant to have a private workspace during the independent tasks, but also enabled easy communication during the collaborative tasks. Participants wore Shure Beta

54 supercardiod microphone headsets, and their conversations were recorded using an M-Audio MobilePre recording interface and Audacity software.

Figure 1: Experimental setup.

Materials

Trivia questions were collected from a variety of crossword puzzles from www.bestcrosswords.com. Questions were all straight-forward (i.e., not "cryptic") type clues. In total, 140 questions were collected with types that were categorized as culture (n = 23), general knowledge (n = 21), definitions (n = 27), logic (n = 22), fill-in-the-blank (FITB, n = 20), categories (n = 16), and sayings (n = 11). Table 1 gives examples of each type.

Table 1: Example trivia types.

Type	Question	Answer
Culture	"Kill Bill" star Thurman	Uma
Knowledge	U.S. spy organization	CIA
Definition	Gift to charity	Donation
Logic	Hour subunits	Minutes
FITB	"If all ____ fails"	Else
Categories	Tulips and irises, for example	Flowers
Sayings	"Rolling in dough" meaning	Rich

The trivia questions were normalized for difficulty. 449 University of California, Merced undergraduate students with an average age of 18.4 (SD = 1.4; 200 male, 249 female) were given surveys containing trivia questions. There were 10 versions of the survey, each of which contained 14 trivia questions with lines indicating the number of letters the answers. Participants were allowed to leave answers blank, but were instructed to do the best they could to answer to each question, guessing when possible. Results showed that questions varied widely in difficulty (see Fig. 2). For the present study, 40 questions were chosen that were answered correctly about half of the time. As shown in Figure 2, these trivia questions were solved by 45-77% of participants, and they contained all types: culture (n = 6), general knowledge (n = 8), definitions (n = 4), logic (n = 8), fill-in-the-blank (FITB, n = 8), categories (n = 2), and sayings (n = 4). The examples in Table 1 were each used.

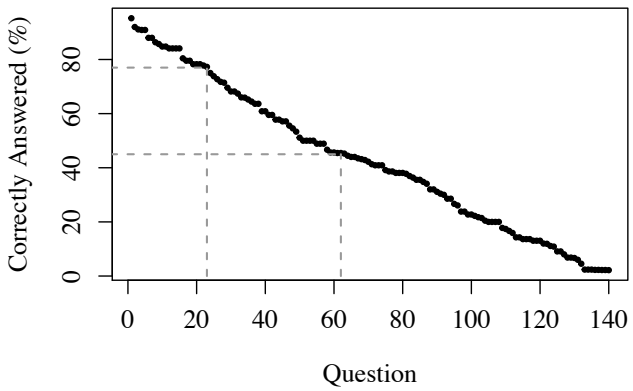

Figure 2: Question norming. Potential questions are ranked by the percentage of participants who answered correctly. Dotted lines show the question rankings we used.

Trivia Program The experimental interface was programmed by the authors using Adobe Flash CS5. The program led participants through four experimental blocks (rounds) containing ten questions each. For each round, the program instructed participants to work either individually (I) or collaboratively (C). During collaborative sessions, participants were asked to work together and discuss each answer as a team. Across all subjects, the order of questions and condition (I-C-I-C or C-I-C-I) was randomized and counterbalanced between dyads, but was kept the same within each dyad.

Each question was provided alone on the screen with a sequence of blank squares indicating the number of letters in the answer. The space-bar was used to submit answers, and subjects were given feedback about their submission. If correct, a green checkmark appeared briefly before moving on to the next question. If incorrect or missing, a red "X" marked the incorrect or blank boxes. Subjects were given 20 seconds to correctly answer each question (with as many tries as necessary) before being automatically moved on to the next question. Between blocks, subjects were given the new condition and asked to wait for their partners before moving on. Progress was indicated using flip cards with "Working" on one side, and "Ready when you are!" on the other (see Fig. 1).

Procedure

Participants were given five minutes to introduce themselves at the beginning of the study, in order to facilitate comfort and camaraderie (consistent with previous findings that more familiar groups tend to perform better on collaborative tasks; Barnier et al., 2008). After this brief familiarization period, headsets were fitted and the Flash program was started. The program began with instructions, which the researcher read aloud and subjects read on their respective screens, then the researcher left the room. After

completion of the four trivia rounds, subjects removed their headsets and summoned the researcher. The trivia program was closed and each subject was given a blank text editor. Subjects were instructed to recall and type as many of the answers to the previous trivia questions as possible. They were given five minutes and asked to work individually.

Results

Thirty-one dyads participated in the experiment, but one dyad's audio was not recorded due to equipment error. Thus, task performance results are given for thirty-one dyads, while the audio results reflect thirty dyads.

For each question, the Flash program recorded (1) whether a correct answer was submitted before time ran out. If a correct answer was achieved, it also recorded (2) how much time elapsed from the beginning of the trial to the submission of the correct answer, in milliseconds, and (3) the number of incorrect attempts before the final, correct submission. Because each participant worked on his own computer, two independent data sets were collected for each dyad. For purposes of data analysis, results for each trial were averaged over the members of the dyad. These aggregated results were used to compare each dyad's performance on individual versus collaborative rounds. Dyads are independently sampled (though, individual performance is not, as one is not independent of one's partner), and hence at the dyad level, conditions (I vs. C) can be compared using paired-samples *t*-tests (unless otherwise noted below).[1]

Trivia Performance

On all three aggregate measures, collaborative dyads outperformed their non-collaborative counterparts. Out of the twenty questions presented in each condition, the average correctly answered by collaborative dyads was 14.94 (*SD* = 3.77), while the average correctly answered by non-collaborative dyads was 12.35 (*SD* = 3.11). This difference was significant, $t(30) = 5.58$, $p < .0001$. Dyads were also faster to submit correct answers while they were collaborating (*M* = 5527ms, *SD* = 1212ms) as compared to when they were not collaborating (*M* = 6611ms, *SD* = 1181ms), and this difference was also significant, $t(30) = 3.17$, $p < .005$. Finally, the number of incorrect attempts made before achieving a correct answer was smaller for collaborative dyads (*M* = .26, *SD* = .16) than for non-collaborative dyads (*M* = .61, *SD* = .27), which is also significant, $t(30) = 7.19$, $p < .0001$.

Thus, working collaboratively conferred benefits on all three measures of task performance: it increased the number, speed, and accuracy of successful submissions. Figure 3 shows the performance gain results, where gain for each dyad is calculated as average performance on collaborative rounds, minus average performance on non-collaborative rounds.

[1] We also examined individual-level performance across most measures, and results are consistent with the dyad-level analyses.

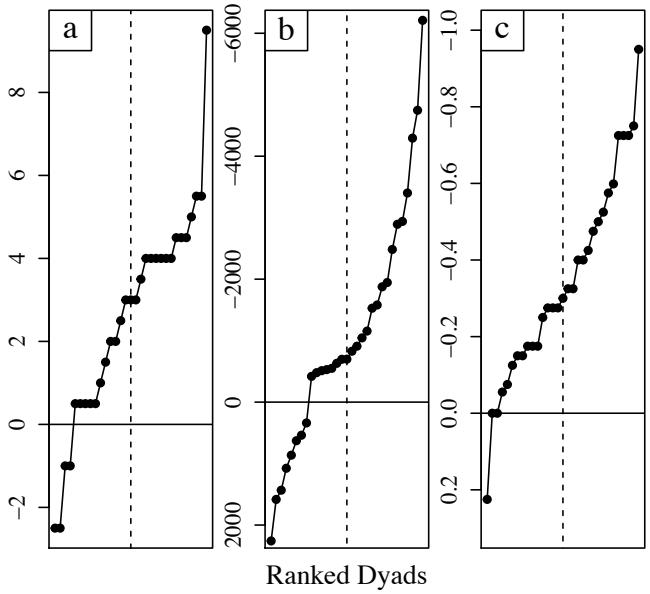

Figure 3: Collaboration gains for the following measures: (**a**) average number of correct answers, (**b**) average time taken to achieve a correct answer (ms), (**c**) average number of incorrect attempts, per question. Gain for each dyad is calculated as the difference between aggregate performance on collaborative versus non-collaborative rounds. All points above *x=0* show dyads benefitting from collaboration. For illustration, dotted lines show median ranked dyads.

Recall

The list of recalled items for each participant was first checked for accuracy and incorrect recalls were removed. This was relatively rare, however, as incorrect recalls represented only 5.7% of the total recalled items across participants (36 out of 629). Each recalled item was matched to the round and condition in which it was encountered. At the group level (i.e., averaged within dyads), the average number of items recalled from each round was, respectively, 1.60 (*SD* = .74), 2.27 (*SD* = 1.35), 1.97 (*SD* = .91), 3.71 (*SD* = 1.57). Items from the last round were recalled significantly more often than any other round, $t(30) = 4.25$, $p < .001$, indicating a serial position effect of recency. Although the mean recall from the first round was the lowest, there was also evidence of a serial position effect from primacy. This pattern is shown in Figure 4, which plots the number of recalled words from each round, binned by the number of individuals recalling each number of items. A generalized linear model, fit to the data, shows both the recency and the (more subtle) primacy effects.

In general, subjects tended to remember more items from the rounds in which they participated collaboratively. Figure 5 shows ranked, aggregated dyads' recall from each round, separated by condition. For each round there was a tendency towards enhanced recall from collaboration, but this difference was only significant in the fourth round, $t(28.88)$, $p < .05$ (Welch's two-sample *t*-test). Overall, group level

recall was not significantly better for items from collaborative rounds (M = 5.24, SD = 2.35) compared to non-collaborative rounds (M = 4.31, SD = 2.00). At the individual level, however, where dyad members are *not* aggregated and are instead treated as independent, there *was* a significant effect of condition. That is, individuals recalled more items they had encountered during collaborative rounds (M = 5.24, SD = 2.63) than during independent rounds (M = 4.31, SD = 2.47), $t(61) = 2.03$, $p < .05$. Thus, there appears to be a tendency for enhanced recall from collaboration. Admittedly, these effects are smaller than the performance measures, though more power may bear this out.

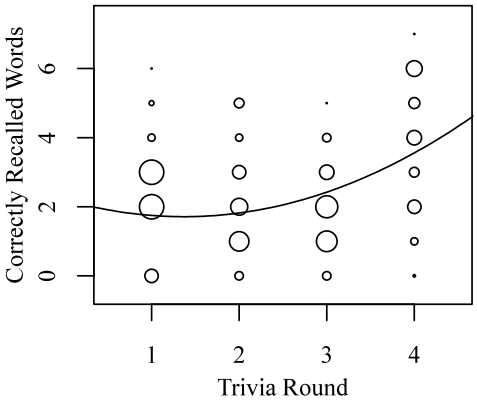

Figure 4: Binned individual-level recall per round. Circle sizes illustrate the number of individuals that recalled the corresponding number of items from each round. The line shows the fit of a generalized linear model with quadratic term.

Conversation Analysis

In order to further quantify the effects of collaboration on performance, conversations during the collaborative sessions were recorded. A coarse analysis of these recordings allowed us to collect information on the total amount of time each dyad spent in the collaborative sessions, as well as the amount of this time that was spent talking. On average, dyads spent 241.13 seconds (SD = 71.37) in the (summed) collaborative rounds, and used, on average, 109.29 of these seconds (SD = 34.72) conversing. Because the amount of time spent in the collaborative part of the task varied between dyads, a measure of percent talking was also calculated for each dyad. This percent talking measure varied from about 27% to 70% (M = 46.54, SD = 10.76).

As in the previous analyses, results were aggregated over dyads and each data point represents the group-level mean, across a dyad's participants. The total amount of time each dyad spent talking was negatively correlated with their performance, as measured by the number of correct answers they submitted during the collaborative rounds, $r(28) = -.77$, $p < .0001$. That is, the more talking they did, the worse they performed. This negative correlation may reflect the fact

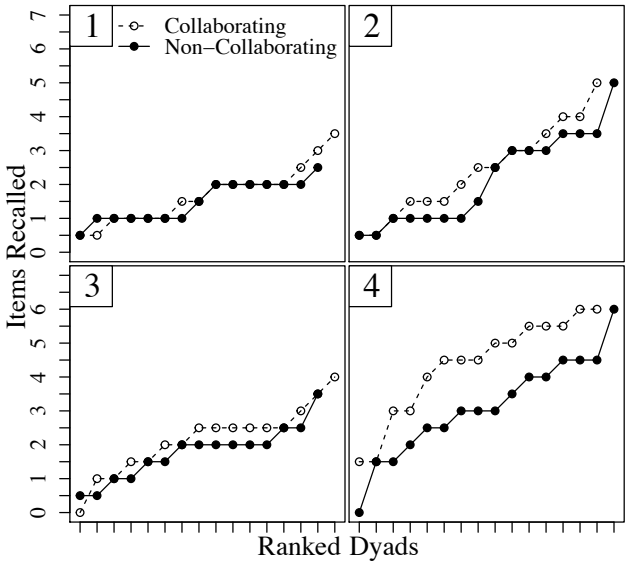

Figure 5: Recall for items from **Rounds 1-4** for each dyad, ranked in order of performance. Dotted lines with empty circles show the aggregated number of items recalled by dyads working collaboratively; Solid lines with filled circles show recall by dyads working non-collaboratively.

that when uncertain of an answer, dyads spend more time in discussion in order to figure it out. Indeed, when considering the *percentage* of time spent talking, there was a positive correlation with performance, $r(28) = .27$, although this trend did not achieve significance. Figure 6 shows the relationship between talking and performance, as measured by both absolute and percentage metrics of talking.

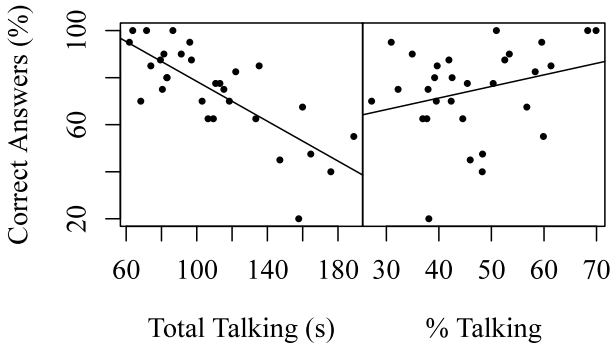

Figure 6: Relationship between talking and performance. The scatterplot on the left shows each dyad's performance (percentage of questions answered correctly) as a function of the total number of seconds spent talking (regression line m = -0.4229). On the right, performance is shown as a function of the percentage of time spent talking (regression line m = 0.4724).

General Discussion

On all measures of performance for the trivia task, there appeared to be a collaborative benefit. Aggregate dyads achieved more correct answers in the collaborative rounds than in the independent rounds, and they did so with greater accuracy. Interestingly, aggregate dyads were actually faster in the collaborative rounds than in the independent rounds, despite the fact that they had the added task of communicating with their partner for each question. With respect to the terminology described earlier (Hare, 1976), we observed that the group product, produced by collaborating individuals, was better than the individual product, produced by individuals working alone. The recall task also suggested a benefit from collaboration. Previous work has shown that participating collaboratively in recall enhances future independent recall (Basden, Basden & Henry, 2000), but our results also suggest that collaborative encoding could enhance independent recall: the individual recall product of collaborating individuals was (slightly) greater than the individual recall products of individuals acting alone.

It must be noted, however, that the present study was specifically designed to enable us to look for evidence of a collaborative gain. The collaborative benefit apparent in this situation may not apply to other situations, as previous work described earlier has found that the degree of collaborative gain is highly influenced by social context. Future work is needed to elaborate on the specifics of the social, environmental and task contexts which allow for these collaborative gains. We would also like to address the current findings in the context of interpersonal alignment, in future work. It was noted earlier that the use of similar task-relevant linguistic forms benefits dyadic cooperation, (Fusaroli et al., 2012), and a growing body of research addresses how interpersonal interactions can cause automatic alignment to spread between physical, linguistic, and other cognitive states (Tollefsen & Dale, 2012). This begs the question of whether collaborative performance on knowledge-based and memory tasks can be influenced or indicated by various levels of behavioral, linguistic, and cognitive alignment.

Acknowledgments

This research was supported partially by NSF grants BCS-0826825 and BCS-0926670. We would like to thank Jacqueline Pagobo and Maxine Varela for their assistance with data collection.

References

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329(5995)*, 1081-1085.

Barnier, A. J., Sutton, J., Harris, C. B., Wilson, R. A. (2008). A conceptual and empirical framework for the social distribution of cognition: The case of memory. *Cognitive Systems Research, 9(1-2)*, 33-51.

Basden, B. H., Basden, D. R., Bryner, S., & Thomas III, R. L. (1997). A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(5)*, 1176-1189.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559-583.

Esser, J. K. (1998). Alive and well after 25 years: A review of groupthink research. *Organizational Behavior and Human Decision Processes, 73(2-3)*, 116-141.

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(8)*, 931-939.

Galton, F. (1907). Vox populi. *Nature, 75(1949)*, 450-451.

Hare, A. P. (1976). *Handbook of small group research*. New York: Free Press.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review, 112(2)*, 494-508.

Hill, G. W. (1982). Group versus individual performance: Are $N + 1$ heads better than one? *Psychological Bulletin, 91(3)*, 517-539.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55*, 623-655.

Loftus, E. F. (1996). *Eyewitness testimony*. Cambridge: Harvard University Press.

Rajaram, S., & Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on Psychological Science, 5(6),* 649-663.

Roediger, H. L., Meade, M. L., & Bergman, E. T. (2001). Social contagion of memory. *Pschonomic Bulletin & Review, 8(2),* 365-371.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Random House.

Tollefsen, D. P. (2006). From extended mind to collective mind. *Cognitive Systems Research, 7(2-3)*, 140-150.

Tollefsen, D. P., & Dale, R. (2012). Naturalizing joint action: A process-based approach. *Philosophical Psychology, 25(3)*, 385-407.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19(7)*, 645-647.

Weldon, M. S., Blair, C., Huebsch, P. D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(6)*, 1568-1577.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior,* (pp. 185-208). New York: Springer-Verlag.

Wegner, D. M., Giuliano, T., & Hertel, P. T. (1985). Cognitive interdependence in close relationships. In W. Ickes (Ed.), *Compatible and incompatible relationships,* (pp. 253-276). New York: Springer-Verlag.